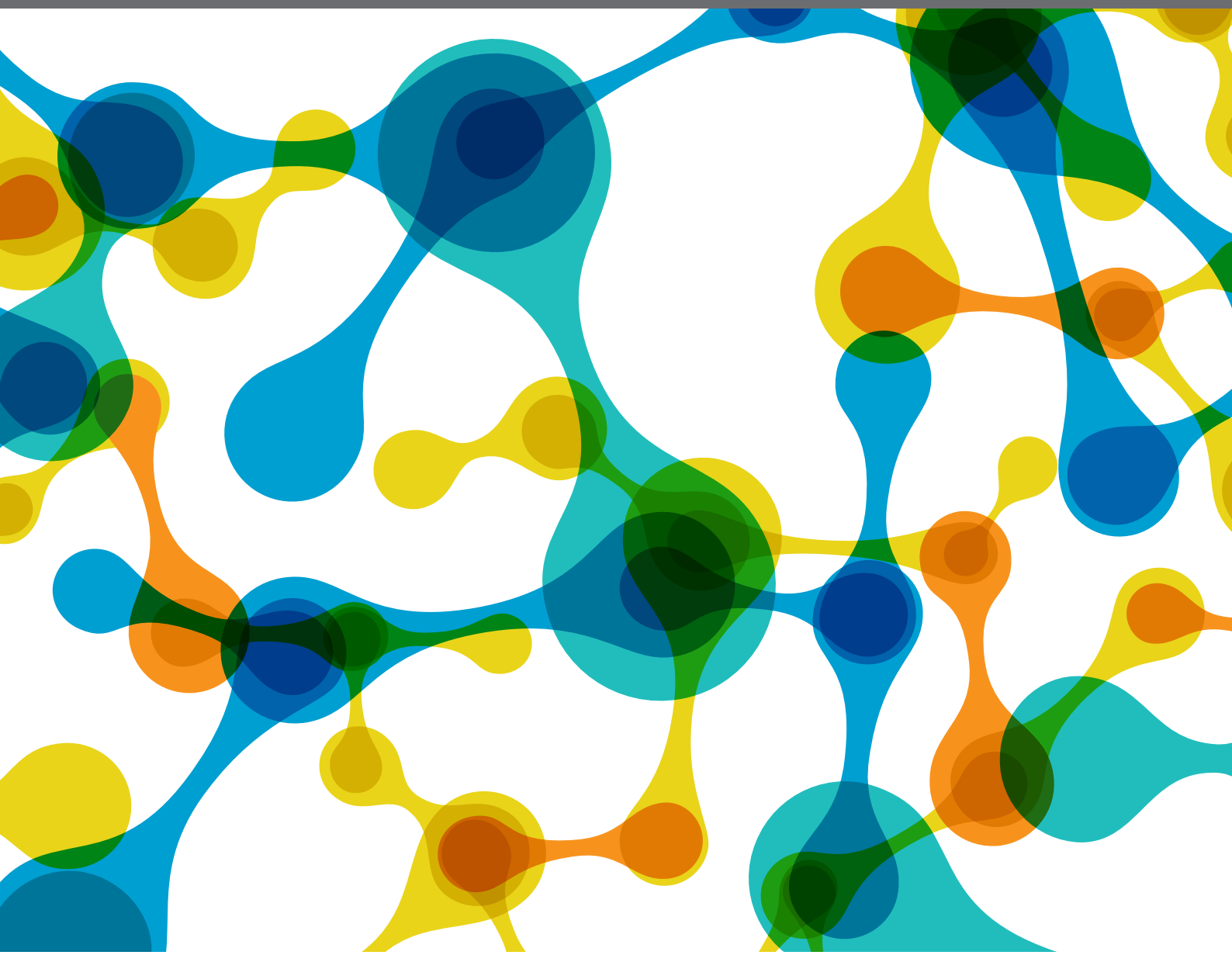
INTEGRATION OF MACHINE LEARNING AND COMPUTER SIMULATION IN SOLVING COMPLEX PHYSIOLOGICAL AND MEDICAL QUESTIONS

EDITED BY: Nicole Y. K. Li-Jessen, Gary An and Michael Döllinger
PUBLISHED IN: Frontiers in Physiology



frontiers Research Topics

frontiers

ISSN 1664-8714
ISBN 978-2-88976-673-4
DOI 10.3389/978-2-88976-673-4

About Frontiers

Frontiers is more than just an open-access publisher of scholarly articles: it is a pioneering approach to the world of academia, radically improving the way scholarly research is managed. The grand vision of Frontiers is a world where all people have an equal opportunity to seek, share and generate knowledge. Frontiers provides immediate and permanent online open access to all its publications, but this alone is not enough to realize our grand goals.

Frontiers Journal Series

The Frontiers Journal Series is a multi-tier and interdisciplinary set of open-access, online journals, promising a paradigm shift from the current review, selection and dissemination processes in academic publishing. All Frontiers journals are driven by researchers for researchers; therefore, they constitute a service to the scholarly community. At the same time, the Frontiers Journal Series operates on a revolutionary invention, the tiered publishing system, initially addressing specific communities of scholars, and gradually climbing up to broader public understanding, thus serving the interests of the lay society, too.

Dedication to Quality

Each Frontiers article is a landmark of the highest quality, thanks to genuinely collaborative interactions between authors and review editors, who include some of the world's best academicians. Research must be certified by peers before entering a stream of knowledge that may eventually reach the public - and shape society; therefore, Frontiers only applies the most rigorous and unbiased reviews.
Frontiers revolutionizes research publishing by freely delivering the most outstanding research, evaluated with no bias from both the academic and social point of view. By applying the most advanced information technologies, Frontiers is catapulting scholarly publishing into a new generation.

What are Frontiers Research Topics?

Frontiers Research Topics are very popular trademarks of the Frontiers Journals Series: they are collections of at least ten articles, all centered on a particular subject. With their unique mix of varied contributions from Original Research to Review Articles, Frontiers Research Topics unify the most influential researchers, the latest key findings and historical advances in a hot research area! Find out more on how to host your own Frontiers Research Topic or contribute to one as an author by contacting the Frontiers Editorial Office: frontiersin.org/about/contact

INTEGRATION OF MACHINE LEARNING AND COMPUTER SIMULATION IN SOLVING COMPLEX PHYSIOLOGICAL AND MEDICAL QUESTIONS

Topic Editors:
Nicole Y. K. Li-Jessen, McGill University, Canada
Gary An, University of Vermont, United States
Michael Döllinger, University Hospital Erlangen, Germany

Citation: Li-Jessen, N. Y. K., An, G., Döllinger, M., eds. (2022). Integration of Machine Learning and Computer Simulation in Solving Complex Physiological and Medical Questions. Lausanne: Frontiers Media SA. doi: 10.3389/978-2-88976-673-4

Table of Contents

frontiers | Frontiers in Physiology

EDITORIAL
published: 05 July 2022
doi: 10.3389/fphys.2022.949771

Editorial: Integration of Machine Learning and Computer Simulation in Solving Complex Physiological and Medical Questions

Gary An[1], *Michael Döllinger*[2] *and Nicole Y. K. Li-Jessen*[3,4,5]*

[1]Department of Surgery, University of Vermont Larner College of Medicine, Burlington, VT, United States, [2]Department of Otorhinolaryngology Head and Neck Surgery, Medical School, Division of Phoniatrics and Pediatric Audiology, University Hospital Erlangen, Friedrich-Alexander-University Erlangen-Nürnberg, Erlangen, Germany, [3]School of Communication Sciences and Disorders, McGill University, Montreal, QC, Canada, [4]Department of Otolaryngology-Head and Neck Surgery, McGill University, Montreal, QC, Canada, [5]Department of Biomedical Engineering, McGill University, Montreal, QC, Canada

Keywords: machine learning, computer simulation, complex disease, personalized medcine, high fidelity computational method, multi-scale modeling

Editorial on the Research Topic

Integration of Machine Learning and Computer Simulation in Solving Complex Physiological and Medical Questions

OPEN ACCESS

Edited and reviewed by:
Raimond L. Winslow, Northeastern University, United States

***Correspondence:**
Nicole Y. K. Li-Jessen
nicole.li@mcgill.ca

Specialty section:
This article was submitted to Computational Physiology and Medicine, a section of the journal Frontiers in Physiology

Received: *21 May 2022*
Accepted: *07 June 2022*
Published: *05 July 2022*

Citation:
An G, Döllinger M and Li-Jessen NYK (2022) Editorial: Integration of Machine Learning and Computer Simulation in Solving Complex Physiological and Medical Questions. Front. Physiol. 13:949771. doi: 10.3389/fphys.2022.949771

BACKGROUND

This Research Topic, "*Integration of Machine Learning and Computer Simulation in Solving Complex Physiological and Medical Questions*", brings together two powerful computational approaches to investigate complex disease processes: the use of high-fidelity, mechanism-based simulation models (MSMs), and the training of artificial neural networks (ANNs) via machine learning (ML) and artificial intelligence (AI). These two approaches represent distinct aspects of the scientific process: ML/AI involves correlation identification/hypothesis generation whereas MSMs provide an in silico means for hypothesis testing and conceptual model verification, with capabilities that can complement and address each other's limitations. High-fidelity MSMs can contain very large numbers of parameters, which poses challenges to effective parameterization and/or parameter space exploration, and can present prohibitive computational costs in terms of executing simulation experiments. Alternatively, ML/AI approaches are notoriously data-hungry (a considerable issue when dealing with biological data sets that are generally orders of magnitude more sparse compared to other ML applications), are highly limited in terms of testing inferred causal relationships, and are often "black boxes" in terms of interpreting why the ANNs do what they do. This Research Topic brings together work that integrates MSM and ML in a complementary fashion. We have organized these papers in the following general classes of investigation.

APPLICATIONS OF INTEGRATED ML AND MSM IN PERSONALIZED MEDICINE

The ostensible goal of the practice of medicine is to treat sick individuals with the right drug and the right time, and be able to have such a treatment regimen for every sick patient. MSMs can serve as "digital twins" of individual patients and provide a means of virtually forecasting their future disease

course, or, with future developments, aid in personalizing potential therapies. Implicit in this process is the need to capture disease trajectories over time (i.e., integrating time series data), which challenges data-hungry pure ML approaches, but also requires tuning a simulation model to a specific person's "parameters." Kuruvila et al. combined convolutional neural network (CNN) and long short-term memory (LTSM) models to infer a listener's auditory attention in noisy acoustic environments. CNNs were trained with experimental data of electroencephalography (EEG) and speech spectrograms from speakers. The CNN outputs then parsed to the bidirectional LSTM and the auditory attention to speakers were classified. Their results supported the integration of listener-specific EEG signals into ML-powered hearing aids that will help listeners attend to speech signals in noisy scenarios. Schafer et al. applied physics-based network diffusion models to simulate the propagation of misfolded tau proteins in three brain regions of patients with Alzheimer's disease. Hierarchical Bayesian Inference models were used to obtain posterior probability distributions for two personalized model parameters, namely, the diffusion coefficient and production rate of tau proteins. Personalized models of tau pathology with capability of predicting tau evolution and their associated cognitive functions would be of great use in creating virtual patient controls for clinical trials. van Duuren et al. combined bi-objective evolutionary algorithm (EA) and an established microsimulation model for personalized colorectal cancer screening. EA was used to find personalized screening policies in minimizing the costs while maximizing the number of Quality-Adjusted Life Years gained. Their study results supported the use of computer models to guide policy making and implementation of personalized colorectal cancer screening.

MACHINE LEARNING AS SURROGATE MODELS OF COMPLEX MECHANISTIC MODELS

Developing "lighter weight" surrogate models of complex MSMs would enhance the computational efficiency of simulation experiments. ANNs, as governed by the universal Approximation Theorem (Hornik et al., 1989) are able to recapitulate any generative function and are therefore appealing means of creating surrogate models. Quetzalcóatl Toledo-Marín et al. applied this principle to partial differential equation (PDE) models of biological diffusion. In this case, there is considerable improvement in performance with the surrogate ANN, which allows for both greater complexity of the MSM and more extensive exploration of possible behaviors with simulation experiments. Alternatively, there are types of MSMs that do not have a readily accessible equation form, primarily agent-based models (ABMs). Larie et al. uses ABM simulation data to create a surrogate ANN, but with certain caveats related to properties often found in biomedical ABMs. Firstly, in contrast to deterministic equation-based models, instead of specific trajectories ANN surrogates of ABMs generate a probabilistic "cone" of future trajectories (ala hurricane path prediction). As such, any attempt to use such surrogate models needs to account for this projected uncertainty with updating to produce a rolling forecast horizon. Secondly, the ANN of the ABM also shows the property of path non-uniqueness, which has implications regarding attempts to "reverse engineer" particular pathway or causal network structures from biological data.

ML-BASED PARAMETER SPACE CHARACTERIZATION METHODS FOR HIGH-FIDELITY MSMS

Complex medical problems require complex solutions. However, there is a tension between using models simple enough to readily parameterize but do not capture key details necessary for clinical utility versus sufficiently expressive but highly complex models with a host of parameters that may not be accessible experimentally. ML methods have thus been applied to the problem of parameter space characterization and uncertainty quantification (Granato and Li-Jessen, 2020) through Model Exploration (ME) (Ozik et al., 2018) methods, such as Random Forest (Garg et al., 2019). Alarid-Escudero et al. utilized the Extreme-scale Model Exploration with Swift (EMEWS) framework for high performance computing (HPC) enabled ME to characterize how experimentally unidentifiable parameters affected the performance of microsimulation decision models regarding the natural history of colorectal cancer (CRC). Leaving these known factors out of a decision model would lead to an intuitively inferior model, and therefore this group used EMEWS to infer regions of identifiable parameter space that produced clinically relevant alterations in the decision model outputs. A different perspective is presented in the paper by Cockrell & An using genetic algorithms (GA) to calibrate a complex ABM. This study introduces a formal mathematical object, the Model Rule Matrix (MRM), intended to account for the inherent "incompleteness" of any mechanism-based simulation model by accounting for all the possible "missing connections" as model parameters. Therefore, as opposed to "parameter fitting" that attempts to reduce experimental/clinical data variation, this approach expands the range of allowable model parameterizations given real-world observations.

CONCLUSION

Future work will invariably continue leveraging the strengths of MSMs and ML to offset their inherent limitations. Moving forward, we note multiple open challenges remain, two of which we briefly note:

- The use of synthetic data is ubiquitous in most non-biomedical applications of ML/AI. This need is even more pronounced given the relative sparsity of biological data. However, given the universal Approximating capabilities of

ANNs, care must be taken when generating biological time series data such that the ANN does not only "learn" to the generative model. Therefore, developing means to "hide" the generative model from the ANN is a crucial area of investigation and development. The paper in this Research Topic by Cockrell & An begins to address this issue.

- One main concern regarding the use of ML/AI in biomedicine is the opacity of these systems. "Explainable" or "interpretable" AI is a key research topic in the general AI community. The use of MSMs to generate synthetic data can aid in addressing the transparency issues, as the MSMs are explicitly transparent (by virtue of their programmed structure) and essentially represent the conceptual-symbolic model that many consider a necessary component of next generation AI systems. (Garcez and Lamb, 2020).

We hope that the papers in this Research Topic will help spur additional developments and applications in what we consider to be an essential set of methods to better understand and treat complex medical diseases.

REFERENCES

Garcez, Ad. A., and Lamb, L. C. (2020). Neurosymbolic AI: The 3rd Wave. *arXiv preprint arXiv:201205876*.

Garg, A., Yuen, S., Seekhao, N., Yu, G., Karwowski, J., Powell, M., et al. (2019). Towards a Physiological Scale of Vocal Fold Agent-Based Models of Surgical Injury and Repair: Sensitivity Analysis, Calibration and Verification. *Appl. Sci.* 9 (15), 2974. doi:10.3390/app9152974

Granato, B., and Li-Jessen, N. Y. (2020). Sensitivity Analysis for Dimensionality Reduction in Agent-Based Modeling. In ECAI 2020. IOS Press 2905–2906.

Hornik, K., Stinchcombe, M., and White, H. (1989). Multilayer Feedforward Networks Are Universal Approximators. *Neural Netw.* 2 (5), 359–366. doi:10.1016/0893-6080(89)90020-8

Ozik, J., Collier, N. T., Wozniak, J. M., Macal, C. M., and An, G. (2018). Extreme-Scale Dynamic Exploration of a Distributed Agent-Based Model with the EMEWS Framework. *IEEE Trans. Comput. Soc. Syst.* 5 (3), 884–895. doi:10.1109/tcss.2018.2859189

AUTHOR CONTRIBUTIONS

All authors listed have made a substantial, direct, and intellectual contribution to the work and approved it for publication.

FUNDING

GA was sponsored in part by the National Institutes of Health Award UO1EB025825. GA is also sponsored by the Defense Advanced Research Projects Agency (DARPA) through Cooperative Agreement D20AC00002 awarded by the United States Department of the Interior (DOI), Interior Business Center. NL-J is supported by the National Sciences and Engineering Research Council of Canada (RGPIN-2018-03843 and ALLRP 548623-19), Compute Canada and Canada Research Chair research stipend. The content of the information does not necessarily reflect the position or the policy of the Government, and no official endorsement should be inferred.

Conflict of Interest: The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

Publisher's Note: All claims expressed in this article are solely those of the authors and do not necessarily represent those of their affiliated organizations, or those of the publisher, the editors and the reviewers. Any product that may be evaluated in this article, or claim that may be made by its manufacturer, is not guaranteed or endorsed by the publisher.

frontiers in Physiology

ORIGINAL RESEARCH
published: 03 March 2021
doi: 10.3389/fphys.2021.628502

Deep Learning in Automatic Sleep Staging With a Single Channel Electroencephalography

*Mingyu Fu[1], Yitian Wang[1], Zixin Chen[2], Jin Li[3], Fengguo Xu[4], Xinyu Liu[1] and Fengzhen Hou[1]**

[1] School of Science, China Pharmaceutical University, Nanjing, China, [2] College of Engineering, University of California, Berkeley, Berkeley, CA, United States, [3] College of Physics and Information Technology, Shaanxi Normal University, Xi'an, China, [4] Key Laboratory of Drug Quality Control and Pharmacovigilance, China Pharmaceutical University, Nanjing, China

This study centers on automatic sleep staging with a single channel electroencephalography (EEG), with some significant findings for sleep staging. In this study, we proposed a deep learning-based network by integrating attention mechanism and bidirectional long short-term memory neural network (AT-BiLSTM) to classify wakefulness, rapid eye movement (REM) sleep and non-REM (NREM) sleep stages N1, N2 and N3. The AT-BiLSTM network outperformed five other networks and achieved an accuracy of 83.78%, a Cohen's kappa coefficient of 0.766 and a macro F1-score of 82.14% on the PhysioNet Sleep-EDF Expanded dataset, and an accuracy of 81.72%, a Cohen's kappa coefficient of 0.751 and a macro F1-score of 80.74% on the DREAMS Subjects dataset. The proposed AT-BiLSTM network even achieved a higher accuracy than the existing methods based on traditional feature extraction. Moreover, better performance was obtained by the AT-BiLSTM network with the frontal EEG derivations than with EEG channels located at the central, occipital or parietal lobe. As EEG signal can be easily acquired using dry electrodes on the forehead, our findings might provide a promising solution for automatic sleep scoring without feature extraction and may prove very useful for the screening of sleep disorders.

Keywords: deep learning, single channel electroencephalography, automatic sleep staging, bidirectional long short-term memory, attention mechanism

OPEN ACCESS

Edited by:
Michael Döllinger, University Hospital Erlangen, Germany

Reviewed by:
Guanghao Sun, The University of Electro-Communications, Japan
Kaare Bjarke Mikkelsen, Aarhus University, Denmark

****Correspondence:***
Fengzhen Hou
houfz@cpu.edu.cn

Specialty section:
This article was submitted to Computational Physiology and Medicine, a section of the journal Frontiers in Physiology

Received: *12 November 2020*
Accepted: *01 February 2021*
Published: *03 March 2021*

Citation:
Fu M, Wang Y, Chen Z, Li J, Xu F, Liu X and Hou F (2021) Deep Learning in Automatic Sleep Staging With a Single Channel Electroencephalography. Front. Physiol. 12:628502. doi: 10.3389/fphys.2021.628502

INTRODUCTION

Sleep is important for the optimal functioning of the brain and the body (Czeisler, 2015). However, a large number of people suffer from sleep related disorders, such as sleep apnea, insomnia and narcolepsy (Ohayon, 2002). Effective and feasible sleep assessment is essential for recognizing sleep problems and making timely interventions.

Sleep assessment is generally based on the manual staging of overnight polysomnography (PSG) signals, including electroencephalogram (EEG), electrooculogram (EOG), electromyogram (EMG), electrocardiogram (ECG), blood oxygen saturation and respiration (Weaver et al., 2005), by trained and certified technicians. According to the American Academy of Sleep Medicine (AASM) manual (Iber et al., 2007), sleep can be staged as wakefulness (WAKE), rapid eye movement (REM) sleep and non-REM (NREM) sleep, which is further divided into three stages, N1, N2 and N3. Usually, it takes about 2–4 h for a technician to mark an overnight (lasting about 8 h)

PSG. The time-consuming nature of manual sleep staging hampers its application on very large datasets and limits related research in this field (Hassan and Bhuiyan, 2016a). Moreover, the inter-scorer agreement is less than 90% and its improvement remains a challenge (Younes, 2017). The multiple channels of PSG also present drawbacks preventing wider usage for the general population, due to complicated preparation and disturbance to participants' normal sleep. Therefore, the past decades have witnessed the growth of automatic sleep staging based on single-channel EEG (Liang et al., 2012; Ronzhina et al., 2012; Aboalayon et al., 2014; Radha et al., 2014; Zhu et al., 2014; Wang et al., 2015; Hassan and Bhuiyan, 2016a, 2017; Boostani et al., 2017; Phan et al., 2017; Silveira et al., 2017; Tian et al., 2017; Lngkvist and Loutfi, 2018; Seifpour et al., 2018; Sors et al., 2018; Tripathy and Acharya, 2018). These methods may eventually lead to a sufficiently accurate, robust, cost-effective and fast means of sleep scoring (Wang et al., 2015).

In the field of machine learning, deep networks are drawing more and more attention because they can learn from data directly without manual feature extraction (Lecun et al., 2015; Tsinalis et al., 2015; Dong et al., 2016; Supratak et al., 2017; Zhang and Wu, 2017; Bresch et al., 2018; Malafeev et al., 2018; Stephansen et al., 2018). There are many useful and well-established deep networks for the data mining of time series, such as the convolutional neural network (CNN) (Lecun and Bengio, 1997) and recurrent neural network (RNN) (Elman, 1990). Although CNN has mainly been applied in automated recognition of images, its application in the analysis of time series has also been notable (Chambon et al., 2018; Cui et al., 2018; Zhang and Wu, 2018; Yildirim et al., 2019). That said, it is generally demonstrated that RNN has better performance than CNN for the analysis of time series (Fiorillo et al., 2019). One of the most widely used RNN is the Long Short-Term Memory (LSTM) neural network, which is capable of capturing the long-term dependent information underlying the temporal structure of the time series (Hochreiter and Schmidhuber, 1997). Furthermore, bidirectional LSTM (BiLSTM), composed of two unidirectional LSTMs, can read data from both ends of the time series and is able to make full use of information embedded in both directions of the time series (Schuster and Paliwal, 1997). Moreover, the concept of attention is arguably one of the most powerful in the deep learning field nowadays. It is based on a common sense intuition that we "attend to" a certain part when processing a large amount of information. This simple yet powerful concept has led to many breakthroughs, not only in natural language processing tasks, such as speech recognition (Jo et al., 2010) and machine translation (Ferri et al., 2012; Karpathy and Fei-Fei, 2014; Hassan and Bhuiyan, 2017), but also in time series analysis. Recently, Zhang et al. (2019) proposed an attention-based LSTM model for financial time series prediction and a comparative analysis conducted by Hollis et al. (2018) further demonstrates that an LSTM with attention indeed outperforms a standalone LSTM for forecasting financial time series.

The application of deep neural networks for automatic sleep staging is soaring (**Table 1**). The PhysioNet Sleep-EDF Expanded (PSEE) dataset (Goldberger et al., 2000; Kemp et al., 2000) was the most widely employed dataset in related studies. As shown in **Table 1**, Tsinalis et al. (2016) and Phan et al. (2019) reported an accuracy of 74.0% and 81.9% respectively, for 5-class sleep staging of the PSEE dataset with a CNN algorithm, while Supratak found that the combination of CNN and BiLSTM increased the accuracy to 82.4% (Supratak et al., 2017). There are also some datasets aside from PSEE that are routinely employed in studies of automatic sleep staging with a single-channel EEG and deep learning algorithms. Hsu et al. (2013) built an RNN model on the PhysioNet Sleep-EDF (PSE) dataset and achieved an accuracy of 87.2%. On the Montreal Archive of Sleep Studies (MASS) dataset, Phan et al. (2019) built a CNN model and achieved an accuracy of 83.6% while Supratak et al. (2017) built a CNN-LSTM model and obtained an accuracy of 86.2%. A CNN was also applied on the Sleep Heart Health Study (SHHS) dataset, yielding an accuracy of 87% (Sors et al., 2018). However, few works investigated whether the performance of sleep staging can be further improved by the combination of BiLSTM and the attention mechanism. Aside from that, there is a lack of comparison between the performance of deep learning based and conventional feature extraction based models.

Although deep learning algorithms have shown themselves promising in automatic sleep staging with a single-channel EEG, few studies investigated whether the performance of such algorithms is sensitive to the choice of EEG channel. Therefore, in this study, the PSEE dataset and the DREAMS Subjects (DRM-SUB) dataset (Devuyst, 2005) were used. Both datasets have more than one channel of EEG and the DRM-SUB dataset was involved in many automatic sleep staging studies with conventional feature extraction (Hassan and Bhuiyan, 2016a, 2017; Ghimatgar et al., 2019; Shen et al., 2019). A neural network named AT-BiLSTM was proposed, which uses the neural attention mechanism of the BiLSTM to classify sleep stages. For comparison, five other networks, CNN, LSTM, BiLSTM, the combination of CNN and LSTM (CNN-LSTM), and the combination of CNN and BiLSTM (CNN-BiLSTM) were also trained and tested. Our aims are threefold: first, to investigate whether AT-BiLSTM can achieve the highest performance among these networks; second, to confirm whether RNN algorithms (i.e., LSTM and BiLSTM) outperform CNN in sleep staging with single channel EEG; third, to explore whether the method of making hybrid networks further improves the performance of sleep staging.

MATERIALS AND METHODS

Datasets

The data analyzed in this study were obtained from two open-access datasets: the DRM-SUB dataset and the PSEE dataset. The DRM-SUB consists of 20 whole-night PSG recordings (lasting 7–9 h) obtained from 20 subjects (four males and 16 females, 20–65 years old). Three EEG channels located in different lobes (Cz-A1, Fp1-A1 and O1-A1) were included in DRM-SUB, with a sampling rate of 200 Hz. To investigate the impact of the choice of EEG derivations on the performance of automatic sleep staging, EEG

TABLE 1 | An overview of the application of deep networks on sleep staging.

Authors	Dataset	Channel	Model	Accuracy
Tsinalis et al.	PSEE	Fpz-Cz	CNN	74.0%
Phan et al.	PSEE	Fpz-Cz	CNN	81.9%
Supratak et al.	PSEE	Fpz-Cz	CNN-BiLSTM	82.4%
Hsu et al.	PSE	Fpz-Cz	RNN	87.2%
Phan et al.	MASS	C4-A1	CNN	83.6%
Supratak et al.	MASS	F4-EOG (Left)	CNN-BiLSTM	86.2%
Sors et al.	SHHS	C4-A1	CNN	87.0%

TABLE 2 | Data distribution of sleep stages in both datasets.

Dataset	Total epochs	WAKE (%)	N1 (%)	N2 (%)	N3 (%)	REM (%)
PSEE	41663	19.2	6.6	42.2	13.4	18.5
DRM-SUB	20265	17.6	7.3	40.7	19.4	14.9

signals from all three channels were used separately for the following analysis.

Twenty healthy subjects (10 males and 10 females, 25–34 years old) from the PSEE dataset were also included. There are two EEG channels (Fpz-Cz and Pz-Oz) available in the PSEE dataset, with a sampling rate of 100 Hz. For each subject, two PSGs of about 20 h each were recorded during two subsequent day-night periods at the subjects' homes. In order to remain consistent with previous studies (Supratak et al., 2017), for each subject and each PSG, only the data from 30 min before sleep-onset (i.e., the first sleep epoch after light-off in the evening) and 30 min after the last sleep epoch in the morning were included. Both channels were investigated separately.

For both datasets, labels of sleep staging for each 30-s EEG epoch were provided by the data distributors according to AASM rules. Five staging classes, i.e., WAKE, N1, N2, N3, and REM were used in this study. The distribution of 30-s EEG epochs of both datasets is illustrated in **Table 2**.

Construction of the AT-BiLSTM Network

The proposed AT-BiLSTM network architecture for automatic sleep staging is illustrated in **Figure 1**. It is composed of two main components, three stacked BiLSTM layers for feature exacting and one attention layer to weight the most relevant parts of the input sequence. According to a preset parameter, called the input dimension m, each raw 30-s EEG epoch is divided into multiple vectors, which are fed into the BiLSTM part sequentially to construct a feature matrix. Then to emphasize the different importance of different vectors, an attention layer is applied in the intra-epoch feature learning and summarizes the outputs of the BiLSTM part with different weights. Finally, the probability of each sleep stage can be derived from a fully connected (FC) layer and a softmax layer.

Given a 30-s EEG epoch $X \quad [x_1, x_2, , x_N]$ with N data points, a moving window with input dimension of m is applied to $\mathbf{X}$ without overlap, leading to the matrix form of $\mathbf{X}$, as shown in Equation 1, where n equals to N/m and $\mathbf{X_t}$ represents the vector in time step t.

$$\begin{bmatrix} \mathbf{X_1} \\ \mathbf{X_2} \\ \vdots \\ \mathbf{X_t} \\ \vdots \\ \mathbf{X_n} \end{bmatrix} = \begin{bmatrix} x_1 & x_2 & \cdots & x_m \\ x_{m+1} & x_{m+2} & \cdots & x_{2m} \\ \vdots & \vdots & \cdots & \vdots \\ x_{(t-1)m+1} & x_{(t-1)m+2} & \cdots & x_{t\times m} \\ \vdots & \vdots & \cdots & \vdots \\ x_{(n-1)m+1} & x_{(n-1)m+2} & \cdots & x_{n\times m} \end{bmatrix} t \in [1, n] \quad (1)$$

All the vectors are fed into the first BiLSTM layer, forward and backward respectively. For time step t, the output of the forward or backward network, denoted as $\mathbf{h_t^f}$ or $\mathbf{h_t^b}$, can be obtained, respectively, according to Equation 2 or 3.

$$\mathbf{h_t^f} = \sigma(\mathbf{W_{fx}x_t} + \mathbf{W_{ff}h_{t-1}^f} + b_f) \quad (2)$$

$$\mathbf{h_t^b} = \sigma(\mathbf{W_{bx}x_t} + \mathbf{W_{bb}h_{t-1}^b} + b_b) \quad (3)$$

where σ is the logistic sigmoid function, $\mathbf{W}$ is the weight matrix (e.g., subscription "fx" in $\mathbf{W}$ represents the forward network of $\mathbf{x_t}$) and b is the bias vector of the network (b_f and b_b represents the bias vector of forward and backward network, respectively).

The weighted sum of $\mathbf{h_t^f}$ and $\mathbf{h_t^b}$, denoted as $\mathbf{h_t}$, is computed as the output of the first BiLSTM layer following Equation 4.

$$\mathbf{h_t} = \mathbf{W_{hf}h_t^f} + \mathbf{W_{hb}h_t^b} + b_h \quad (4)$$

The output of the previous BiLSTM layer is fed into the next layer in the same way. The third layer gives the final output of the BiLSTM part, which is weighted by the attention layer before feeding into the FC layer. Considering that EEG signal in different time steps should contribute differently to the classification task, it is rational to give strong weights to the more discriminative parts and vice versa. Formally, the attention weight $\mathbf{a_t}$ at the time step t is computed according to Formula (5) – (6).

$$u_t = \tanh(\mathrm{w_w h_t} + \mathrm{b_w}) \quad (5)$$

$$a_t = \frac{\exp(u_t^T u_w)}{\sum_t \exp(u_t^T u_w)} \quad (6)$$

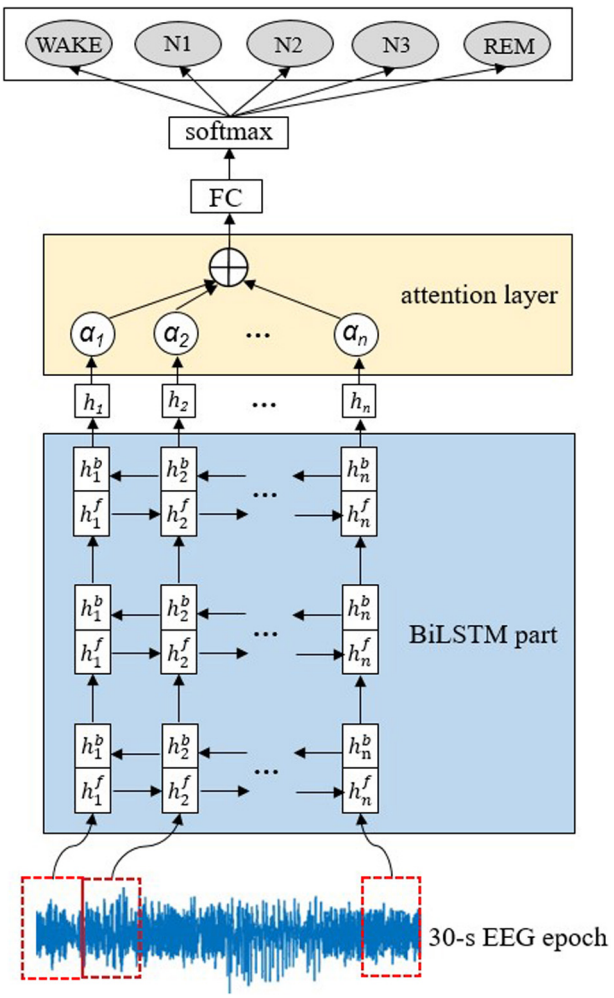


FIGURE 1 | Illustration of the proposed AT-BiLSTM network architecture for automated sleep staging. The network consists of a BiLSTM part, an attention layer, a full-connected (FC) layer and a softmax layer. The input of the network is a raw 30-s EEG time series and the output is the probability of each sleep stage. The dashed rectangle on the EEG time series represents a vector of EEG signals at a time step.

In Formula (5)–(6), $\mathbf{u_t}$ represents the state of the hidden layer obtained from a simple neural network, $\mathbf{u_w}$ represents a weight vector randomly initialized, $\mathbf{a_t}$ represents the similarity between$\mathbf{u_t}$ and $\mathbf{u_w}$ obtained by softmax function.

$$\mathbf{s_t} = \sum_{\mathbf{t}} \mathbf{a_t h_t} \tag{7}$$

By weighting and summing the output of the BiLSTM part, the attention vector, denoted as $\mathbf{s_t}$, can be obtained and fed into FC layer, preceding to the softmax layer which finally yields the probability of each sleep stage.

Construction of Baseline Networks

Apart from the proposed AT-BiLSTM network, we also constructed five baseline networks, including three single networks, i.e., CNN, LSTM and BiLSTM, and two hybrid networks, i.e., CNN-LSTM and CNN-BiLSTM.

Single Networks

Figure 2A illustrated the CNN topology used in this study, which is fed with a matrix reconstructed from a raw 30-s EEG epoch according to Equation 1. It consists of three convolution blocks and three max pooling layers. Each convolutional block contains a one-dimensional convolutional layer and a rectified linear unit (ReLU) activation layer. The input matrix is padded with zeros to ensure that the number of rows in the matrix is constant during the convolutional process. The output of CNN is fed into a FC layer, then activated by softmax function to obtain the sleep stage probability.

Two scenarios were considered in single RNN network. In the first scenario, three layers of LSTMs were stacked, also followed by a FC layer and a softmax layer. The second scenario employed stacked BiLSTM layers instead of the LSTM layers.

Hybrid Networks With CNN and RNN

As shown in **Figure 2B,C**, a CNN part followed by an RNN part was adopted in the hybrid networks, in order to make use of RNN for further processing the features extracted by CNN. The structures of the CNN part and RNN part are the same with the single networks aforementioned.

Datasets Splitting Strategy

Machine learning algorithms require independent training and test sets for model training and performance evaluation. Also, k-fold cross validation is preferred in application. Generally, there are two types of training data partitioning for clinic data: subject-wise and epoch-wise (**Figure 3**). For the subject-wise method, all the subjects were split into k folds equally and onefold is taken as the test set in turn while the remains as the training set. For the epoch-wise method, all the 30-s EEG epochs from all the subjects were merged and then split into k equal folds for each stage randomly. That is, for each sleep stage, all the 30-s EEG epochs from all the subjects were collected and divided into k folds. Consequently, the epochs of a subject may appear in both the training and test set, violating the independence between the training and test set and contributing to a virtual high performance. Thus, in the present study, the subject-wise method with fivefold cross validation was adopted. The model was trained using the training set and evaluated using the test set. Finally, all evaluation results were combined.

Experimental Setting and Network Optimization

Using the first fold as the test set, the network parameters, such as the input dimension, the number of hidden units in each LSTM/BiLSTM/convolutional layer, and the filter/stride size of each convolutional layer and pooling layer, were determined by a grid-search to minimize the errors of networks with Python 3.6 and TensorFlow v1.15.0 (Abadi et al., 2016). The standard cross-entropy loss was used as the loss function in model training due to its good performance in measuring the errors of networks with

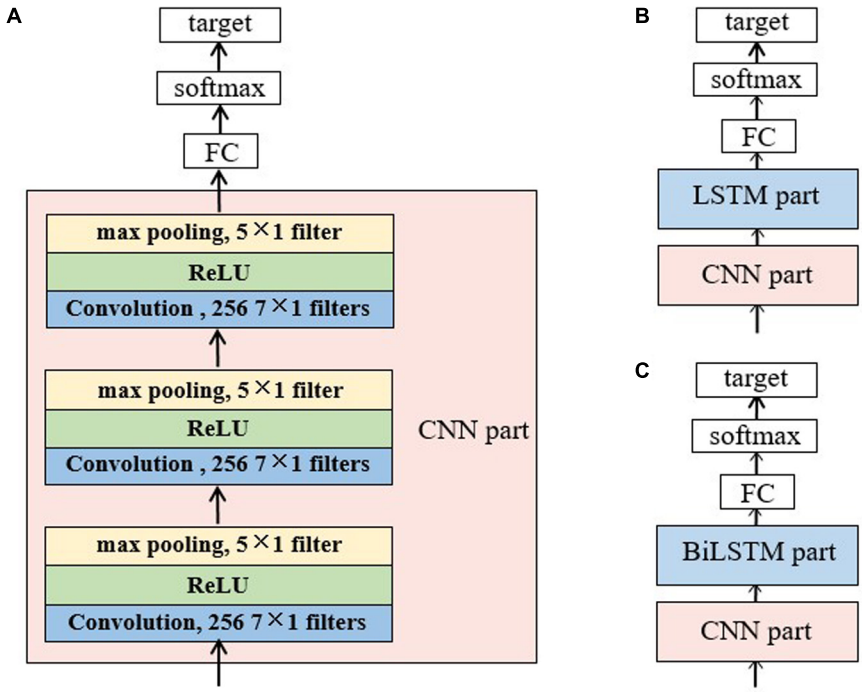


FIGURE 2 | Structure of the baseline networks for sleep staging: **(A)** the CNN network, **(B)** the CNN-LSTM network and **(C)** CNN-BiLSTM network. The CNN network consists of a CNN part, a full-connected layer and a softmax layer. In the CNN part, there are three convolution layers and three max pooling layers. Each convolution layer has 256 filters with a size of 7 × 1 each and each pooling layer has one filter of size 5 × 1. A rectified linear unit (ReLU) follows the convolution layer and precedes the pooling layer. The CNN part in panels **(B,C)** has the same topology with panel (A). For the LSTM/BiLSTM part, there are three stacked LSTM/BiLSTM layers with each layer consists of 256 memory cells. The target for all the networks was the probability of each sleep stage.

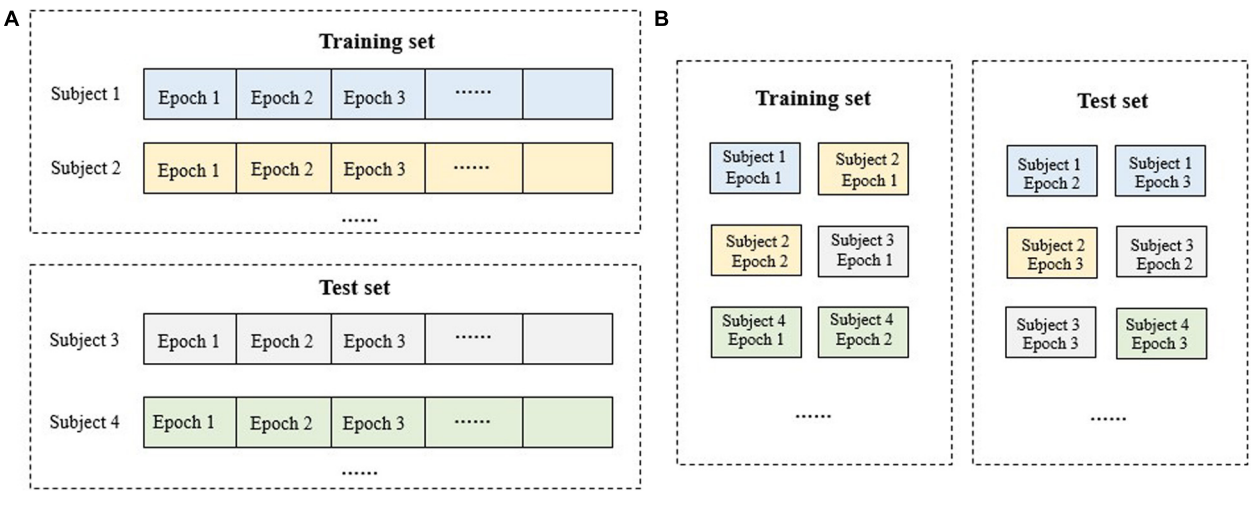


FIGURE 3 | Schematic diagram for the dataset splitting of training and test set: **(A)** subject-wise method; **(B)** epoch-wise method. For the subject-wise method, all the 30-s EEG epochs from a subject will be included in the training set or the test set as a whole while for the epoch-wise method, the epochs of a subject may appear in both the training and test set.

discrete targets (Boer et al., 2005). Each network was trained for 30 epochs with a mini batch size of 64 sequences. As a result, the input dimension *m* was set as 5, the number of hidden units as 256, and the stride size for both convolution layers and max pooling layers as 1 × 1. The filter size of each convolutional layer and max pooling layer in CNN were set to 1 × 7 and 1 × 5 respectively.

For backpropagation, the adaptive moment estimation (ADAM) algorithm was adopted because it solves the optimization problem in non-stationary conditions and

works faster than the standard gradient descent algorithm and the root mean square propagation (Kingma and Ba, 2017). The main hyper-parameters used for ADAM algorithm were set as: learning rate (α = 0.001), gradient decay factor ($\beta 1$ = 0.9), squared gradient decay factor ($\beta 2$ = 0.999), and epsilon (ε = 10–8) for numerical stability. Moreover, a dropout layer before the last FC layer was used to avoid over-fitting and its dropout rate was set to 0.2, leading to 20% of the weights dropped during the training phase.

Performance Metrics

Overall metrics, including accuracy, macro F1-score (MF1) and Cohen's kappa (κ) were used to evaluate the performance of each model. Performance on individual sleep stages was also assessed *via* class-wise precision and sensitivity.

Cohen's kappa coefficient is a statistical measure of inter-rater agreement for categorical items (Cohen, 1960). When two persons (algorithms or raters) try to evaluate the same data, Cohen's Kappa coefficient, κ, is used as a measure of agreement between their decisions. In this study, it measures the amount of agreement between the output of the proposed algorithm and the provided labels of sleep stages.

Another metric used for performance evaluation here is the area under the receiver operating characteristics (ROC) curve, called AUC. The ROC curve is a graphical tool and demonstrates the classification performance by plotting the true positive rate (TPR) against the false positive rate (FPR) at different classification thresholds (Zweig and Campbell, 1993). Furthermore, it provides a convenient way for selecting the threshold that provides the maximum classification TPR while not exceeding a maximum allowable FPR level (Kim et al., 2019). For an *n*- class classification task, *n* ROC curves can be obtained by splitting the task into *n* binary classification tasks. For each binary classification task, its AUC value can be used as a class-wise measure of performance and the macro-average AUC of these tasks can be regarded as an overall metric for the performance evaluation.

RESULTS

Table 3 shows the overall performance of different networks on the PSEE dataset. The proposed AT-BiLSTM network outperforms other networks with overall accuracy, κ, MF1 and MAUC of 83.78%, 0.766, 82.14% and 97.45% on channel Fpz-Cz, respectively and an overall accuracy, κ, MF1 and MAUC of 80.79%, 0.731, 79.27% and 96.33% on channel Pz-Oz, respectively. The AT -BiLSTM network performs better than the other networks overall. For the single networks, the RNN-based networks outperform the CNN network while the results of BiLSTM and LSTM are comparable. The hybrid networks further improve the overall performance compared to the single models. Moreover, AT-BiLSTM achieves better precision and sensitivity on N3 and REM than the hybrid networks with CNN and RNN, although they have a comparable performance on stages Wake, N1 and N2. Furthermore, better performance is found in Fpz-Cz than Pz-Oz channel, regardless of the network topology used, indicating EEG derived from the frontal lobe is more valuable than those from the parietal lobe in sleep staging.

Table 4 shows the performance of different networks on the DRM-SUB dataset. The AT -BiLSTM network still outperforms other networks, suggesting its good generalization in sleep staging. Consistent with the results in PSEE dataset, the frontal EEG channel (Fp1-A1 here) achieves the best performance. The results are in line with a recent work, which found that EEG signals from an Fp1-A1 channel yielded higher accuracy values in automatic sleep staging than those of a Cz-A1 or O1-A1 channel (Ghimatgar et al., 2019).

Figure 4 illustrates the hypnograms labeled manually by a clinical technician of sleep and by the trained AT-BiLSTM model. The corresponding EEG recoding was obtained from the first person in PSEE dataset (SC4001E0), who spent 7 h during sleep. Noting that the subject is located in the test set for the trained model. The accuracy of the automatic sleep staging for this subject is 87.30%, showing considerable reliability of the proposed AT-BiLSTM network. Most of the wrong classifications were made during the transitions from one stage to another.

Table 5 shows the class-wise performances obtained on the PSEE dataset. For most stages, better performance is achieved by the AT-BiLSTM model than the baseline networks and Fpz-Cz channel outperforms the Pz-Oz one. Although the classification accuracy of stage N1 is significantly lower than that of the other stages, which might due to the small percentage of N1 during sleep, it is higher than those reported in previous studies (Hsu et al., 2013; Supratak et al., 2017). Similar findings can be found on the DRM-SUB (**Table 6**).

Furthermore, ROC curves were used to compare the performances of the proposed AT-BiLSTM model

TABLE 3 | The overall performance of different networks on the PSEE dataset (value in bold represents for the best among all the networks).

Networks	Fpz-Cz				Pz-Oz			
	Acc.	κ	MF1	AUC	Acc.	κ	MF1	AUC
AT-BiLSTM	**83.78**	**0.766**	**82.14**	**96.08**	**80.79**	**0.731**	**79.27**	**93.63**
CNN	78.84	0.706	76.10	92.89	76.45	0.669	74.56	89.91
LSTM	81.59	0.747	79.25	95.36	79.02	0.706	75.92	92.14
BiLSTM	81.48	0.740	80.13	93.78	78.95	0.707	77.44	91.81
CNN-LSTM	82.58	0.759	80.40	93.96	79.51	0.718	76.44	92.36
CNN-BiLSTM	82.58	0.759	81.15	94.67	79.37	0.710	77.92	92.70

TABLE 4 | The overall performance of different networks on the DRM-SUB dataset (value in bold represents for the best among all the networks).

Networks	Fp1-A1				Cz-A1				O1-A1			
	Acc.	κ	MF1	AUC	Acc.	κ	MF1	AUC	Acc.	κ	MF1	AUC
AT-BiLSTM	**81.72**	**0.751**	**80.74**	**94.99**	**81.62**	**0.749**	**80.76**	**95.25**	**77.09**	**0.685**	**75.98**	**94.91**
CNN	77.84	0.732	67.17	89.90	75.82	0.664	73.98	91.00	72.45	0.617	70.84	90.59
LSTM	80.19	0.738	70.96	94.29	80.53	0.733	80.23	94.65	74.13	0.641	72.32	92.97
BiLSTM	80.31	0.739	70.41	93.84	80.41	0.733	79.58	94.66	74.22	0.644	72.98	92.85
CNN-LSTM	80.55	0.738	71.96	94.43	80.87	0.736	79.24	94.50	75.78	0.665	74.52	93.73
CNN-BiLSTM	80.61	0.737	71.62	93.87	80.83	0.738	80.71	94.68	75.94	0.666	74.42	93.79

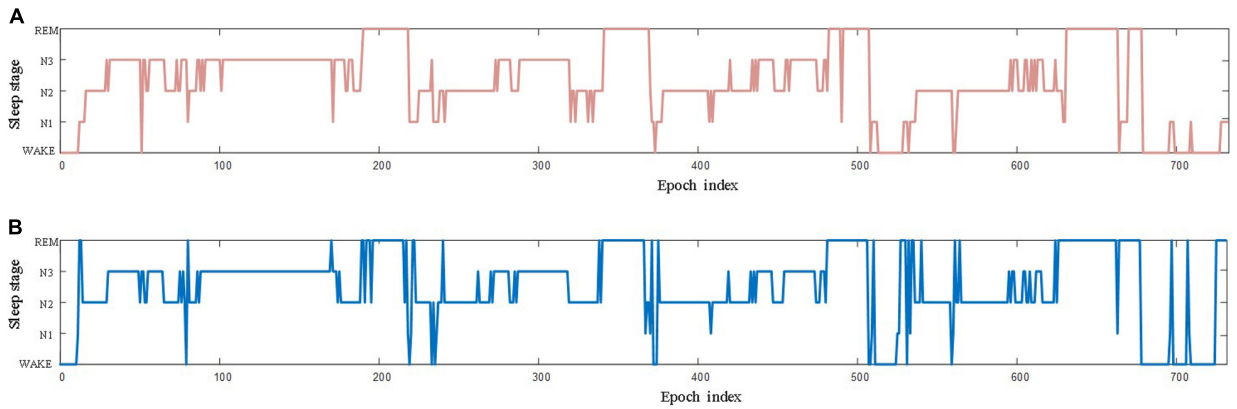


FIGURE 4 | The hypnograms labeled **(A)** manually by a clinical technician of sleep and **(B)** by the trained AT-BiLSTM model. The corresponding EEG recoding was obtained from PSEE dataset (SC4001E0).

TABLE 5 | The class-wise performance obtained on the PSEE dataset (value in bold represents for the best among all the networks).

EEG signal	Networks	Precision					Sensitivity					Class-wise AUC				
		W	N1	N2	N3	REM	W	N1	N2	N3	REM	W	N1	N2	N3	REM
Fpz-Cz	AT-BiLSTM	**86.38**	**45.06**	87.82	**88.88**	**76.84**	89.42	**25.76**	**89.18**	**89.62**	**82.18**	97.58	**88.85**	**97.31**	**99.34**	**97.33**
	CNN	84.68	34.32	84.78	82.66	67.08	83.88	19.34	84.98	82.63	77.22	95.98	82.00	94.18	97.99	94.32
	LSTM	84.72	43.28	85.98	88.68	71.92	87.74	20.68	88.16	85.12	80.14	**98.72**	84.85	97.25	99.08	96.94
	BiLSTM	81.34	41.23	86.95	86.71	69.90	88.32	20.84	86.88	89.42	81.43	96.70	81.18	96.54	99.04	95.45
	CNN-LSTM	85.72	45.16	**88.46**	87.22	71.04	**90.14**	12.46	86.54	88.46	79.22	96.52	81.29	96.77	99.17	96.09
	CNN-BiLSTM	85.16	42.51	87.56	87.42	75.72	88.72	25.34	87.42	88.64	79.76	97.19	83.26	96.96	99.22	96.74
Pz-Oz	AT-BiLSTM	**82.58**	40.24	**84.64**	**84.84**	**71.76**	**82.58**	40.24	**84.64**	**84.84**	**71.76**	**96.23**	**82.54**	**96.08**	**98.76**	**94.52**
	CNN	78.48	24.18	81.16	79.42	63.28	78.48	24.18	81.16	79.42	63.28	94.59	75.96	92.70	95.17	91.14
	LSTM	79.84	41.82	82.94	82.36	64.82	79.84	41.82	82.94	82.36	64.81	95.34	78.34	94.80	98.35	93.88
	BiLSTM	80.64	42.94	83.78	82.26	66.70	80.64	42.94	84.28	82.26	66.74	95.19	76.55	95.16	98.46	93.69
	CNN-LSTM	80.95	42.65	83.95	82.55	69.75	80.95	42.65	83.95	82.55	69.75	95.31	78.72	95.26	98.48	94.04
	CNN-BiLSTM	79.52	**44.62**	84.26	83.04	70.37	79.55	**44.62**	84.26	83.04	70.38	95.79	80.53	95.16	98.39	93.63

for different sleep stages with the frontal channels in both datasets (**Figure 5**). As shown in **Figure 5**, AT-BiLSTM is sufficient to identify WAKE, N3 and REM, but insufficient to identify N1.

Table 7 illustrates the results of a comparison between the proposed AT-BiLSTM model and the state-of-the-art works using the same dataset of DRM-SUB (Hassan and Bhuiyan, 2016a,b; Hassan and Subasi, 2017; Ghimatgar et al., 2019; Shen et al., 2019). With the same dataset, same EEG channel and same dataset splitting strategy, the proposed AT-BiLSTM model achieves the highest accuracy.

DISCUSSION

In this study, we proposed an AT-BiLSTM network for automatic sleep staging with single-channel EEG. The main findings were: (1) the frontal EEG derivations contribute to better

TABLE 6 | The class-wise performance obtained on the DRM-SUB dataset (value in bold represents for the best among all the networks).

EEG signal	Networks	Precision					Sensitivity					Class-wise AUC				
		W	N1	N2	N3	REM	W	N1	N2	N3	REM	W	N1	N2	N3	REM
Fp1-A1	AT-BiLSTM	**88.48**	45.92	**84.98**	**89.08**	**68.26**	89.34	**25.88**	**85.12**	85.18	**83.06**	**99.50**	**88.80**	**93.05**	**96.69**	**96.93**
	CNN	83.54	40.82	79.76	87.14	63.02	84.22	11.54	85.92	78.52	78.96	97.85	82.77	87.05	86.85	94.96
	LSTM	83.63	44.16	82.58	88.26	68.14	89.92	19.18	84.82	84.74	75.28	99.35	87.06	92.33	96.40	96.33
	BiLSTM	85.24	43.82	83.28	86.82	66.28	86.78	17.36	85.31	86.83	77.74	98.90	85.83	92.01	96.39	96.07
	CNN-LSTM	87.12	42.26	84.52	86.26	66.44	87.04	23.34	84.52	**87.66**	79.91	99.26	88.22	91.94	96.61	96.13
	CNN-BiLSTM	86.62	**46.92**	83.88	86.74	66.38	**89.98**	20.74	84.54	85.88	79.86	99.21	86.30	92.04	96.28	95.54
Cz-A1	AT-BiLSTM	**88.02**	42.02	**84.98**	87.56	69.22	90.96	22.22	86.32	**86.28**	**80.92**	**99.40**	**89.24**	**93.62**	**97.10**	**96.87**
	CNN	83.68	23.44	75.18	86.14	65.12	86.55	8.92	85.36	73.28	72.68	98.18	82.81	88.39	90.85	94.77
	LSTM	87.24	43.88	83.44	87.76	69.26	89.54	22.34	86.12	86.26	77.04	99.14	88.59	92.39	97.06	96.09
	BiLSTM	86.46	39.68	83.82	88.58	67.12	89.24	20.62	85.36	83.38	78.42	99.27	88.56	92.57	96.77	96.14
	CNN-LSTM	84.34	**48.72**	81.44	**89.58**	70.64	**91.34**	14.16	87.66	82.46	78.06	98.99	88.43	92.42	96.49	96.15
	CNN-BiLSTM	86.22	42.96	82.32	87.98	**70.88**	90.36	**23.68**	**86.78**	84.26	76.58	99.30	88.68	92.48	96.66	96.29
O1-A1	AT-BiLSTM	88.86	**46.58**	**78.76**	83.20	**62.56**	**91.36**	**18.53**	81.94	78.67	**72.74**	**99.55**	**89.76**	**92.40**	**96.67**	**96.18**
	CNN	86.21	38.78	73.48	82.92	52.02	84.38	7.56	80.62	72.40	65.38	98.25	82.82	87.94	92.30	91.63
	LSTM	83.66	28.76	74.96	**83.98**	55.44	90.44	8.12	81.88	75.96	62.66	99.17	84.52	91.31	96.09	93.75
	BiLSTM	89.92	39.72	75.06	80.52	56.02	87.72	13.14	79.64	81.02	62.96	99.24	84.52	91.28	95.89	93.33
	CNN-LSTM	**90.52**	42.81	76.32	82.25	59.18	88.76	13.56	**82.34**	78.86	68.73	99.23	87.46	91.39	95.13	95.46
	CNN-BiLSTM	88.04	45.96	76.26	82.44	60.84	89.92	11.72	82.18	**79.62**	67.68	99.26	86.43	91.41	96.40	95.44

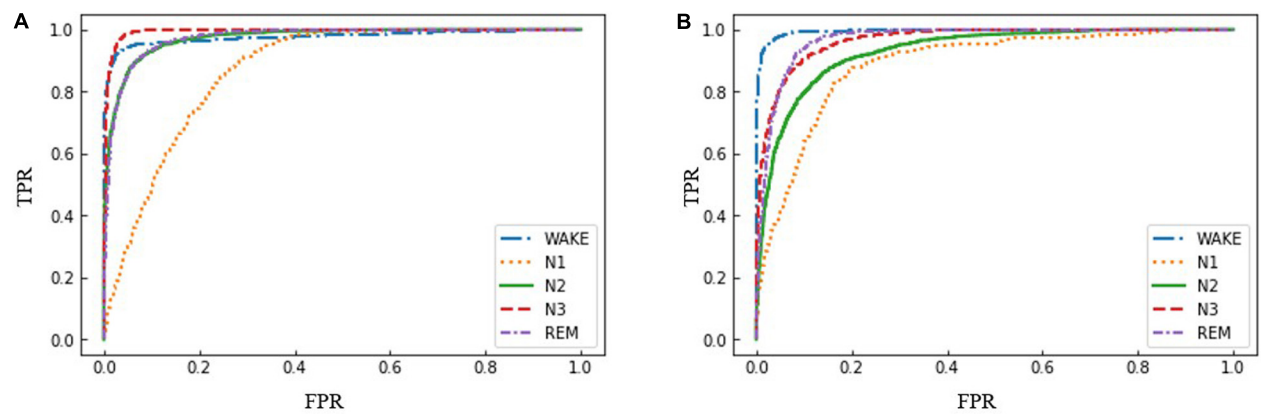


FIGURE 5 | ROC curves for sleep stages using the proposed AT-BiLSTM models trained with **(A)** Fpz-Cz channel of PSEE dataset and **(B)** Fp1-A1 channel of DRM-SUB dataset.

TABLE 7 | Comparison of sleep staging performance on the DRM-SUB dataset between the proposed method and previous works based on conventional feature extraction.

Authors	Year	Methodology	Dataset splitting strategy	Channel	Accuracy
Hassan and Bhuiyan, 2016a	2016	Tunable Q-factor wavelet transform, random forest (Hassan and Bhuiyan, 2016a)	Epoch-wise	Fp1-A1	72.28%
Hassan and Bhuiyan, 2016b	2016	Implementation of ensemble empirical mode decomposition in conjunction with random under sampling boosting (Hassan and Bhuiyan, 2016a)	Epoch-wise	Fp1-A1	74.59%
Hassan and Subasi, 2017	2017	Tunable Q-factor wavelet transform, bagging (Hassan and Subasi, 2017)	Epoch-wise	Fp1-A1	78.95%
Shen et al.	2019	Essence features extraction method (Shen et al., 2019)	Subject-wise	Cz-A1	80.90%
Ghimatgar et al.	2019	Features in time domain, frequency domain, cepstral domain, wavelet features, autoregressive coefficients and non-linear features with Hidden Markov Model (Ghimatgar et al., 2019)	Subject-wise	Fp1-A1	81.22%
Proposed method		Raw EEG signal and AT-BiLSTM	Subject-wise	Fp1-A1	81.72%

performance of sleep staging than those located in the central, occipital or parietal lobe; (2) the proposed AT-BiLSTM network outperforms the other networks based on CNN or RNN; (3) The proposed deep learning network achieves higher accuracy than conventional feature extraction methods.

Two EEG datasets, i.e., PSEE and DRM-SUB, with different EEG derivations were used in our study. To clarify the influence of the EEG channel on automatic sleep staging, here we applied the proposed method to all the EEG channels in both datasets. The results obtained from both datasets are similar: the model adopting frontal derivation behaved better than those from other lobes. Such a finding indicated that the performance of sleep scoring was sensitive to the selection of EEG channel and the derivations from the frontal region are the optimal choices. Physiologically, the prefrontal cortex is deactivated and reactivated during the sleep cycle, indicating its involvement in the wake–sleep cycle (Maquet et al., 1996). With the development of wearable EEG devices, EEG signals can be easily obtained using dry electrodes on the forehead (Hassan and Bhuiyan, 2016a); the proposed method would be promising in supporting people monitoring sleep.

In recent years, many automated sleep staging methods based on deep neural networks used CNNs for feature extraction and RNNs to capture temporal information. These approaches have significantly improved the accuracy of sleep staging (Hassan and Bhuiyan, 2016a; Boostani et al., 2017; Sors et al., 2018). In general, for the sequence-to-label model based on RNN, only the output vector at the last time step is retained for classification, e.g., *via* a softmax layer (Phan et al., 2017). However, it is reasonable to combine the output vectors of different time steps by some weighting schemes. Intuitively, those parts of the input sequence which are essential to the classification task at hand should be associated with strong weights, and those with less importance should be weighted correspondingly less. Ideally, these weights should be automatically learned by the network. This can be accomplished with an attention layer (Luong et al., 2015). Besides, previous works demonstrated that the performance of classification or regression can be further improved by stacking multiple BiLSTM in neural networks (Liu et al., 2017; Wang et al., 2018; Liu et al., 2018). Aside from that, we found the overall performance of the RNN based model to be better than that of the CNN models in automatic sleep staging, which might indicate that the RNNs are promising in capturing the temporal nature of an EEG time series. From such a perspective, the highest performance achieved by the proposed AT-BiLSTM might further confirm the role of stacking layers and attention mechanism in feature extracting of time series.

In this study, all experiments were performed on a server configured with 64 CPUs [Intel(R) Xeon(R) CPU @ 2.10 GHz), 64 GB memory, a GPU (NVIDIA GeForce GTX 1,080 Ti] and a Windows Server 2016 system. A CNN network has the lowest computational cost as its training time for each batch was 0.16 s on average. LSTM and CNN-LSTM networks take similar times (8.46 and 8.60 s respectively) for each batch in training. The computational cost of BiLSTM based networks is twice that of LSTM based networks because they must calculate the input sequence in two directions and set up double parameters. Moreover, approximately 1.3 s more is required for each batch with the attention layer.

Our study demonstrated that a deep learning approach without manual feature extraction can also provide sufficient accuracy for sleep staging, which is even better than conventional methods based on manual feature extraction. Therefore, the proposed method is a promising choice for computer-aided detection of sleep stages and similar 1-D signal classification problems. In conclusion, our findings provide a possible solution for automatic sleep scoring without manual signal preprocessing and feature extraction. With the development of wearable EEG devices, such a solution would be valuable in the screening of sleep disorders at home for the general population.

DATA AVAILABILITY STATEMENT

The datasets presented in this study can be found in online repositories. The names of the repository/repositories and accession number(s) can be found below: Dreams Subjects: https://zenodo.org/record/2650142#.X6tbymgzZdg. Sleep-EDF Database Expanded: https://physionet.org/content/sleep-edfx/1.0.0/.

ETHICS STATEMENT

The studies involving human participants were reviewed and approved by the institutional review board of two open-access databsets, i.e., the Sleep-EDF Expanded dataset available at Physionet and the DREAMS Subjects dataset. The patients/participants provided their written informed consent to participate in this study.

AUTHOR CONTRIBUTIONS

FH, XL, FX, and JL designed this study. MF and YW analyzed the data. MF, FH, and ZC wrote the article. All authors contributed to the article and approved the submitted version.

FUNDING

This research was funded by the National Natural Science Foundation of China (Grant No. 11974231) and the Double First-Class University Project of China.

REFERENCES

Abadi, M., Agarwal, A., Barham, P., Brevdo, E., Chen, Z., Citro, C., et al. (2016). Tensorflow: large-scale machine learning on heterogeneous distributed systems. *arXiv* [Preprint]. Available online at: https://arxiv.org/abs/1603.04467 (accessed February 15, 2020).

Aboalayon, K. A., Ocbagabir, H. T., and Faezipour, M. (2014). Efficient sleep stage classification based on EEG signals. *IEEE LISAT* 2014, 978–983. doi: 10.1109/LISAT.2014.6845193

Boer, P., Kroese, D. P., Mannor, S., and Rubinstein, R. Y. (2005). A tutorial on the cross-entropy method. *Ann. Operat. Res.* 134, 19–67. doi: 10.1007/s10479-005-5724-z

Boostani, R., Karimzadeh, F., and Nami, M. J. (2017). A comparative review on sleep stage classification methods in patients and healthy individuals. *Comput. Methods Prog. Biomed.* 140, 77–91. doi: 10.1016/j.cmpb.2016.12.004

Bresch, E., Groekathfer, U., and Garcia-Molina, G. (2018). Recurrent deep neural networks for real-time sleep stage classification from single channel EEG. *Front. Comput. Neurosci.* 12:85. doi: 10.3389/fncom.2018.00085

Chambon, S., Galtier, M., Arnal, P. J., and Wainrib, G. (2018). A deep learning architecture for temporal sleep stage classification using multivariate and multimodal time series. *IEEE Trans. Neural Syst. Rehabil. Eng.* 26, 758–769. doi: 10.1109/TNSRE.2018.2813138

Cohen, J. A. (1960). A coefficient of agreement for nominal scales. *Educ. Psychol. Meas.* 20, 37–46. doi: 10.1177/001316446002000104

Cui, Z. H., Zheng, X. W., Shao, X. X., and Cui, L. Z. (2018). Automatic sleep stage classification based on convolutional neural network and fine-grained segments. *Complexity* 2018, 13. doi: 10.1155/2018/9248410

Czeisler, C. A. (2015). Duration, timing and quality of sleep are each vital for health, performance and safety. *Sleep Health* 1, 5–8. doi: 10.1109/Trustcom.2015.524

Devuyst, S. (2005). *The DREAMS Databases and Assessment Algorithm*. Genève: Zenodo. doi: 10.5281/zenodo.2650142

Dong, H., Supratak, A., Pan, W., Wu, C., Matthews, P. M., and Guo, Y. (2016). Mixed neural network approach for temporal sleep stage classification. *IEEE Trans. Neural Syst. Rehabil. Eng.* 26, 324–333. doi: 10.1109/TNSRE.2017.2733220

Elman, J. (1990). Finding structure in time. *Trends Cogn. Sci.* 14, 179–211. doi: 10.1016/0364-0213(90)90002-E

Ferri, R., Rundo, F., Novelli, L., and Terzano, M. G. (2012). A new quantitative automatic method for the measurement of non-rapid eye movement sleep electroencephalographic amplitude variability. *J. Sleep Res.* 21, 212–220. doi: 10.1111/j.1365-2869.2011.00981.x

Fiorillo, L., Puiatti, A., Papandrea, M., and Ratti, P. L. (2019). Automated sleep scoring: a review of the latest approaches. *Sleep Med. Rev.* 48, 101204–101204. doi: 10.1016/j.smrv.2019.07.007

Ghimatgar, H., Kazemi, K., Helfroush, M. S., and Aarabi, A. (2019). An automatic single-channel EEG-based sleep stage scoring method based on hidden Markov model. *J. Neurosci. Methods* 324:108320. doi: 10.1016/j.jneumeth.2019.108320

Goldberger, A. L., Amaral, L. A., Glass, L., Hausdorff, J. M., Ivanov, P. C., Mark, R. G., et al. (2000). PhysioBank, PhysioToolkit, and PhysioNet: components of a new research resource for complex physiologic signals. *Circulation* 101, E215–E220. doi: 10.1161/01.cir.101.23.e215

Hassan, A. R., and Bhuiyan, M. I. (2016a). A decision support system for automatic sleep staging from EEG signals using tunable Q-factor wavelet transform and spectral features. *J. Neurosci. Methods* 271, 107–118. doi: 10.1016/j.jneumeth.2016.07.012

Hassan, A. R., and Bhuiyan, M. I. (2016b). Automatic sleep scoring using statistical features in the EMD domain and ensemble methods. *Biocybern. Biomed. Eng.* 1, 248–255. doi: 10.1016/j.bbe.2015.11.001

Hassan, A. R., and Bhuiyan, M. I. (2017). Automated identification of sleep states from EEG signals by means of ensemble empirical mode decomposition and random under sampling boosting. *Comput. Methods Prog. Biomed.* 140, 201–210. doi: 10.1016/j.cmpb.2016.12.015

Hassan, A. R., and Subasi, A. (2017). A decision support system for automated identification of sleep stages from single-channel EEG signals. *Knowl. Based Syst.* 128, 115–124. doi: 10.1016/j.knosys.2017.05.005

Hochreiter, S., and Schmidhuber, J. (1997). Long short-term memory. *Neural Comput.* 9, 1735–1780. doi: 10.1162/neco.1997.9.8.1735

Hollis, T., Viscardi, A., and Yi, S. E. (2018). A comparison of LSTMs and attention mechanisms for forecasting financial time series. *arXiv [Preprint]*. Available online at: https://arxiv.org/abs/1812.07699 (accessed March 7, 2020).

Hsu, Y. L., Yane, Y. T., Wane, T. S., and Hsu, C. (2013). Automatic sleep stage recurrent neural classifier using energy features of EEG signals. *Neurocomputing* 104, 105–114. doi: 10.5555/2438096.2438127

Iber, C., Israel, S. A., Chesson, A. L., and Quan, S. F. (2007). *The AASM Manual for the Scoring of Sleep and Associated Events: Rules, Terminology and Technical Specifications*. Darien, IL: AASM.

Jo, H. G., Park, J. Y., Lee, C. K., and An, S. K. (2010). Genetic fuzzy classifier for sleep stage identification. *Comput. Biol. Med.* 40, 629–634. doi: 10.1016/j.compbiomed.2010.04.007

Karpathy, A., and Fei-Fei, L. (2014). Deep visual-semantic alignments for generating image descriptions. *arXiv [Preprint]*. Available online at: https://arxiv.org/abs/1412.2306v2 (accessed March 20, 2020).

Kemp, B., Zwinderman, A. H., Tuk, B., Kamphuisen, H. A. C., and Oberye, J. J. L. (2000). Analysis of a sleep-dependent neuronal feedback loop: The slow-wave microcontinuity of the EEG. *IEEE Trans. Biomed. Eng.* 47, 1185–1194. doi: 10.1109/10.867928

Kim, J., ElMoaqet, H., Tilbury, D. M., Ramachandran, S. K., and Penzel, T. (2019). Time domain characterization forsleep apnea in oronasal airflow signal: a dynamic threshold classification approach. *Physiol. Meas.* 40:5. doi: 10.1088/1361-6579/aaf4a9

Kingma, D. P., and Ba, J. (2017). Adam: a method for stochastic optimization. *arXiv [Preprint]*. Available online at: http://arxiv.org/abs/1412.6980 (accessed March 20, 2020).

Lecun, Y., and Bengio, Y. (1997). *Convolutional Networks for Images, Speech, and Time-Series*. New York, NY: ACM.

Lecun, Y., Bengio, Y., and Hinton, G. (2015). Deep learning. *Nature* 521, 436–444. doi: 10.1038/nature14539

Liang, S. F., Kuo, C. E., Hu, Y. H., Pan, Y. H., and Wang, Y. H. (2012). Automatic stage scoring of single-channel sleep EEG by using multiscale entropy and autoregressive models. *IEEE Trans. Instrum. Meas.* 61, 1649–1657. doi: 10.1109/TIM.2012.2187242

Liu, T., Yu, S., Xu, B., and Yin, H. (2018). Recurrent networks with attention and convolutional networks for sentence representation and classification. *Appl. Intell.* 48, 3797–3806. doi: 10.1007/s10489-018-1176-4

Liu, Z., Yang, M., and Wang, X. (2017). Entity recognition from clinical texts via recurrent neural network. *BMC Med. Inform. Decis.* 17:67. doi: 10.1186/s12911-017-0468-7

Lngkvist, M., and Loutfi, A. (2018). A deep learning approach with an attention mechanism for automatic sleep stage classification. *arXiv [Preprint]*. Available online at: https://arxiv.org/abs/1805.05036 (accessed March 20, 2020).

Luong, M. T., Pham, H., and Manning, C. D. (2015). Effective approaches to attention-based neural machine translation. *arXiv [Preprint]*. Available online at: https://arxiv.org/abs/1508.04025 (accessed March 20, 2020).

Malafeev, A., Laptev, D., Bauer, S., Omlin, X., Wierzbicka, A., Wichniak, A., et al. (2018). Automatic human sleep stage scoring using deep neural networks. *Front. Neurosci.* 12:781. doi: 10.3389/fnins.2018.00781

Maquet, P., Péters, J.-M., Aerts, J. L., Delfiore, G., Egueldre, C. D., and Luxen, A. (1996). Functional neuroanatomy of human rapid-eye-movement sleep and dreaming. *Nature* 383, 163–166. doi: 10.1038/383163a0

Ohayon, M. M. (2002). Epidemiology of insomnia: what we know and what we still need to learn. *Sleep Med. Rev.* 6, 97–111. doi: 10.1053/smrv.2002.0186

Phan, H., Andreotti, F., Cooray, N., Chén, O. Y., and Vos, M. D. (2019). Joint classification and prediction CNN framework for automatic sleep stage classification. *IEEE Trans. Bio Med. Eng.* 66, 1285–1296. doi: 10.1109/TBME.2018.2872652

Phan, H., Koch, P., Katzberg, F., Maass, M., Mazur, R., and Mertins, A. (2017). Audio scene classification with deep recurrent neural networks. *arXiv [preprint]*. Available online at: https://arxiv.org/abs/1703.04770 (accessed April 5, 2020).

Radha, M., Garcia-Molina, G., Poel, M., and Tononi, G. (2014). Comparison of feature and classifier algorithms for online automatic sleep staging based on a single EEG signal. *IEEE EMB* 2014, 1876–1880. doi: 10.1109/EMBC.2014.6943976

Ronzhina, M., Janoušek, O., Kolářová, J., and Nováková, M. (2012). Sleep scoring using artificial neural networks. *Sleep Med. Rev.* 16, 251–263. doi: 10.1016/j.smrv.2011.06.003

Schuster, M., and Paliwal, K. K. (1997). Bidirectional recurrent neural networks. *IEEE Trans. Signal. Process.* 45, 2673–2681. doi: 10.1109/78.650093

Seifpour, S., Niknazar, H., Mikaeili, M., and Nasrabadi, A. M. (2018). A new automatic sleep staging system based on statistical behavior of local extrema using single channel EEG signal. *Expert Syst. Appl.* 104, 277–293. doi: 10.1016/j.eswa.2018.03.020

Shen, H. M., Xu, M. H., Guez, A., and Li, A. (2019). An accurate sleep stages classification method based on state space model. *IEEE Access.* 4, 1–12. doi: 10.1109/ACCESS.2019.2939038

Silveira, T. L., Kozakevicius, A. J., and Rodrigues, C. R. (2017). Single-channel EEG sleep stage classification based on a streamlined set of statistical features in wavelet domain. *Med. Biol. Eng. Comput.* 55, 343–352. doi: 10.1007/s11517-016-1519-4

Sors, A., Bonnet, S., Mirek, S., and Vercueil, L. (2018). A convolutional neural network for sleep stage scoring from raw single-channel EEG. *Biomed. Signal. Proces.* 42, 107–114. doi: 10.1016/j.bspc.2017.12.001

Stephansen, J. B., Olesen, A. N., Olsen, M., Ambati, A., Leary, E. B., Moore, H. E., et al. (2018). Neural network analysis of sleep stages enables efficient diagnosis of narcolepsy. *Nat. Commun.* 9:1. doi: 10.1038/s41467-018-07229-3

Supratak, A., Dong, H., Wu, C., and Guo, Y. (2017). "DeepSleepNet: a model for automatic sleep stage scoring based on raw single-channel EEG," in *IEEE Transactions on Neural Systems and Rehabilitation Engineering*, (Piscataway, NJ: IEEE), 99. doi: 10.1109/TNSRE.2017.2721116

Tian, P., Hu, J., Qi, J., Ye, X., Che, D., Ding, Y., et al. (2017). A hierarchical classification method for automatic sleep scoring using multiscale entropy features and proportion information of sleep architecture. *Biocybern. Biomed. Eng.* 37, 263–271. doi: 10.1016/j.bbe.2017.01.005

Tripathy, R., and Acharya, U. R. (2018). Use of features from RR-time series and EEG signals for automated classification of sleep stages in deep neural network framework. *Biocybern. Biomed. Eng.* 38, 890–902. doi: 10.1016/j.bbe.2018.05.005

Tsinalis, O., Matthews, P. M., and Guo, Y. (2015). Automatic sleep stage scoring using time-frequency analysis and stacked sparse autoencoders. *Ann. Biomed. Eng.* 44, 1587–1597. doi: 10.1007/s10439-015-1444-y

Tsinalis, O., Matthews, P. M., Guo, Y., and Zafeiriou, S. (2016). Automatic sleep stage scoring with single-channel EEG using convolutional neural networks. *arXiv [Preprint]*. Available online at: https://arxiv.org/abs/1610.01683 (accessed May 10, 2020).

Wang, C., Yang, H., and Meinel, C. (2018). Image captioning with deep bidirectional LSTMs and multi-task learning. *ACM Trans. Multim. Comput.* 14, 1–20. doi: 10.1145/3115432

Wang, Y., Loparo, K. A., Kelly, M. R., and Kaplan, R. F. (2015). Evaluation of an automated single-channel sleep staging algorithm. *Nat. Sci. Sleep* 7, 101–111. doi: 10.2147/NSS.S77888

Weaver, E. M., Woodson, B. T., and Steward, D. L. (2005). Polysomnography indexes are discordant with quality of life, symptoms, and reaction times in sleep apnea patients. *Otolaryngol. Head Neck Surg.* 132, 255–262. doi: 10.1016/j.otohns.2004.11.001

Yildirim, O., Baloglu, U. B., and Acharya, U. R. (2019). A deep learning model for automated sleep stages classification using PSG signals. *Int. J. Environ. Res. Public Health* 16, 599–599. doi: 10.3390/ijerph16040599

Younes, M. (2017). The case for using digital EEG analysis in clinical sleep medicine. *Sleep Sci. Prac.* 1:2. doi: 10.1186/s41606-016-0005-0

Zhang, J., and Wu, Y. (2017). A new method for automatic sleep stage classification. *IEEE Trans. Biomed. Circuits Syst.* 5, 1097–1110. doi: 10.1109/TBCAS.2017.2719631

Zhang, J., and Wu, Y. (2018). Complex-valued unsupervised convolutional neural networks for sleep stage classification. *Comput. Methods Prog. Biomed.* 164, 181–191. doi: 10.1016/j.cmpb.2018.07.015

Zhang, X., Liang, X., Zhiyuli, A., and Zhang, S. (2019). AT-LSTM: an attention-based LSTM model for financial time series prediction. *IOP Conf. Ser. Mater. Sci. Eng.* 569:052037. doi: 10.1088/1757-899X/569/5/052037

Zhu, G., Li, Y., and Wen, P. P. (2014). Analysis and classification of sleep stages based on difference visibility graphs from a single-channel EEG signal. *IEEE J. Biomed. Health* 18, 1813–1821. doi: 10.1109/JBHI.2014.2303991

Zweig, M. H., and Campbell, G. (1993). Receiver-operating characteristic (ROC) plots: a fundamental evaluation tool in clinical medicine. *Clin. Chem.* 39, 561–577. doi: 10.1093/clinchem/39.4.561

Conflict of Interest: The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

frontiers in Physiology

ORIGINAL RESEARCH
published: 08 March 2021
doi: 10.3389/fphys.2021.616985

3D-FV-FE Aeroacoustic Larynx Model for Investigation of Functional Based Voice Disorders

Sebastian Falk[1], Stefan Kniesburges[1], Stefan Schoder[2], Bernhard Jakubaß[1], Paul Maurerlehner[2], Matthias Echternach[3], Manfred Kaltenbacher[2] and Michael Döllinger[1]*

[1] Division of Phoniatrics and Pediatric Audiology, Department of Otorhinolaryngology, Head & Neck Surgery, Friedrich-Alexander-University Erlangen-Nürnberg, Erlangen, Germany, [2] Institute of Fundamentals and Theory in Electrical Engineering, Division Vibro- and Aeroacoustics, Graz University of Technology, Graz, Austria, [3] Division of Phoniatrics and Pediatric Audiology, Department of Otorhinolaryngology, Munich University Hospital (LMU), Munich, Germany

OPEN ACCESS

Edited by:
Rajat Mittal, Johns Hopkins University, United States

Reviewed by:
Byron Erath, Clarkson University, United States
Haoxiang Luo, Vanderbilt University, United States
Xudong Zheng, University of Maine, United States

***Correspondence:**
Sebastian Falk
sebastian.falk@uk-erlangen.de

Specialty section:
This article was submitted to Computational Physiology and Medicine, a section of the journal Frontiers in Physiology

Received: *13 October 2020*
Accepted: *09 February 2021*
Published: *08 March 2021*

Citation:
Falk S, Kniesburges S, Schoder S, Jakubaß B, Maurerlehner P, Echternach M, Kaltenbacher M and Döllinger M (2021) 3D-FV-FE Aeroacoustic Larynx Model for Investigation of Functional Based Voice Disorders. Front. Physiol. 12:616985. doi: 10.3389/fphys.2021.616985

For the clinical analysis of underlying mechanisms of voice disorders, we developed a numerical aeroacoustic larynx model, called *simVoice*, that mimics commonly observed functional laryngeal disorders as glottal insufficiency and vibrational left-right asymmetries. The model is a combination of the Finite Volume (FV) CFD solver Star-CCM+ and the Finite Element (FE) aeroacoustic solver CFS++. *simVoice* models turbulence using Large Eddy Simulations (LES) and the acoustic wave propagation with the perturbed convective wave equation (PCWE). Its geometry corresponds to a simplified larynx and a vocal tract model representing the vowel /a/. The oscillations of the vocal folds are externally driven. In total, 10 configurations with different degrees of functional-based disorders were simulated and analyzed. The energy transfer between the glottal airflow and the vocal folds decreases with an increasing glottal insufficiency and potentially reflects the higher effort during speech for patients being concerned. This loss of energy transfer may also have an essential influence on the quality of the sound signal as expressed by decreasing sound pressure level (SPL), Cepstral Peak Prominence (CPP), and Vocal Efficiency (VE). Asymmetry in the vocal fold oscillations also reduces the quality of the sound signal. However, *simVoice* confirmed previous clinical and experimental observations that a high level of glottal insufficiency worsens the acoustic signal quality more than oscillatory left-right asymmetry. Both symptoms in combination will further reduce the quality of the sound signal. In summary, *simVoice* allows for detailed analysis of the origins of disordered voice production and hence fosters the further understanding of laryngeal physiology, including occurring dependencies. A current walltime of 10 h/cycle is, with a prospective increase in computing power, auspicious for a future clinical use of *simVoice*.

Keywords: computational fluid dynamics, computational aero acoustics, glottal insufficiency, left-right asymmetry, posterior gap, simVoice (numerical larynx model)

1. INTRODUCTION

The human voice as a prerequisite for speech production is our most important tool to communicate with other people. Moreover, people heavily rely on oral communication in their professional life. Disorders of the ordinary communication system have severe consequences on concerned persons' employments and even on the whole economic system (Ruben, 2000). The phonatory process, the prerequisite for human speech, describes the production of the human voice and depends on various factors as age, gender, training, and health status (Titze, 2000; Aronson and Bless, 2009).

The human voice results from a periodic oscillation of the vocal folds (VF) in the larynx, see **Figure 1**. The oscillations are caused by a complex fluid-structure interaction between the tracheal airflow and the elastic tissue of the vocal folds. Thereby, the airflow is the main sound generating source, that is subsequently modulated by the vocal tract consisting of the upper airway structures and is then emitted from the lips as an audible signal.

This process is supposed to be most efficient when (1) the vocal folds close the gap in between (called glottis) completely in each oscillation cycle and (2) when they oscillate symmetrically and periodic (Titze, 2000). An incomplete glottis closure or glottal insufficiency and asymmetric oscillations of the vocal folds cause a reduced voice quality with decreased tonal and increased broadband sound in the voice signal (Park and Mongeau, 2008; Hoffman et al., 2012; Yamauchi et al., 2016). The voice is then described as aspirated/breathy and hoarse. However, as shown by Inwald et al. (2010) and Schneider and Bigenzahn (2003), these underlying symptoms do not only occur in pathologic (e.g., scars, paresis, or paralysis) cases (Bhatt and Verma, 2014), but also in apparently organically healthy larynges (Rammage et al., 1992; Inwald et al., 2010; Patel et al., 2012) and with advancing age of the patients (Södersten et al., 1995; Vaca et al., 2017).

The scientific investigation and the clinical diagnostics suffer from the restricted location of the vocal folds inside the larynx, especially during phonation. To compensate this restriction, experimental (*ex/in vivo*), and numerical models have been developed. *In vivo* studies on glottal insufficiency were done by Södersten et al. (1995), Södersten and Lindestad (1990), and Yamauchi et al. (2014) and on the asymmetric vocal fold oscillations by Eysholdt et al. (2003). Whereas *in vivo* studies are difficult to perform and are mainly restricted to pure observation of the vocal fold oscillations (Inwald et al., 2010; Döllinger et al., 2012), *ex vivo* experiments with excised cadaver larynges (e.g., canine, porcine, human) provide better access to the laryngeal area and enable to manipulate the larynx (Hoffman et al., 2012; Birk et al., 2017b). *Ex vivo* studies about different levels of glottal insufficiency were reported by Döllinger et al. (2018) and Thornton et al. (2019) using rabbit larynges and Birk et al. (2017b) who used porcine larynges. Moreover, Oren et al. (2016) investigated asymmetric vocal fold oscillations in excised canine larynges.

Besides excised larynges, synthetic vocal fold models with silicone vocal folds were carried out with the focus on the glottal insufficiency (Park and Mongeau, 2008; Kirmse et al., 2010; Kniesburges et al., 2013, 2016). Pickup and Thomson (2009) and Zhang et al. (2012) investigated asymmetric vocal fold oscillations with a silicone model. Such models can mimic specific physiological and disordered motion patterns of the vocal folds for which they have been developed for and are therefore well-established in voice science (Zhang et al., 2004; Thomson et al., 2005; Park and Mongeau, 2008; Kirmse et al., 2010; Murray and Thomson, 2012; Kniesburges et al., 2013, 2016; Van Hirtum and Pelorson, 2017; Motie-Shirazi et al., 2019; Taylor et al., 2019; Romero et al., 2020). However, both *ex vivo* and synthetic larynx models are restricted regarding the spatial resolution of the measuring data of fluid flow, the vocal fold dynamics, and their interaction.

Thus, numerical models based on Finite-Elements and/or Finite-Volumes have great potential to be applied in the clinical routine, e.g., diagnostics and treatment control. Numeric simulations, regarding the effect of the glottal insufficiency on the human voice, were done by Zörner et al. (2016) and on the asymmetric vocal fold oscillations by Xue et al. (2010) and Samlan et al. (2014). In contrast to experimental models, computer models provide the complete 3D data of the flow field

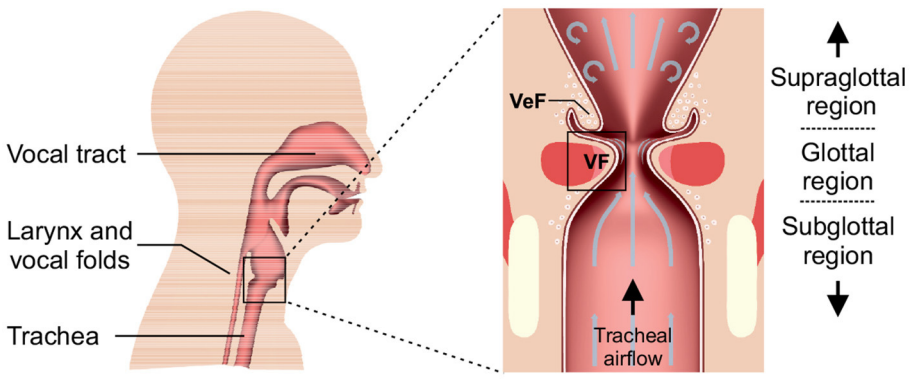


FIGURE 1 | 2D view of a human head **(left)** with an enlargement of the larynx **(right)** and its embedded structures that are important for the phonatory process. The vocal folds (VF) and the above arranged ventricular folds (VeF) are indicated.

TABLE 1 | Parameter reported for normal male phonation in *in vivo* and *ex vivo* studies compared with the experimental silicone model *synthVOICE* (Kniesburges et al., 2013, 2016, 2020; Kniesburges, 2014) (validation cases) and the performed numerical validation simulations by *simVoice* (Sadeghi et al., 2018, 2019a,b; Sadeghi, 2019; Schoder et al., 2020).

Parameter	*In vivo* (male)	*Ex vivo* (male)	Silicone model (*synthVOICE*)	Numerical simulation (*simVoice*)
Fundamental Frequency (F_0) [Hz]	103–220 (Larsson and Hertegård, 2004; Sundberg et al., 2005)	97–200 (Döllinger et al., 2005, 2016; Döllinger and Berry, 2006b)	148	148
Vocal fold length (anterior–posterior) [mm]	14–17 (Schuberth et al., 2002; Hoppe et al., 2003; Larsson and Hertegård, 2004; Rogers et al., 2014)	13–18 (Lagier et al., 2017)	15	15
Glottal gap diameter (d_G) [mm]	1.49–2.8 (Hoppe et al., 2003; George et al., 2008; Semmler et al., 2018)	2.3–5.6 (Döllinger et al., 2005, 2016; Döllinger and Berry, 2006a,b; Boessenecker et al., 2007)	4.66	4.66
Speed Quotient (*SQ*) [a.u.]	0.59–1.978 (Holmberg et al., 1988; Baken and Orlikoff, 2000)	0.8–1.6 (Döllinger et al., 2014)	0.67	0.67
Open Quotient (*OQ*) [a.u.]	0.37–1.00 (Holmberg et al., 1988; Baken and Orlikoff, 2000)	0.42–1.00 (Mendelsohn et al., 2015)	0.93	0.93
Mean flow rate ($\overline{Q}$) [$\frac{l}{min}$]	4.5–18 (Holmberg et al., 1988; Baken and Orlikoff, 2000)	6–108 (Döllinger et al., 2005, 2014, 2016; Döllinger and Berry, 2006a,b; Boessenecker et al., 2007)	65–115	37.8–132
Mean subglottal pressure (P_{sub}) [Pa]	157–3510 (Holmberg et al., 1988; Sundberg et al., 1993, 2005; Alku et al., 2006)	600–4300 (Döllinger et al., 2005, 2014, 2016; Döllinger and Berry, 2006b)	2449–3251	2450–3251

(Sciamarella and Le Quéré, 2008; Zörner et al., 2013; Sadeghi et al., 2018) and in case of coupled models the fluid-structure interaction (FSI) between flow, tissue and the aeroacoustic sound generation and propagation during phonation (de Oliveira Rosa et al., 2003; Luo et al., 2008, 2009; Tao and Jiang, 2008; Link et al., 2009; Kaltenbacher et al., 2014; Xue et al., 2014; Jo et al., 2016).

The large drawback of these numerical models are the large computational costs to perform the simulations (Sadeghi et al., 2018). Thus, they are not applicable in the clinical environment yet, where a short wall time with sufficient accuracy is needed. However, computational fluid dynamic (CFD) models with prescribed vocal fold movements and a prospective increasing computational power already keeps the simulation time adequately small (Sadeghi et al., 2019b).

For the development of our hybrid (sound propagation is calculated based on aeroacoustic source terms from the flow simulation) 3D aeroacoustic numeric larynx model *simVoice* (Sadeghi et al., 2018, 2019a,b; Schoder et al., 2020) for future clinic usefulness, it is essential to replicate normal and disordered glottal closures and dynamical asymmetries. A method to set up a workflow containing the import of various physiological and disordered glottal geometries into *simVoice* is shown in this study. We concentrate on modeling four disordered cases of glottal insufficiency based on high-speed video data of porcine *ex vivo* experiments performed by Birk et al. (2017b). Moreover, symmetric and asymmetric vocal fold motions are modeled. Our hypotheses for this study are:

- Hypothesis 1: Our existing and validated 3D-FV-FE numerical larynx model *simVoice* can accurately mimic and simulate realistic glottis geometries and vocal fold motions based on experimental high-speed video data.
- Hypothesis 2: *simVoice* can qualitatively and quantitatively mimic typical glottal parameters quantifying the different levels of glottal insufficiency that are reported in the literature.
- Hypothesis 3: Typical parameters of the acoustic voice signal computed from the simulated sound signal show typical characteristics for glottal insufficiency and asymmetric vocal fold oscillations.

2. METHODS: HYBRID AEROACOUSTIC NUMERICAL LARYNX MODEL—*SIMVOICE*

The 3D aeroacoustic numeric larynx model *simVoice* is a combination of the Finite Volume (FV) CFD solver Star-CCM+ and the Finite Element (FE) solver CFS++ (Kaltenbacher, 2015). The basic *simVoice* model was validated against a silicone model that provided an extensively large set of experimental data, including the vocal fold motion, the flow field, and produced sound field (Kniesburges et al., 2013, 2016, 2020; Lodermeyer et al., 2015, 2018). Characteristic parameters of the silicone model performance and corresponding physiological male values are shown in **Table 1**. Validation parameters in detail were: (1) Flow dynamic properties as pressure measurements and the

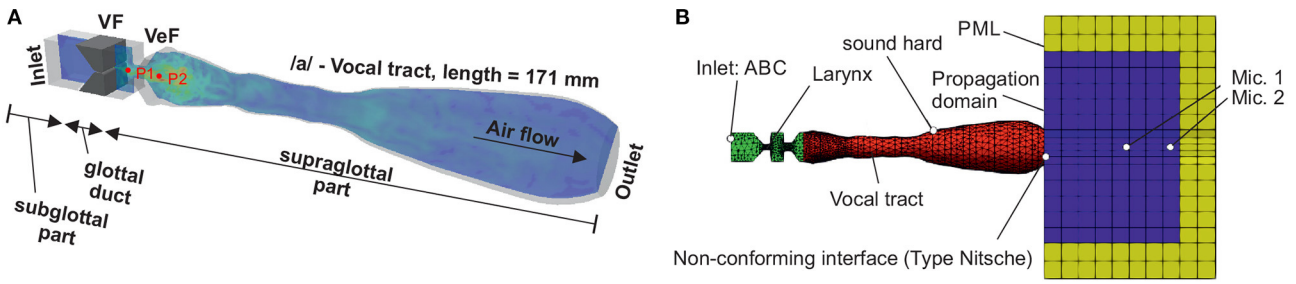


FIGURE 2 | (A) 3D representation of *simVoice*, including a velocity field in the mid-coronal plane, the vocal folds (VF), the ventricular folds (VeF), and the vowel /a/-vocal tract. Points P1, and P2 are located 6 mm, and 20 mm in distance to the vocal folds. **(B)** Geometry and domain of the CAA model of *simVoice* as introduced by Schoder et al. (2020). Mic.1 and Mic.2 are located 5 and 8 cm in distance of the vocal tract exit (mouth).

velocity field with the glottal jet in the supraglottal region using particle image velocimetry (PIV) by Sadeghi et al. (2018, 2019a), and (2) the acoustic signal by Schoder et al. (2020). In this study, the investigated configurations of glottal insufficiency and asymmetric vocal fold oscillations are synthetic cases that were derived as combination from *ex vivo* (Birk et al., 2016, 2017a) and silicone model experiments (Kniesburges et al., 2013). Thus, there are no experimental data for validation purposes.

2.1. *simVoice*—CFD

2.1.1. Geometric Dimensions

The CFD model *simVoice* represents three main parts: the subglottal section upstream of the vocal folds, the glottal duct with the two vocal folds (VF) and the supraglottal part with the ventricular folds (VeF) and an MRI-based vocal tract (VT), see **Figure 2A**. The vocal folds are based on the well-known M5 model (Scherer et al., 2001; Thomson et al., 2005) and the numerical domain dimension is obtained from the experimental setup of a synthetic vocal fold model (Becker et al., 2009; Kniesburges et al., 2013, 2016; Lodermeyer et al., 2015). All dimensions of the larynx structures are in the human length scale (Titze, 2000). The basic development of *simVoice* is described in (Sadeghi et al., 2018, 2019a,b). The gap between the VeF is 5 mm as in (Sadeghi et al., 2019a). The vocal tract represents the vowel /a/ and was developed by Probst et al. (2019) based on MRI data of 6 professional tenors (Echternach et al., 2011). Probst et al. (2019) simplified the single tenors' VTs with the method introduced by Story et al. (1996) and generated a mean vocal tract model by averaging the six single vocal tracts. The resulting staged vocal tract model was subsequently smoothed with linear interpolation. Arnela et al. (2016) showed, that using a simplified vocal tract instead of a realistic vocal tract is an appropriate approach. The distance between the vocal folds and the outlet of the vocal tract is 171 mm.

2.1.2. Modeling the Glottis Geometry

In this study, four types of clinically seen glottis closures (GC1 to GC4) were designed that are based on high-speed recordings obtained from experiments with *ex vivo* porcine larynges by Birk et al. (2016, 2017a), see **Figure 3**. Furthermore, an additional type GC5 with a rectangular glottis shape, similar to a midmembranous gap (Södersten et al., 1995), was modeled. GC1 to GC4 represent posterior gaps with an increasing glottal insufficiency, whereas GC5 represents a complete glottal insufficiency. A glottal insufficiency can not only occur in pathological phonation cases as a result of scars, paresis, paralysis, or age-related atrophy (Bhatt and Verma, 2014; Vaca et al., 2017), but also in physiological phonation of women or children with a triangular-shaped gap located at the posterior part of the glottis (Södersten and Lindestad, 1990; Rammage et al., 1992; Södersten et al., 1995; Inwald et al., 2010; Döllinger et al., 2012; Patel et al., 2012). All GC types are modeled by two parameters: (1) the initial glottal gap area and (2) the length of the closed part of the glottis divided by the entire glottis length. As shown in **Figure 3**, the modeled glottis is either fully closed (GC1: 100% $Length_{VF}$), partly closed (GC2: 60% and GC3: 30% $Length_{VF}$), or completely open (GC4 and GC5: 0% $Length_{VF}$) at the initial glottal gap. The initial glottal gaps for GC2 to GC4 are based on the glottal gap index of Birk et al. (2016, 2017a) and the initial glottal gap of GC5 is half the maximum GAW of the synthetic model (Kniesburges et al., 2016). As described by Sadeghi et al. (2018), there must be a small area between both vocal folds of 0.5 mm^2 at GC1 to reach a numerically stable simulation. Nevertheless, this small gap still interrupts the flow through the glottis during the closed phase, as shown by Sadeghi et al. (2019b). For GC2, GC3, and GC4, the initial glottal gaps possess a triangular and for GC5 a rectangular shape, see **Figure 3**.

2.1.3. Vocal Fold Motion

The lower part of **Figure 3** shows the phases of the synthetic vocal folds during one oscillation cycle (Lodermeyer et al., 2015; Kniesburges et al., 2020) and the corresponding glottal area waveform (GAW). The GAW is computed as the change of glottal area over time and is a common measure for the description of laryngeal dynamics. Based on the GAW of the synthetic model (Kniesburges et al., 2016), the oscillation of the vocal folds is modeled in *simVoice* as proposed by Sadeghi et al. (2018). In the right part of **Figure 3**, the five GC types combined with the respective modified GAWs are shown. The GAW for GC1 is equal to that used by (Sadeghi et al., 2018). The GAWs for GC2 to GC5 were computed as follows:

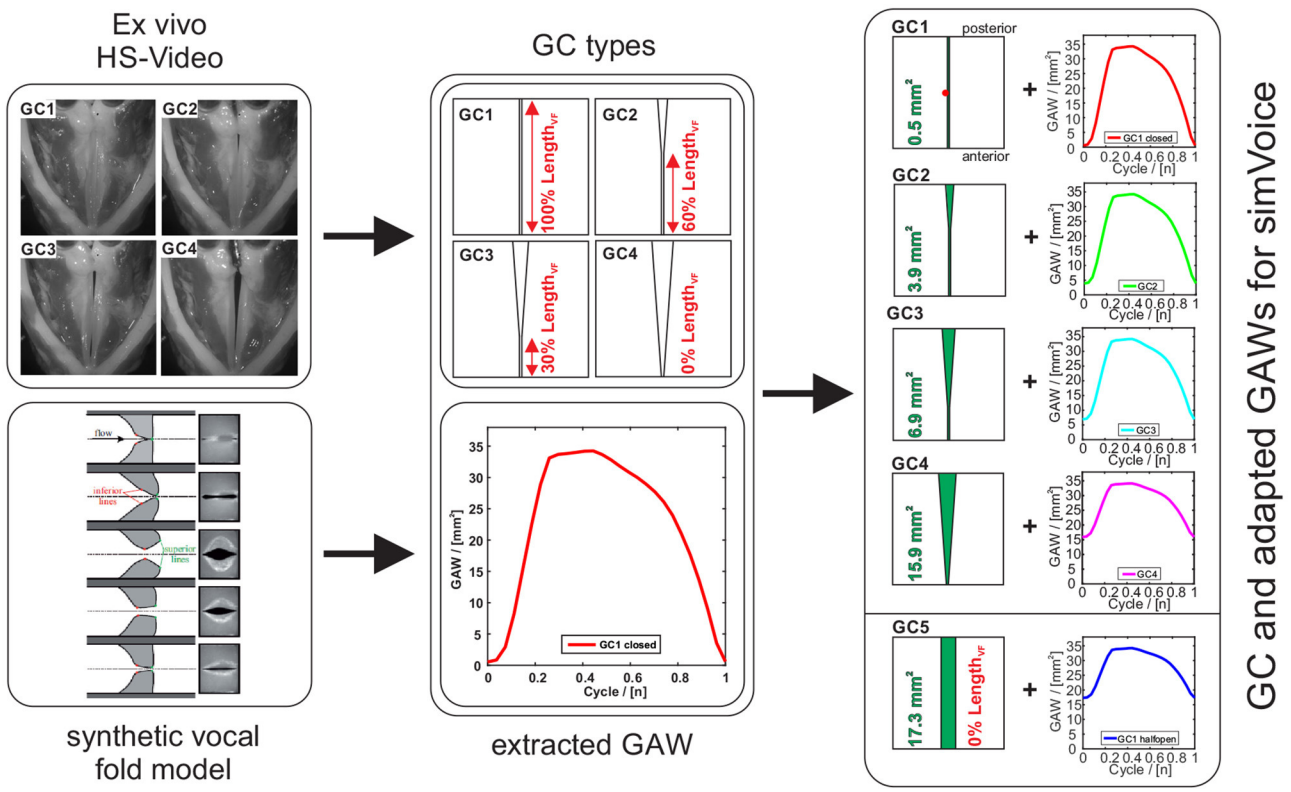


FIGURE 3 | Workflow of vocal folds modeling. Upper left part: four GC types of *ex vivo* experiments based on high-speed videos (Birk et al., 2017b) and the corresponding schematic numeric GC geometries (superior view). Lower left part: phases of the vocal fold motion of the synthetic vocal fold model (view on the coronal plane) during one oscillation cycle (Lodermeyer et al., 2015) and the GAW was taken from high-speed videos (Sadeghi et al., 2018). Right part: Four plus one additional GC types with the adapted GAWs.

$$A_i(t) = A_i^0 + \frac{A_0^{max} - A_i^0}{A_0^{max}} \cdot A_0(t) \tag{1}$$

where $A_i(t)$ is the modified GAW (of the individual GC type), subscript $i = 0$ indicates the GAW of the synthetic model of Kniesburges et al. (2016), and subscript $i = 1$ to 5 indicates GC1 to GC5. A_i^{max} is the maximum value of the GAW and A_i^0 represents the initial glottal gap area, see **Figure 3**.

We explicitly selected one motion pattern in combination with the five increasing levels of glottal insufficiency (GC1-GC5). With this strategy, we avoided to include individual effects of the patient-specific motion that may overlap the effects of the glottal insufficiency in the acoustic results.

To reduce the computational costs of the CFD simulations, the vocal fold dynamics are externally forced with characteristic dynamic patterns according to the modified GAWs. The computation of the elliptic shaped vocal fold motion is generated by a sinusoidal function along the two vocal folds (Sadeghi et al., 2018). Additionally, Sadeghi et al. (2018) added a simple convergent-divergent standard mucosal wave-like motion model based on experiments (time periods of convergent and divergent glottal duct shapes) and the literature for typical angles of the glottal duct during oscillation (Titze, 2000). It contains a convergent shaped glottal duct during the opening (0.1 T to 0.32 T) with an angle range of 0° to 5° and a divergent duct (0.32 T to 0.9 T) with angles of −10° to 0°. The glottis is closed between 0.9 T and 0.1 T of the next cycle. The 3D vocal fold motion is realized by moving wall boundaries of the vocal folds that form the glottal duct, see **Supplementary Video 1**. For all GC types, the vocal folds oscillate with a fundamental frequency of $f_0 = 148\,Hz$. The maximum glottis width of 4.66 mm is in the range as reported for *ex vivo* male larynx studies (up to 5.6 mm)(Döllinger et al., 2005; Döllinger and Berry, 2006a,b; Boessenecker et al., 2007) but higher than reported for *in vivo* measurements (up to 2.8 mm) (George et al., 2008; Semmler et al., 2018).

For the symmetric motion, both vocal folds move equally but in opposite directions. The left-right asymmetric vocal fold motion is realized by reducing the amplitude of one vocal fold to 50% (of the original amplitude), see **Figure 4** and **Supplementary Video 2**. Subsequently, the asymmetric motion reduces the corresponding maxima of the GAWs to 75% compared to the symmetric cases.

2.1.4. Boundary Conditions

At all walls of the *simVoice* model, no-slip no-injection boundary conditions were applied. The walls of the moving vocal folds were defined as moving wall boundaries. For all simulation cases, the

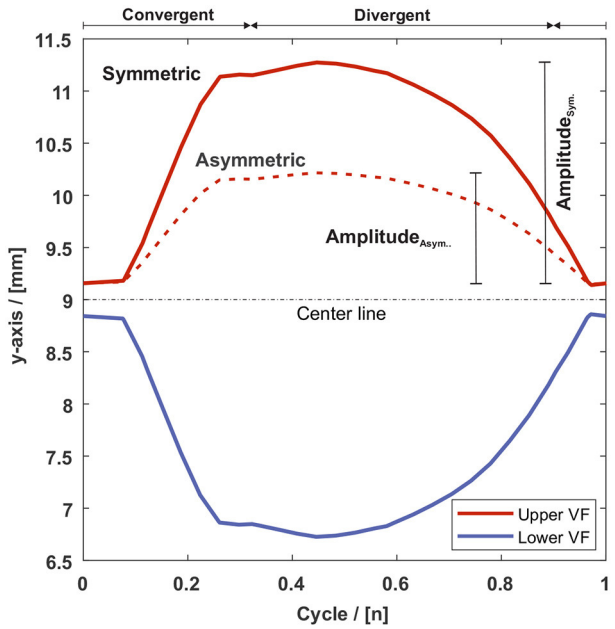


FIGURE 4 | Exemplary vocal fold motion of GC1 for the symmetric and asymmetric case along the y-axis (medial-lateral direction) for a point on the medial plane of the VF surface, see the red mark at GC1 in **Figure 3**. The solid red line represents the motion on the y-axis of the upper vocal fold for the symmetric and the dashed line for the asymmetric motion. The blue line represents the motion on the y-axis of the lower vocal fold.

mean pressure of the subglottal inlet boundary is $P_{inlet} = 775Pa$ that is in the physiologic range of human lunge pressures during normal phonation (Titze, 2000). The mean pressure at the outlet, which represents the mouth, is $P_{outlet} = 0\,Pa$. The kinematic viscosity of air was specified as $\nu = 1.5666 \cdot 10^{-5}\,\frac{m^2}{s}$ and the density of air constant at $\rho = 1.18415\,\frac{kg}{m^3}$ as the Mach number is $Ma < 0.3$ (Kniesburges et al., 2011).

2.1.5. Numerical Methods

The numerical setup is identical to the previous studies (Sadeghi et al., 2018, 2019a,b). To perform the simulations of *simVoice*, we use the software package STAR-CCM+ (Siemens, PLM Software, Plano, TX, USA) with a finite-volume cell-centered non-staggered grid. For modeling the turbulence, Large Eddy Simulations (LES) in combination with a Wall-Adapting Local Eddy-Viscosity (WALE) subgrid-scale model (Nicoud and Ducros, 1999) were carried out. The convective and diffusive terms of the Navier-Stokes equations were discretized with a central difference scheme with second-order accuracy. Subsequently, the pressure-correction PISO algorithm (Pressure-Implicit with Splitting Operators) solves the pressure-velocity linked equations non iteratively. Finally, an Algebraic Multigrid (AMG) method with a Gauss-Seidel relaxation scheme was applied to solve the final linear system of equations.

2.1.6. Mesh Generation

The mesh consists of hexahedral cells and is based on the mesh presented by Sadeghi et al. (2019b). For the mesh independence study, GC1 and a symmetric vocal fold motion was conducted. Starting with the base mesh (MB) with 2.9 million cells, three more meshes (M1-M3) with a decreasing number of cells were generated, see **Supplementary Table 1**. The limit for the mesh coarsening was set by the Taylor micro-scale $\lambda_T = 0.085\,mm$ according to Mihaescu et al. (2010). **Figure 5A**) shows the flow rate for one oscillation cycle. M1-M3 produced a similar trend and the mean relative deviation to MB ranges between −1.3% and +2.6%, see **Supplementary Table 1**. Whereas M3 shows the best accordance with MB in the cycle range 1.4 to 1.8, M1 and M2 deviate from the trend of MB. **Figure 5B** shows an instantaneous pressure evolution at point P1 with a good agreement of meshes M1-M3 in comparison with mesh MB. Small deviations at the beginning and the end of the cycle are visible, which are the result of different instantaneous turbulent fluctuations at point P1 (Sadeghi et al., 2019b), see **Figure 5B**. Summarizing, M3 with the lowest number of cells shows good agreement with the base mesh MB. The resulting mesh M3 is assembled of 1.3 million cells with a basic cell size of 0.68 mm, see **Supplementary Table 1**.

The near-wall flow is modeled by the all-y+ model of Star-CCM+ that can handle fine and coarse meshes (Reichardt, 1951). The first cell layers on the vocal fold walls have a $y+ = 1$. The time step size is set to $1.36 \cdot 10^{-6}\,s$, and the corresponding mean CFL number is 3.5 that is appropriate for implicit solvers (Anderson, 1995; Hirsch, 2007). *simVoice* uses the overset mesh approach of STAR-CCM+ to realize the vocal fold motion. This chimera method combines a fixed Eulerian background mesh with an Arbitrary Lagrangian-Eulerian (ALE) overlapping mesh (Hadzic, 2005). In *simVoice*, the mesh around both vocal folds represents the overlapping or overset mesh. Consequently, the total number of cells changes over time and depends on the GC type and the distance between the vocal folds during the oscillation.

2.2. *simVoice*—CAA Model

2.2.1. Geometry Dimensions

The acoustic model of *simVoice* has been introduced by Schoder et al. (2020). According to the hybrid aeroacoustic approach, the acoustic domain captures the CFD domain assembled by the larynx and the vocal tract, where the acoustic sources occur. This region is coupled to a propagation domain in which the microphone points Mic1 and Mic2 are located, see **Figure 2B**. These points are positioned on the centerline of the vocal tract at a distance of 5 cm and 8 cm from the vocal tract exit (mouth). Additionally, perfectly matched layers (PML) surround the propagation domain to ensure free field radiation (Kaltenbacher, 2015). Owing to the plane wave approximation, we use an absorbing boundary condition (ABC) at the inlet that requires less computing power compared to PML (Kaltenbacher, 2015). Furthermore, all solid walls are modeled as acoustically hard.

To preserve mesh flexibility and element quality, the acoustic computation grid is composed of two conforming meshes linked via a non-conforming Nitsche-type mortaring interface. The mesh of the larynx and the vocal tract was generated for each GC type separately, representing the geometry of the maximum VF opening. It consists of tetrahedral finite elements with a maximum cell size of 5.7 mm. In contrast, the mesh in the

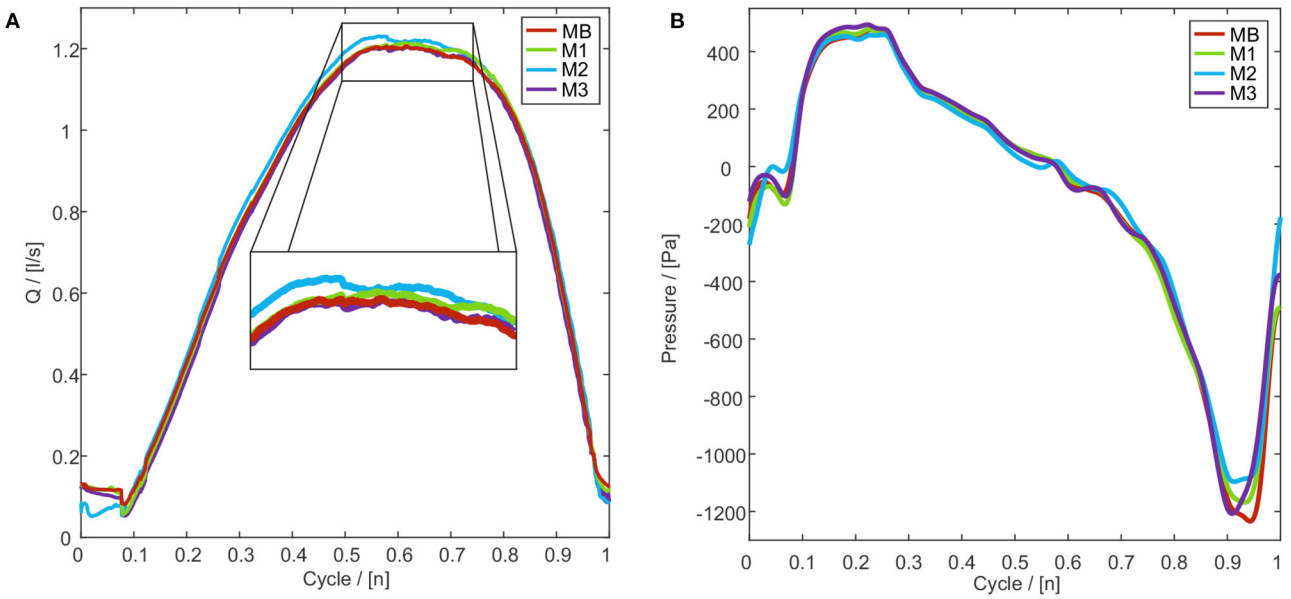


FIGURE 5 | (A) Volume flow rate through the glottis for one oscillation cycle for different mesh resolutions MB-M3. **(B)** Instantaneous pressure evolution for mesh resolutions MB-M3 for one oscillation cycle at point P1, see **Figure 2**. The pressure evolutions were smoothed by a low-pass filter (Butterworth), with a cut-off frequency of 2,000 Hz, to reduce the numerical noise.

propagation domain is the same for all GC types and has hexahedral elements with a cell size of about 10.9 mm.

2.2.2. Numerical Methods

The aeroacoustic sound generation and acoustic wave propagation is described by the perturbed convective wave equation (PCWE) (Kaltenbacher et al., 2016), which is solved via the finite element solver CFS++ (Schoder et al., 2020). To compute the acoustic source term for the PCWE, the incompressible pressure field from the CFD is transferred onto the CAA mesh by a conservative interpolation scheme based on a cut cell algorithm (Schoder et al., 2019, 2020). The acoustic source term is then computed on the CAA grid as the partial time derivative of the incompressible pressure. We modeled a one-way coupling from the flow to the acoustic sources which was found to be valid for normal voice production (Schoder et al., 2020). A back-coupling effect from the acoustics to the flow field was not considered.

2.3. *simVoice*—Data Acquisition and Analysis

A total of 20 oscillation cycles of the vocal folds were simulated. In a first step, the *simVoice* CFD simulations were executed for 10 oscillation cycles to produce a fully developed flow field. After these 10 initializing oscillations, another 10 oscillation cycles were simulated to provide valid data for the analysis. As shown by **Supplementary Figure 1** the model has achieved repeatable periodic oscillations with the flow field fully converged. The mean cyclic pressure at P1 fluctuates in the range of −7.1 and 9.1% and for P2 in the range of −9.1 and 6.5%, see **Supplementary Figure 1A**). These small fluctuations highly depend on the turbulent characteristic and the small cycle-to-cycle changes of the fluid flow in the supraglottal region (Kniesburges et al., 2016). The mean volume flow $\overline{Q}$ of the 10 initial oscillations is nearly constant and fluctuates in the range of −0.4 and 1.2%, see **Supplementary Figure 1B**). For the analysis, the complete 3D pressure and velocity fields were exported at every 10th time-step. These flow field data are then imported into CFS++ to determine the acoustic sources and to run the simulation of sound propagation. Finally, the acoustic signals at the two microphone positions were used. The sound pressure level (SPL) was calculated at a reference sound pressure of $p_0 = 20\,\mu Pa$ using a Matlab (Mathworks, USA) routine. Therefore, the acoustic potential of Mic.2, see **Figure 2B**, was extrapolated to a distance of 20 cm far from the vocal tract outlet to match the distance of *ex vivo* studies (Birk et al., 2016, 2017b). The Vocal Efficiency (VE) is calculated as proposed by Riede et al. (2019) and Titze (1992):

$$\mathrm{VE} = \frac{P_\mathrm{r}}{P_\mathrm{a}} = \frac{4 \cdot \pi \cdot R^2 \cdot 10^{\frac{SPL-120}{10}}}{P_\mathrm{sub} \cdot \overline{Q}} \tag{2}$$

where P_r is the radiated acoustic power, P_a is the aerodynamic power, R is the distance of the microphone to the opening of the vocal tract, P_sub is the subglottal pressure, and $\overline{Q}$ is the mean volume flow through the glottis. Additionally, the computed acoustic pressures were analyzed by the in-house Glottis Analysis Tool (GAT) for obtaining the Cepstral Peak Prominence (CPP) (Hillenbrand et al., 1994). The CPP is a spectra-based, well-established and objective measure to judge for perceived breathiness or vocal fatigue (Hillenbrand et al., 1994; Hillenbrand and Houde, 1996; Brinca et al., 2014;

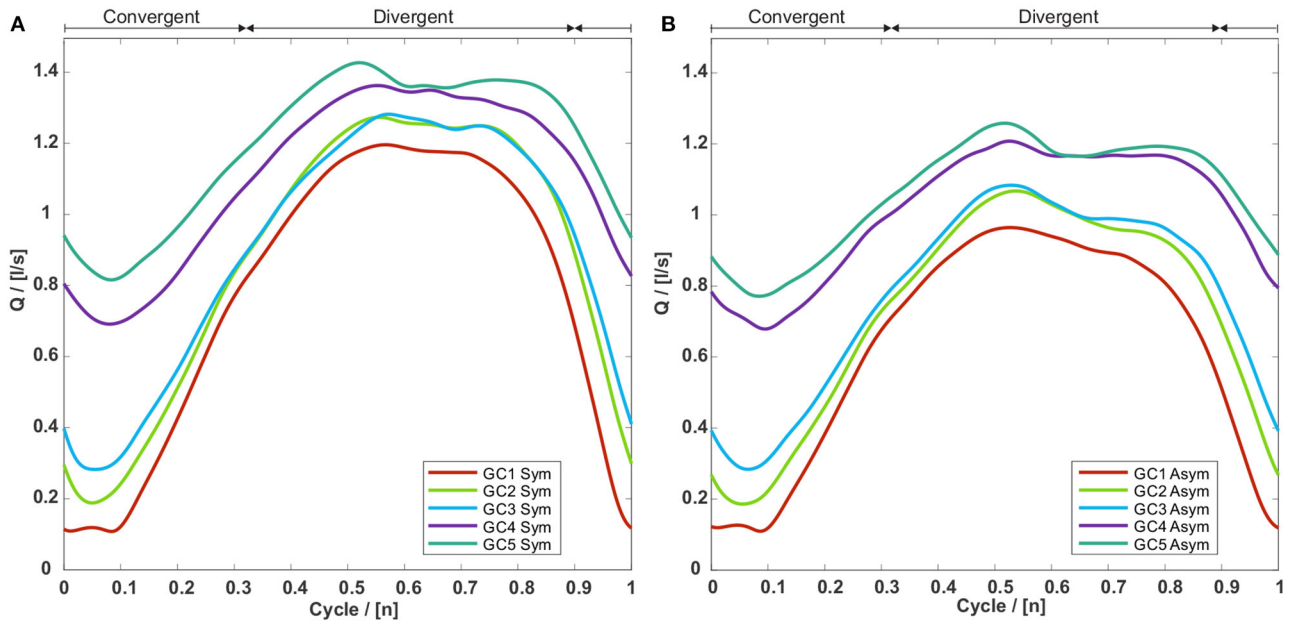


FIGURE 6 | Volume flow through the glottis for the five GC types with **(A)** a symmetric and **(B)** an asymmetric vocal fold motion. For both motion types the volume flows are rising with an increasing glottal insufficiency, whereas the corresponding volume flows of the asymmetric motion are collectively smaller than those of the symmetric motion.

Samlan et al., 2014; Samlan and Story, 2017; Patel et al., 2018; Mahalingam et al., 2020; Murton et al., 2020) and has proven to be a more reliable measure of dysphonia than time-based measures (Heman-Ackah et al., 2003). The exact computation procedure is shown in Birk et al. (2016). The CFD data are evaluated concerning the volume flow through the glottis, the glottis resistance as proposed by van den Berg et al. (1957), and the energy transfer between the airflow and the vocal folds tissue. The energy transfer is defined by the work performed by the aerodynamic forces on the moving VFs according to Thomson et al. (2005).

3. RESULTS

3.1. Aerodynamic Characteristics

3.1.1. Volume Flow

The minimum, maximum, and mean volume flow through the glottis consequently increases with an increasing glottal insufficiency from GC1 to GC5 for symmetric and asymmetric vocal fold motions as shown in **Figure 6** and **Table 2**. The flow rate decrease comparing symmetric and asymmetric motion amounts between 9.0% (GC4) and 18.2% (GC1) as displayed in **Table 2**.

3.1.2. Glottis Resistance

The flow resistance across the glottal duct R_{Glottis} (Kniesburges et al., 2017) decreases with an increasing glottal insufficiency. The reason for this decrease in R_{Glottis} is the rising flow rate $\overline{Q}$, while the P_{sub} remains constant. The direct comparison of R_{Glottis} between symmetric and asymmetric vocal fold motion yielded a larger resistance for the asymmetric motion because $\overline{Q}$ is reduced owing to the smaller glottal gap, see **Table 2**.

TABLE 2 | Mean volume flow through the glottis $\overline{Q}$, the glottis resistance R_{Glottis}, and the net energy W_{net} of all GC types.

Parameter	GC1	GC2	GC3	GC4	GC5
$\overline{Q}^{\text{sym}}$ in $[\frac{l}{s}]$	0.77	0.88	0.91	1.11	1.20
$\overline{Q}^{\text{asym}}$ in $[\frac{l}{s}]$	0.63	0.73	0.78	1.01	1.06
rel.Dev.	−18.2%	−17.0%	−14.3%	−9.0%	−11.7%
$R^{\text{sym}}_{\text{Glottis}}$ in $[\frac{Pa \cdot s}{m^3}]$	1168.9	1044.3	1003.2	845.6	915.0
$R^{\text{asym}}_{\text{Glottis}}$ in $[\frac{Pa \cdot s}{m^3}]$	1397.3	1262.0	1183.5	902.4	1032.3
rel.Dev.	19.5%	20.8%	18.0%	6.7%	12.8%
$W^{\text{sym}}_{\text{net}}$ in $[\mu J]$	165.4	167.1	148.7	73.1	79.1
$W^{\text{asym}}_{\text{net}}$ in $[\mu J]$	114.8	113.1	105.2	54.1	65.2
rel.Dev.	−30.6%	−32.3%	-29.3%	−26.0%	−18.0%

Relative deviation (rel.Dev.) refers to deviation of asymmetric to symmetric motion values. $\overline{Q}$ increases while R_{Glottis} and W_{net} decrease with increasing glottal insufficiency. However, in contrast to R_{Glottis}, W_{net} decreases for asymmetric motion owing to the smaller total amplitude of the glottis oscillation.

3.1.3. Energy Transfer

As proposed by Sadeghi et al. (2019a), the total transferred work (W_{net}) during one oscillation cycle is calculated, see **Table 2**. For both motion types, the total net work during an oscillation cycle is positive, being typical for vocal fold oscillations during phonation (Thomson et al., 2005; Luo et al., 2009). Furthermore,

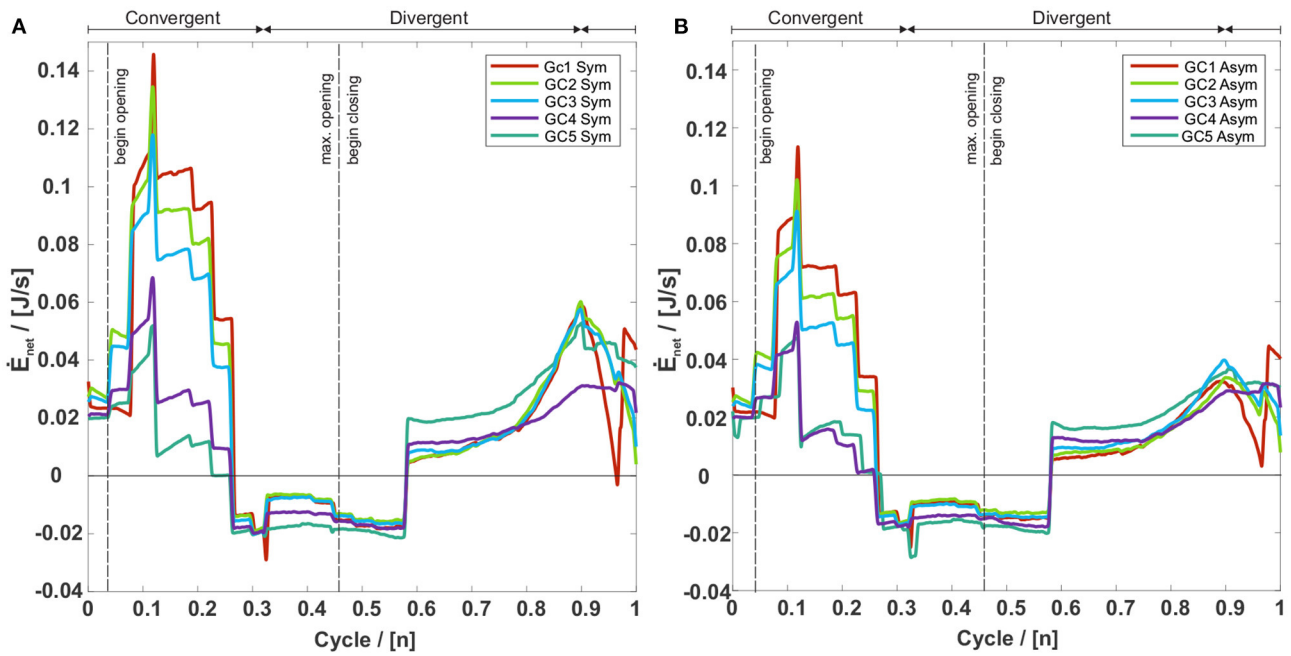


FIGURE 7 | Net rate energy transfer ($\dot{E}_{net}$) of the five GC types with **(A)** a symmetric and **(B)** an asymmetric vocal fold motion. A positive $\dot{E}_{net}$ means an energy flux from the glottal flow toward the vocal folds and a negative $\dot{E}_{net}$ an energy flux from the vocal folds toward the airflow. For both motion types $\dot{E}_{net}$ is positive at the beginning and the end of the oscillation cycle. In these intervals $\dot{E}_{net}$ decreases with an increasing glottal insufficiency, whereas the corresponding values of the asymmetric motion are collectively smaller than those of the symmetric motion.

W_{net} decreases with an increasing glottal insufficiency. **Table 2** shows that W_{net} decreases by 55.8% (symmetric) and 52.9% (asymmetric) from GC1 to GC5 whereas the maximum deviation comparing symmetric and asymmetric motion occurs for GC2 with 32.3%. However, in contrast to $R_{Glottis}$, W_{net} decreases for asymmetric motion owing to the smaller total amplitude of the glottis oscillation. Overall, our data shows that a partially closed glottis (GC2 and GC3) in combination with an asymmetric motion produces a higher W_{net} than a contact-free symmetric oscillation, see **Table 2**.

According to Sadeghi et al. (2019a), the time derivative of the work constitutes the net energy transfer rate $\dot{E}_{net}$ between fluid and tissue. It is shown in **Figure 7** for both symmetric and asymmetric vocal fold motions. A positive $\dot{E}_{net}$ corresponds to an energy flux from the laryngeal flow into the tissue, i.e., the flow deforms the vocal folds (Sadeghi et al., 2019a). During the opening, until 0.25 T, $\dot{E}_{net}$ is positive, which indicates the tissue deformation by the laryngeal flow. Between 0.25 T to 0.58 T, the glottis width reaches its maximum, producing a negative $\dot{E}_{net}$, resulting from the tissue's resistance to deform further (Sadeghi et al., 2019a). After the flow is fully accelerated, the aerodynamic pressure between the vocal folds is minimal, which initiates the glottis's closing motion. The VFs move toward each other, starting at 0.58 T, and again a positive $\dot{E}_{net}$ arises. Although the motion of the vocal folds is prescribed in this model, Luo et al. (2009) show a similar energy transfer rate during a cycle of flow-induced VF oscillations. For clarity, we want to mention that the discrete changes in the energy transfer plots occur due to the frame rate of 4,000 fps of the camera, which was used to record the oscillations of the synthetic vocal folds (Kniesburges et al., 2013). Based on this recording the motion of the vocal folds was modeled without further smoothing and therefore discrete changes in the velocity subsequently occur at multiples of 0.25 ms. **Figure 7** further shows that the positive $\dot{E}_{net}$ during the opening and closing phases decreases with an increasing glottal insufficiency. Furthermore, in the opening and closing phase, $\dot{E}_{net}$ is lower for the asymmetric motion, whereas it is equal for both motion types during the phase of significant tissue resistance (0.25 T - 0.58 T).

3.1.4. Flow Field Structure

Figure 8 shows the supraglottal flow field at two time instances ($t_1 = 0$ and $t_2 = 0.56\,T$) during the oscillation cycle for the symmetric and the asymmetric vocal fold motion. For all GC types, a long jet expands into the supraglottal region. While GC1 fully interrupts this glottal jet at the end of the cycle, GC2 and GC3 only partly interrupt the laryngeal fluid flow at the anterior section of the glottis. For GC4 and GC5, the vocal folds remain open along the entire glottis length during the oscillation cycle, see **Supplementary Videos 3, 4**. This absent interruption of the glottal jet is often related to an aspirated voice signal characterized by lower tonal sound components (Fritzen et al., 1986; Bhatt and Verma, 2014; Kniesburges et al., 2020). As reported by Sadeghi et al. (2018), the VeFs have a stabilizing influence on the glottal jet. Therefore, no jet deflection in the medial-lateral

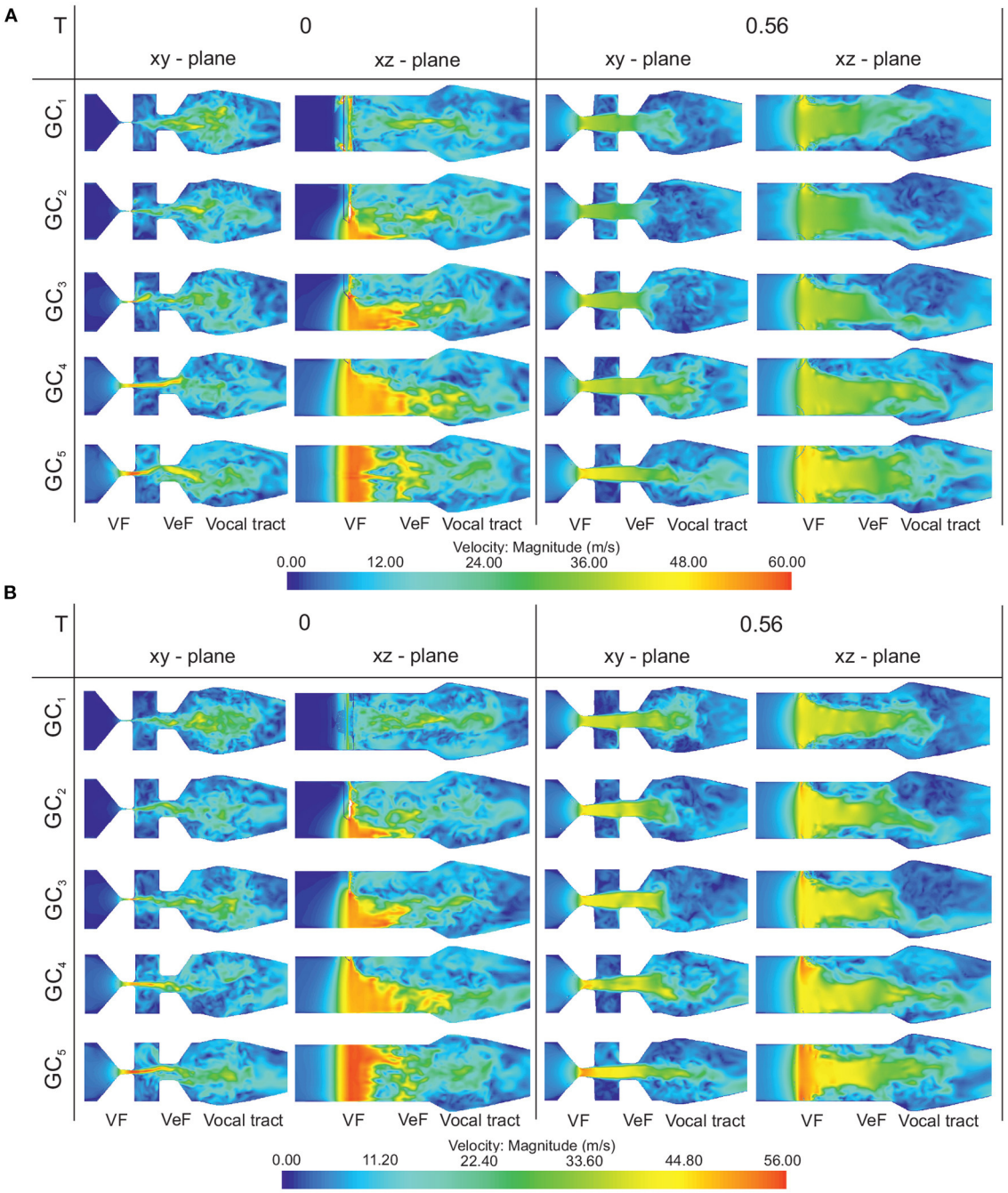


FIGURE 8 | (A) Symmetric vocal fold motion: velocity magnitude in the midcoronal (xy-plane) and the sagittal (xz-plane) plane for the five GC types at two instances ($t_1 = 0$ and $t_2 = 0.56\,T$) of an oscillation cycle. While GC1 fully interrupts the glottal jet at the end of the cycle, GC2 and GC3 only partly, and GC4 and GC5 do not interrupt the laryngeal fluid flow. **(B)** Asymmetric vocal fold motion: Velocity magnitude in the midcoronal (xy-plane) and the sagittal (xz-plane) plane for the five GC types at two instances ($t_1 = 0$ and $t_2 = 0.56\,T$) of an oscillation cycle. The upper vocal fold moves with the 50% amplitude and the glottal jet impinges mainly the lower VeF and subsequently, just a vortex in the lower ventricle occurs.

directions (**Figure 8** in the xy-plane) can be observed, see **Supplementary Videos 5, 6**. However, the glottal opening shape has a strong influence on the posterior-anterior jet shape (**Figure 8** in the xz-plane), see also **Supplementary Videos 7, 8**. As similarly reported by Zörner et al. (2016), triangular glottal orifices deflect the jet toward the larger glottal opening that occurs for GC2 and GC3 at the posterior end of the glottis.

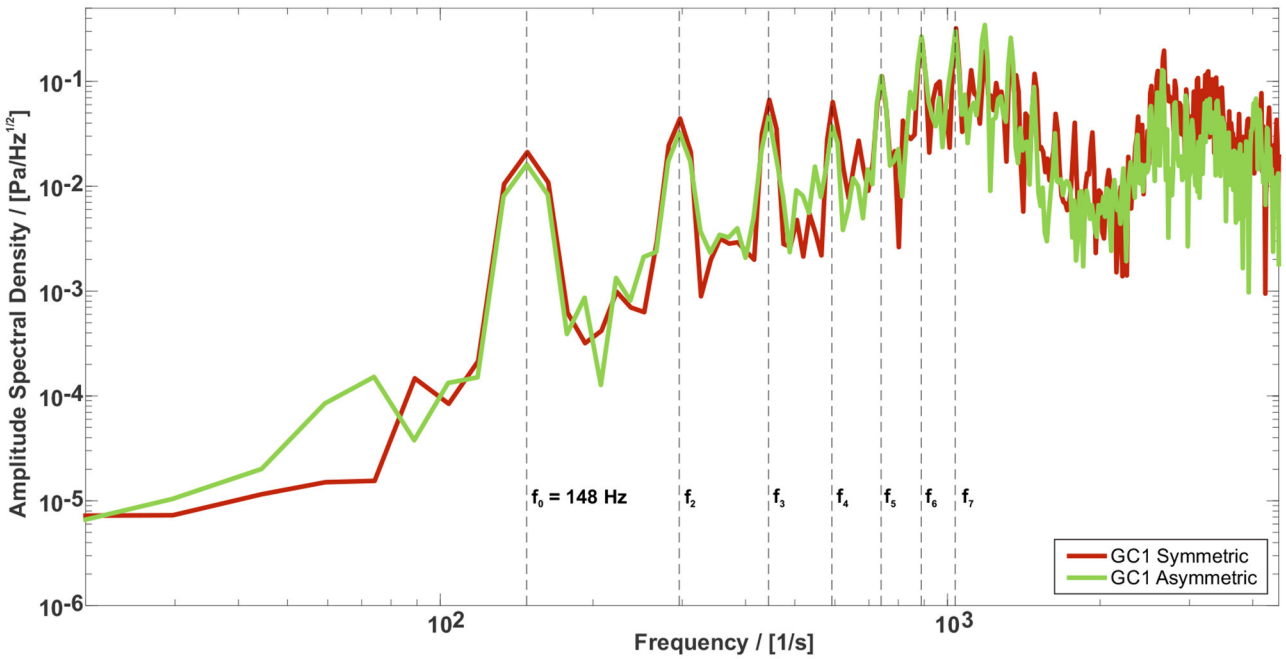


FIGURE 9 | Amplitude Spectral Density (ASD) for the GC1 type for the symmetric and asymmetric vocal fold motion. The spectra of both motions show similar slope and only slight deviations in the amplitudes at the fundamental frequency, whereas more significant differences occur at higher harmonics.

For the symmetric vocal fold motion, the glottal jet impinges both VeF during the oscillation cycle and vortices arise in both ventricles. For the asymmetric case, the glottal jet impinges mainly the lower VeF and subsequently, just a vortex in the lower ventricle occurs, see **Figure 8B** for $t_2 = 0.56\,T$ in the xy-plane. Furthermore, the maximum glottal velocity is higher for the symmetric vocal fold motion than for the asymmetric vocal fold motion due to the larger flow rate in the symmetric cases, see color bars in **Figure 8**.

3.2. Quality of Acoustic Voice Signal

3.2.1. Spectral Analysis and Formant Frequencies

Figure 9 shows the amplitude spectral density (ASD) of the sound signals for GC1 and both symmetric and asymmetric vocal folds motions measured at the Mic.1 position, see **Figure 2B**. Both spectra exhibit the main peak at the oscillation frequency of the vocal folds $f_0 = 148\,Hz$, followed by their harmonics. Comparing the spectra of all GC types shows similar slope and only slight deviations in the amplitudes at the fundamental frequency, whereas more significant differences at the higher harmonics occur, see **Supplementary Figures 2, 3**. Regarding the motion type of the vocal folds, the harmonic tones are more pronounced for the symmetric vocal fold motion, especially in the frequency range between $1,000$ and $2,000\,Hz$. This variance in the acoustic spectra of the radiated sound was also found by Zörner et al. (2016) although the velocity fields of the five GC types are considerably different.

A modal analysis of the vocal tract shows that the first two formants $F_1 = 1,020\,Hz$ and $F_2 = 1,350\,Hz$, see transfer function of /a/ vocal tract in **Supplementary Figure S4**, are well-positioned within the region of the /a/ vowel of the formant chart of Peterson and Barney (1952), shown in **Figure 10**.

3.2.2. Sound Pressure Level (SPL) and Vocal Efficiency (VE)

Figure 11A presents the SPL for all GC types. SPL significantly decreases with an increasing glottal insufficiency: For the symmetric motion type from 91.8 dB for GC1 to 82.4 and 84.2 dB for GC4 and GC5 representing a loss of 10.2 and 8.3%. For the asymmetric motion type, a decrease of about 4.5% for GC2 and GC3, 1.9% for GC4, and 4.9% for GC5, was found compared to SPL = 89.8 dB for GC1. The comparison between both motion types shows only minor differences. A maximum deviation of 6.4% for a higher SPL at the asymmetric motion occurs at GC4. **Figure 11B** shows the VE of all GC types. As mentioned before, the VE is the ratio of radiated acoustic power to aerodynamic power, see Equation (1). According to the SPL, the VE decreases for both vocal fold motion types (symmetric vs. asymmetric) and an increasing degree of glottal insufficiency (GC1 to GC5). The VE decreases from VE = 0.25% for GC1 to VE = 0.03% for GC5 for the symmetric motion and for the asymmetric motion, VE decreases less, from VE = 0.19% (GC1) to VE = 0.04% (GC5).

3.2.3. Cepstral Peak Prominence (CPP)

The CPP is widely used as a quantitative measure for the periodicity of a signal and thereby has proven to be a reliable indicator for the strength of tonal components and therewith the quality of the human voice (Hillenbrand et al., 1994; Hillenbrand

and Houde, 1996; Birk et al., 2017b). It is shown in **Figure 12** for both motion types. The CPP for the symmetric vocal fold motion starts at 17.1 for GC1 and increases to 17.4 for GC2 and GC3. Afterwards, the CPP decreases to 16.2 dB for GC4 and further to 14.4 dB for GC5. For the asymmetric vocal fold motion the CPP decreases for an increasing glottal insufficiency from 17.1 dB for GC1 to 12.5 dB for GC5.

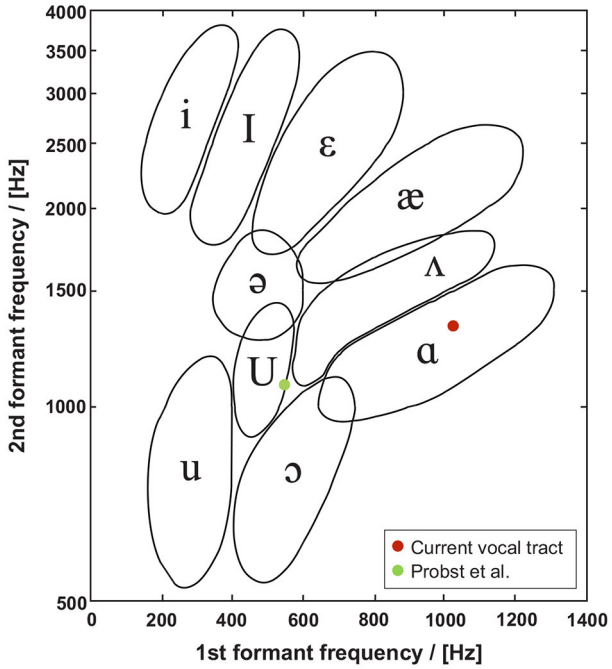


FIGURE 10 | Formant chart as proposed by Peterson and Barney (1952), shows the formant frequencies of the first two formants found in this study and that of Probst et al. (2019). In contrast to Probst et al. (2019), $F1 = 1,020\,Hz$ and $F2 = 1,350\,Hz$ simulated by *simVoice* are well-positioned within the region of the /a/ vowel.

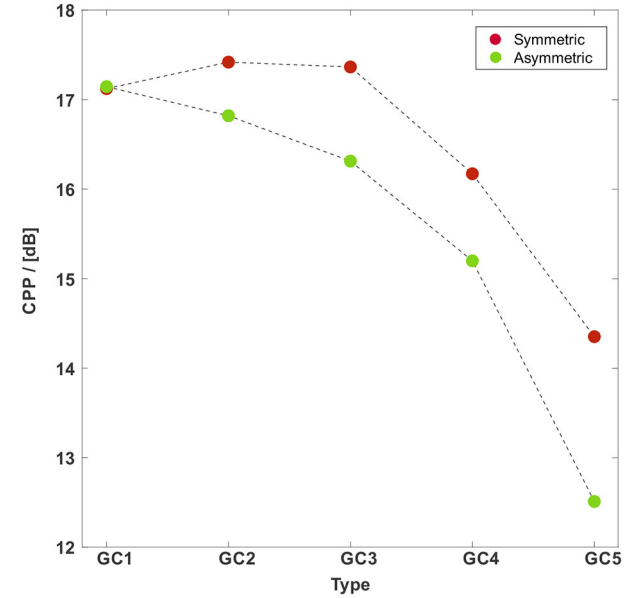


FIGURE 12 | CPP vs. the GC types with a symmetric (red points) and an asymmetric (green points) vocal fold motion. The CPP for the symmetric vocal fold motion almost remains at the same level for GC1 to GC3 followed by a decrease. The CPP for the asymmetric vocal fold motion decreases for an increasing glottal insufficiency. The CPP for the asymmetric motion is collectively smaller than those for the symmetric motion.

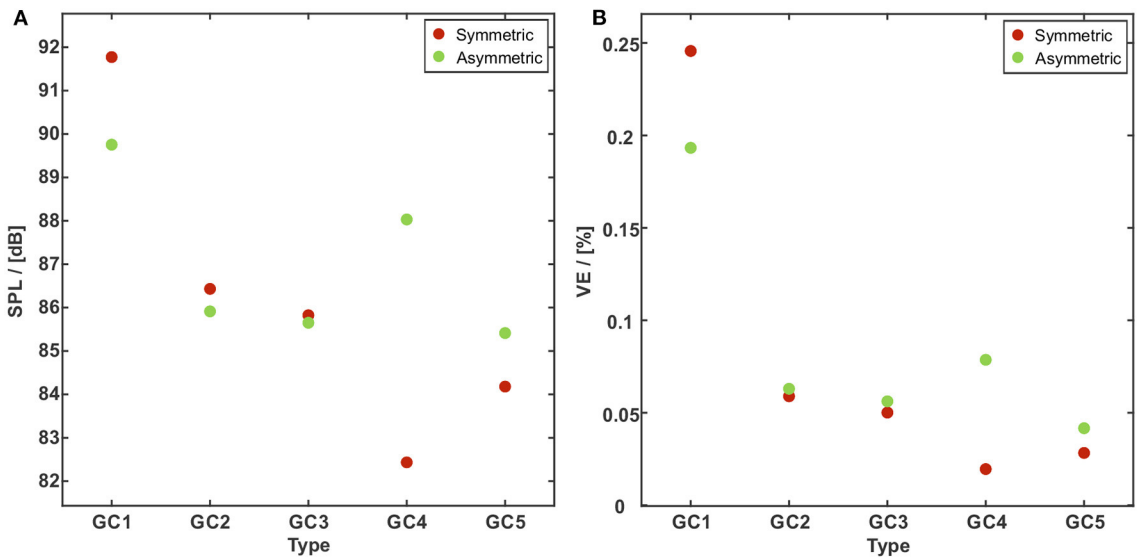


FIGURE 11 | SPL **(A)** and VE **(B)** vs. the GC types with a symmetric (red dots) and an asymmetric (green dots) vocal fold motion. The SPL and the VE significantly decrease with an increasing glottal insufficiency. The comparison between both motion types shows significant differences for GC1 and GC4 and only minor differences for GC2, GC3, andGC5.

4. DISCUSSION

4.1. Aerodynamic Characteristics

Our results of the volume flow through the glottis agree with the study by Zañartu et al. (2014) who reported an airflow rise with an increasing posterior gap. As the maximum glottal gap area of an asymmetric type is smaller than its symmetric equivalent, the mean volume flow $\overline{Q}$ is subsequently decreased, see **Table 2**. The left-right asymmetry does not only affect the maximum glottal area as reported by Pickup and Thomson (2009) but also significantly reduces the volume flow through the glottis for a constant inlet pressure in both motion types.

In phonation, the goal is to increase the energy transfer between the glottal airflow and the VFs as a beneficial mechanism to induce the VF oscillation. Kniesburges et al. (2017) interpreted the flow resistance as a measure of energy transfer from the glottal flow to the VFs. Furthermore, Birk et al. (2017b) reported that the energy transfer from the glottal airstream to the vocal folds, as indicated by the glottal resistance, is strongly dependent on glottal insufficiency. In this context, a complete glottis closure during the VFs oscillation produces a large flow resistance R_{Glottis} and in addition a large energy transfer between flow and tissue. Additionally, our data support the findings of Döllinger et al. (2018) which showed that a partially closed glottis (GC2 and GC3) in combination with an asymmetric motion may be still better than a contact-free symmetric oscillation.

In all GC cases, the interaction of the jet with the flow structures in the immediate supraglottal area causes deflection of the tail of the glottal jet. Zhang and Mongeau (2006) reported that this interaction leads to pronounced shear layers between the jet and the resting fluid with large velocity fluctuations.

4.2. Quality of Acoustic Signal

As described above, the vocal tract model is the smoothed version of the staged model developed by Probst et al. (2019). They reported formant frequencies of $F_1 = 550\,Hz$ and $F_2 = 1,080\,Hz$, being lower than the formants found here. We assume the shift of the first two formants in this study to higher values is due to the vocal tract smoothing. As reported by Jiang et al. (2017) the location of the formants and a resulting shift significantly depends on the area variation along the tract. Probst et al. (2019) and Jiang et al. (2017) found lower frequencies for the first two formants, but Jiang et al. (2017) used a vocal tract, mimicking a neutral vowel /schwa/ superimposed onto a realistic airway centerline from *in vivo* MRI measurements. Comparing the third formant F_3 of our model with that of Probst et al. (2019) shows a good agreement.

Moreover, the results of SPL show good qualitative agreement with those reported by Thornton et al. (2019) and Döllinger et al. (2018), see **Table 3**. They executed *ex vivo* experiments with rabbit larynxes and three different glottal insufficiency grades (complete glottal closure, partial glottal closure, no contact of vocal folds). They measured the SPL at a distance of 20 cm from the glottis. Furthermore, our SPL is higher than the *in vivo* measurements of Södersten et al. (1995) because the microphone in our model is located 30 cm closer to the vocal folds, nevertheless our SPL values are in the human range (Gramming et al., 1988). Our results show that an increasing posterior gap and glottal insufficiency may reduce the SPL as reported by Zañartu et al. (2014).

TABLE 3 | SPL in [dB] of Döllinger et al. (2018) and Thornton et al. (2019).

SPL in [dB]	GC1 closed	GC2/GC3 partially closed	GC4 no contact
Döllinger et al., 2018	79.1 ± 6.4	76.1 ± 7.1	69.4 ± 7.5
Thornton et al., 2019	76.7 ± 6.5	76.0 ± 7.6	59.4 ± 7.5

Even though our SPL values are higher they show good qualitative agreement with the trends reported in this study.

TABLE 4 | CPP in [dB] of Birk et al. (2017b), Döllinger et al. (2018), and Thornton et al. (2019).

CPP in [dB]	GC1 closed	GC2 30% partially closed	GC3 60% partially closed	GC4 no contact
Birk et al., 2017b	24.3 ± 5.82	21.8 ± 4.2	16.4 ± 2.82	15.7 ± 1.94
Döllinger et al., 2018	24.0 ± 4.8	22.8 ± 4.8		19.4 ± 4.9
Thornton et al., 2019	17.9 ± 4.3	15.8 ± 6.5		11.0 ± 3.4

Quantitatively, our CPP values are in the range of values reported there.

Tanaka and Gould (1985) found a low VE with a large glottal gap and a high flow rate. Due to the dependency of the radiated acoustic power from the mouth opening and therefore from the vowels (Titze et al., 2016), our results may be just valid for a vowel /a/. Although the basic trend of the VE for the asymmetric motion coincides with that for symmetric motion, the VE is mostly larger (GC2 to GC5) compared to the symmetric motion and is just lower for GC1. Thus, our results agree for GC1 with the study by Oren et al. (2016), who reported a reduction of VE for asymmetric vocal fold motion (the study does not present the degree of glottal insufficiency). We could not identify a discrete effect that produces the outlier in SPL and subsequently in VE for GC4. We assume a cumulative effect that may occur mainly in the higher frequency range of the acoustic signal.

Both effects, an increasing insufficiency, and an asymmetric vocal fold motion potentially reduce the tonal components of the acoustic signal and the voice quality. The same observations have been made in *in vivo* studies executed by Samlan et al. (2014) and Chen et al. (2011). Furthermore, the qualitative trend of CPP was also found in *ex vivo* studies with human (Birk et al., 2017b) and rabbit larynges (Döllinger et al., 2018; Thornton et al., 2019), as shown in **Table 4**. The high CPP values for GC2 and GC3 for symmetrically oscillating VFs shows, that the acoustic signal is still tonal and physiological for small posterior gaps as often observed in physiological phonation of women and childs (Södersten and Lindestad, 1990; Södersten et al., 1995; Inwald et al., 2010; Patel et al., 2012; Kniesburges et al., 2020). Quantitatively, our CPP values are in the range of values reported by Döllinger et al. (2018) and Thornton et al. (2019).

4.3. Limitations of the Study

The vocal fold vibration in this study is prescribed, neglecting the fluid-structure interaction (FSI), which is a common approach to increase the efficiency of the simulations.

5. CONCLUSION

Glottal insufficiency and asymmetric vocal fold oscillations have been investigated using our numerical aeroacoustic model *simVoice*. Aerodynamically, an increasing degree of glottal insufficiency leads to a decrease in flow resistance and a decrease in the energy transfer rate between flow and tissue. This means a reduction of the stimulation of the vocal fold oscillations and subsequently impairs the acoustic signal. Thus, CPP (Hillenbrand and Houde, 1996; Birk et al., 2017b; Döllinger et al., 2018; Thornton et al., 2019), SPL (Döllinger et al., 2018; Thornton et al., 2019), and VE (Tanaka and Gould, 1985) deteriorate for an increasing degree of glottal insufficiency.

All these findings correlate with symptoms of functional voice disorders such as breathiness, hoarseness, and an enhanced effort needed to phonate, commonly called air loss during phonation (Fritzen et al., 1986; Zhang, 2019). However, a glottis insufficiency can also occur in physiological phonation often observed in women and children who have a triangular-shaped opening located in the posterior glottis (Södersten and Lindestad, 1990; Södersten et al., 1995; Inwald et al., 2010; Patel et al., 2012; Kniesburges et al., 2020). Those persons have often a soft and quiet voice as reported by Fritzen et al. (1986) and Bhatt and Verma (2014).

In principle, the same trend of a deterioration for an increasing degree of glottal insufficiency for CPP, SPL and VE can be seen when comparing symmetric and asymmetric motion of the vocal folds: The energy transfer rate and the acoustic parameters decrease for asymmetric motion. However, this trend is not that distinct as for glottal insufficiency (Birk et al., 2017b). Therefore, a left-right asymmetry must not necessarily result in a salient reduction in voice quality, as similarly reported by Zhang et al. (2012).

From our results, we assume that a high degree of glottal insufficiency potentially displays the most severe symptom for a functional voice disorder, which has to be focused on during clinical treatment [e.g., medialization with hyaluronic acid-based materials or thyroplasty (type 1 thyroplasty)]. Thereby, the asymmetry of the motion of the vocal folds seems to have a reduced role in negatively impacting the voice quality compared to the glottal insufficiency. But both symptoms in combination will further reduce the quality of the sound signal.

Regarding the functionality of *simVoice*, the study shows: (1) *simVoice* can mimic simplified vibration characteristics and glottal geometries, (2) *simVoice* reveals separated and combined effects of aerodynamic and acoustic symptoms of a glottal insufficiency and an asymmetric vocal fold motion, and (3) a current walltime of 10 h/cycle is, with a prospective increase in computing power, very promising for a clinical approach.

Furthermore, CFD data in addition to experimental data are essential to develop, train and validate neural networks as done by Zhang (2020) and Zhang et al. (2020), which will further speed up the computing time of the phonation process and the implementing of numerical models in the clinical environment.

DATA AVAILABILITY STATEMENT

The original contributions presented in the study are included in the article/**supplementary material**, further inquiries can be directed to the corresponding author/s.

AUTHOR CONTRIBUTIONS

MD, MK, and SK conceived the study and contributed to data analysis and interpretation, supervision, and manuscript writing. SF conducted main writing and review editing. SF, BJ, SS, and PM conducted the CFD and CAA simulations, and contributed to data analysis, interpretation and manuscript writing. ME contributed to results interpretation, clinical input, review editing and provided the VT geometry. All authors contributed to the article and approved submitted version.

FUNDING

The authors acknowledge support from the German Research Foundation (DFG) under DO 1247/10-1 (no. 391215328) and the Austrian Research Council (FWF) under no. I 3702.

ACKNOWLEDGMENTS

Furthermore, the authors acknowledge support from the Central Institute for Scientific Computing (ZISC) and computational resources and support provided by the Erlangen Regional Computing Center (RRZE).

SUPPLEMENTARY MATERIAL

The Supplementary Material for this article can be found online at: https://www.frontiersin.org/articles/10.3389/fphys.2021.616985/full#supplementary-material

REFERENCES

Alku, P., Airas, M., Björkner, E., and Sundberg, J. (2006). An amplitude quotient based method to analyze changes in the shape of the glottal pulse in the regulation of vocal intensity. *J. Acoust. Soc. Am.* 120, 1052–1062. doi: 10.1121/1.2211589

Anderson, J. (1995). *Computational Fluid Dynamics: The Basics with Applications*. New York, NY: McGraw-Hill.

Arnela, M., Dabbaghchian, S., Blandin, R., Guasch, O., Engwall, O., Van Hirtum, A., et al. (2016). Influence of vocal tract geometry simplifications on the numerical simulation of vowel sounds. *J. Acoust. Soc. Am.* 140, 1707–1718. doi: 10.1121/1.4962488

Aronson, A., and Bless, D. (2009). *Clinical Voice Disorders, 4th Edn*. New York, NY: Thieme.

Baken, R. J., and Orlikoff, R. F. (2000). *Clinical Measurement of Speech and Voice*. San Diego, CA: Singular.

Becker, S., Kniesburges, S., Müller, S., Delgado, A., Link, G., Kaltenbacher, M., et al. (2009). Flow-structure-acoustic interaction in a human voice model. *J. Acoust. Soc. Am*. 125, 1351–1361. doi: 10.1121/1.3068444

Bhatt, J., and Verma, S. (2014). Management of glottal insufficiency. *Otorinolaringologia* 64, 101–107.

Birk, V., Döllinger, M., Sutor, A., Berry, D. A., Gedeon, D., Traxdorf, M., et al. (2017a). Automated setup for *ex vivo* larynx experiments. *J. Acoust. Soc. Am*. 141, 1349–1359. doi: 10.1121/1.4976085

Birk, V., Kniesburges, S., Semmler, M., Berry, D. A., Bohr, C., Döllinger, M., et al. (2017b). Influence of glottal closure on the phonatory process in *ex vivo* porcine larynges. *J. Acoust. Soc. Am*. 142, 2197–2207. doi: 10.1121/1.5007952

Birk, V., Sutor, A., Döllinger, M., Bohr, C., and Kniesburges, S. (2016). Acoustic impact of ventricular folds on phonation studied in ex vivo human larynx models. *Acta Acust. Unit. Acust*. 102, 244–256. doi: 10.3813/AAA.918941

Boessenecker, A., Berry, D. A., Lohscheller, J., Eysholdt, U., and Döllinger, M. (2007). Mucosal wave properties of a human vocal fold. *Acta Acust. Unit. Acust*. 93, 815–823.

Brinca, L. F., Batista, A. P. F., Tavares, A. I., Gonçalves, I. C., and Moreno, M. L. (2014). Use of cepstral analyses for differentiating normal from dysphonic voices: a comparative study of connected speech versus sustained vowel in European Portuguese female speakers. *J. Voice* 28, 282–286. doi: 10.1016/j.jvoice.2013.10.001

Chen, G., Kreiman, J., Shue, Y.-L., and Alwan, A. (2011). "Acoustic correlates of glottal gaps," in *Twelfth Annual Conference of the International Speech Communication Association*. Florence.

de Oliveira Rosa, M., Pereira, J. C., Grellet, M., and Alwan, A. (2003). A contribution to simulating a three-dimensional larynx model using the finite element method. *J. Acoust. Soc. Am*. 114, 2893–2905. doi: 10.1121/1.1619981

Döllinger, M., and Berry, D. A. (2006a). Computation of the three-dimensional medial surface dynamics of the vocal folds. *J. Biomech*. 39, 369–374. doi: 10.1016/j.jbiomech.2004.11.026

Döllinger, M., and Berry, D. A. (2006b). Visualization and quantification of the medial surface dynamics of an excised human vocal fold during phonation. *J. Voice* 20, 401–413. doi: 10.1016/j.jvoice.2005.08.003

Döllinger, M., Berry, D. A., and Kniesburges, S. (2016). Dynamic vocal fold parameters with changing adduction in *ex-vivo* hemilarynx experiments. *J. Acoust. Soc. Am*. 139, 2372–2385. doi: 10.1121/1.4947044

Döllinger, M., Dubrovskiy, D., and Patel, R. (2012). Spatiotemporal analysis of vocal fold vibrations between children and adults. *Laryngoscope* 122, 2511–2518. doi: 10.1002/lary.23568

Döllinger, M., Gröhn, F., Berry, D. A., Eysholdt, U., and Luegmair, G. (2014). Preliminary results on the influence of engineered artificial mucus layer on phonation. *J. Speech Lang. Hear. Res*. 57, 637–647. doi: 10.1044/2014_JSLHR-S-12-0277

Döllinger, M., Kniesburges, S., Berry, D. A., Birk, V., Wendler, O., Dürr, S., et al. (2018). Investigation of phonatory characteristics using ex vivo rabbit larynges. *J. Acoust. Soc. Am*. 144, 142–152. doi: 10.1121/1.5043384

Döllinger, M., Tayama, N., and Berry, D. A. (2005). Empirical eigenfunctions and medial surface dynamics of a human vocal fold. *Methods Inform. Med*. 44, 384–391. doi: 10.1055/s-0038-1633981

Echternach, M., Sundberg, J., Baumann, T., Markl, M., and Richter, B. (2011). Vocal tract area functions and formant frequencies in opera tenors' modal and falsetto registers. *J. Acoust. Soc. Am*. 129, 3955–3963. doi: 10.1121/1.3589249

Eysholdt, U., Rosanowski, F., and Hoppe, U. (2003). Vocal fold vibration irregularities caused by different types of laryngeal asymmetry. *Eur. Arch. Otorhinolaryngol*. 260, 412–417. doi: 10.1007/s00405-003-0606-y

Fritzen, B., Hammarberg, B., Gauffin, J., and Karlsson, I. (1986). Breathiness and insufficient vocal fold closure. *Elsevier* 14, 549–553. doi: 10.1016/S0095-4470(19)30705-3

George, N. A., de Mul, F. F., Qiu, Q., Rakhorst, G., and Schutte, H. K. (2008). Depth-kymography: high-speed calibrated 3D imaging of human vocal fold vibration dynamics. *Phys. Med. Biol*. 53, 2667–2675. doi: 10.1088/0031-9155/53/10/015

Gramming, P., Sundberg, J., Ternström, S., Leanderson, R., and Perkins, W. H. (1988). Relationship between changes in voice pitch and loudness. *J. Voice* 2, 118–126. doi: 10.1016/S0892-1997(88)80067-5

Hadzic, H. (2005). *Development and application of a finite volume method for the computation of flows around moving bodies on unstructured, overlapping grids* (Ph.D. thesis). Hamburg: Technische Universität Hamburg.

Heman-Ackah, Y. D., Heuer, R. J., Michael, D. D., Ostrowski, R., Horman, M., Baroody, M. M., et al. (2003). Cepstral peak prominence: a more reliable measure of dysphonia. *Ann. Otol. Rhinol. Laryngol*. 112, 324–333. doi: 10.1177/000348940311200406

Hillenbrand, J., Cleveland, R. A., and Erickson, R. L. (1994). Acoustic correlates of breathy vocal quality. *J. Speech Hear. Res*. 37, 769–778. doi: 10.1044/jshr.3704.769

Hillenbrand, J., and Houde, R. A. (1996). Acoustic correlates of breathy vocal quality: dysphonic voices and continuous speech. *J. Speech Hear. Res*. 39, 311–321. doi: 10.1044/jshr.3902.311

Hirsch, C. (2007). *Numerical Computation of Internal and External Flows: The Fundamentals of Computational Fluid Dynamics, 2nd edn*. Oxford: Elsevier Ltd.

Hoffman, M. R., Surender, K., Devine, E. E., and Jiang, J. J. (2012). Classification of glottic insufficiency and tension asymmetry using a multilayer perceptron. *Laryngoscope* 122, 2773–2780. doi: 10.1002/lary.23549

Holmberg, E. B., Hillman, R. E., and Perkell, J. S. (1988). Glottal airflow and transglottal air pressure measurements for male and female speakers in soft, normal, and loud voice. *J. Acoust. Soc. Am*. 84, 511–529. doi: 10.1121/1.396829

Hoppe, U., Rosanowski, F., Döllinger, M., Lohscheller, J., Schuster, M., and Eysholdt, U. (2003). Glissando: laryngeal motorics and acoustics. *J. Voice* 17, 370–376. doi: 10.1067/S0892-1997(03)00019-5

Inwald, E., Döllinger, M., Schuster, M., Eysholdt, U., and Bohr, C. (2010). Multiparametric analysis of vocal fold vibrations in healthy and disordered voices in high-speed imaging. *J. Voice* 25, 576–590. doi: 10.1016/j.jvoice.2010.04.004

Jiang, W., Zheng, X., and Xue, Q. (2017). Computational modeling of fluid-structure-acoustics interaction during voice production. *Front. Bioeng. Biotechnol*. 5:7. doi: 10.3389/fbioe.2017.00007

Jo, Y., Ra, H., Moon, Y. J., and Döllinger, M. (2016). Three-dimensional computation of flow and sound for human hemilarynx. *Comput. Fluids* 134–135:41–50. doi: 10.1016/j.compfluid.2016.04.029

Kaltenbacher, M. (2015). *Numerical Simulation of Mechatronic Sensors and Actuators: Finite Elements for Computational Multiphysics, 3rd Edn*. Berlin: Springer.

Kaltenbacher, M., Hüppe, A., Grabinger, J., and Wohlmuth, B. (2016). Modeling and finite element formulation for acoustic problems including rotating domains. *AIAA J*. 54, 3768–3777. doi: 10.2514/1.J054890

Kaltenbacher, M., Zörner, S., and Hüppe, A. (2014). On the importance of strong fluid-solid coupling with application to human phonation. *Prog. Comput. Fluid Dyn*. 14, 2–13. doi: 10.1504/PCFD.2014.059195

Kirmse, C., Triep, M., Brücker, C., Döllinger, M., and Stingl, M. (2010). Experimental flow study of modeled regular and irregular glottal closure types. *Logoped. Phoniatr. Vocol*. 35, 45–50. doi: 10.3109/14015431003667652

Kniesburges, S. (2014). *Fluid-structure-acoustic interaction during phonation in a synthetic larynx model* (Ph.D. thesis). Aachen: Shaker Verlag.

Kniesburges, S., Birk, V., Lodermeyer, A., Schützenberger, A., Bohr, C., and Becker, S. (2017). Effect of the ventricular folds in a synthetic larynx model. *J. Biomech*. 55, 128–133. doi: 10.1016/j.jbiomech.2017.02.021

Kniesburges, S., Hesselmann, C., Becker, S., Schlücker, E., and Döllinger, M. (2013). Influence of vortical flow structures on the glottal jet location in the supraglottal region. *J. Voice* 27, 531–544. doi: 10.1016/j.jvoice.2013.04.005

Kniesburges, S., Lodermeyer, A., Becker, S., Traxdorf, M., and Döllinger, M. (2016). The mechanisms of subharmonic tone generation in a synthetic larynx model. *J. Acoust. Soc. Am*. 139, 3182–3192. doi: 10.1121/1.4954264

Kniesburges, S., Lodermeyer, A., Semmler, M., Schulz, Y. K., Schützenberger, A., and Becker, S. (2020). Analysis of the tonal sound generation during phonation with and without glottis closure. *J. Acoust. Soc. Am*. 147:3285. doi: 10.1121/10.0001184

Kniesburges, S., Thomson, S. L., Barney, A., Triep, M., Petr, Š., Horáček, J., et al. (2011). *In vitro* experimental investigation of voice production. *Curr. Bioinform*. 6, 305–322. doi: 10.2174/157489311796904637

Lagier, A., Guenoun, D., Legou, T., Espesser, R., Giovanni, A., and Champsaur, P. (2017). Control of the glottal configuration in ex vivo human models: quantitative anatomy for clinical and experimental practices. *Surg. Radiol. Anat*. 39, 257–262. doi: 10.1007/s00276-016-1738-2

Larsson, H., and Hertegård, S. (2004). Calibration of high-speed imaging by laser triangulation. *Logoped. Phoniatr. Vocol*. 29, 154–161. doi: 10.1080/14015430410024353

Link, G., Kaltenbacher, M., Breuer, M., and Döllinger, M. (2009). A 2D finite-element scheme for fluid-solid-acoustic interactions and its application to human phonation. *Comput. Methods Appl. Mech. Eng*. 198, 3321–3334. doi: 10.1016/j.cma.2009.06.009

Lodermeyer, A., Becker, S., Döllinger, M., and Kniesburges, S. (2015). Phase-locked flow field analysis in a synthetic human larynx model. *Exp. Fluids* 56, 1–13. doi: 10.1007/s00348-015-1942-6

Lodermeyer, A., Tautz, M., Becker, S., Döllinger, M., Birk, V., and Kniesburges, S. (2018). Aeroacoustic analysis of the human phonation process based on a hybrid acoustic PIV approach. *Exp. Fluids* 59, 1–15. doi: 10.1007/s00348-017-2469-9

Luo, H., Mittal, R., and Bielamowicz, S. A. (2009). Analysis of flow-structure interaction in the larynx during phonation using an immersed-boundary method. *J. Acoust. Soc. Am*. 126, 816–824. doi: 10.1121/1.3158942

Luo, H., Mittal, R., Zheng, X., Bielamowicz, S. A., Walsh, R. J., and Hahn, J. K. (2008). An immersed-boundary method for flow-structure interaction in biological systems with application to phonation. *J. Comput. Phys*. 227, 9303–9332. doi: 10.1016/j.jcp.2008.05.001

Mahalingam, S., Boominathan, P., Arunachalam, R., Venkatesh, L., and Srinivas, S. (2020). Cepstral measures to analyze vocal fatigue in individuals with hyperfunctional voice disorder. *J. Voice*. doi: 10.1016/j.jvoice.2020.02.007. [Epub ahead of print].

Mendelsohn, A. H., Zhang, Z., Luegmair, G., Orestes, M., and Berke, G. S. (2015). Preliminary study of the open quotient in an ex vivo perfused human larynx. *JAMA Otolaryngol*. 141, 751–756. doi: 10.1001/jamaoto.2015.1249

Mihaescu, M., Khosla, S. M., Murugappan, S., and Gutmark, E. J. (2010). Unsteady laryngeal airflow simulations of the intra-glottal vortical structures. *J. Acoust. Soc. Am*. 127, 435–444. doi: 10.1121/1.3271276

Motie-Shirazi, M., Zañartu, M., Peterson, S. D., Mehta, D. D., Kobler, J. B., Hillman, R. E., et al. (2019). Toward development of a vocal fold contact pressure probe: sensor characterization and validation using synthetic vocal fold models. *Appl. Sci*. 9:3002. doi: 10.3390/app9153002

Murray, P. R., and Thomson, S. L. (2012). Vibratory responses of synthetic, self-oscillating vocal fold models. *J. Acoust. Soc. Am*. 132, 3428–3438. doi: 10.1121/1.4754551

Murton, O., Hillman, R., and Mehta, D. (2020). Cepstral peak prominence values for clinical voice evaluation. *Am. J. Speech Lang. Pathol*. 29, 1596–1607. doi: 10.1044/2020_AJSLP-20-00001

Nicoud, F., and Ducros, F. (1999). Subgrid-scale stress modelling based on the square of the velocity gradient tensor. *Flow Turbul. Combust*. 62, 183–200. doi: 10.1023/A:1009995426001

Oren, L., Khosla, S., and Gutmark, E. (2016). Effect of vocal fold asymmetries on glottal flow. *Laryngoscope* 126, 2534–2538. doi: 10.1002/lary.25948

Park, J. B., and Mongeau, L. (2008). Experimental investigation of the influence of a posterior gap on glottal flow and sound. *J. Acoust. Soc. Am*. 124, 1171–1179. doi: 10.1121/1.2945116

Patel, R. R., Awan, S. N., Barkmeier-Kraemer, J., Courey, M., Deliyski, D., Eadie, T., et al. (2018). Recommended protocols for instrumental assessment of voice: American speech-language-hearing association expert panel to develop a protocol for instrumental assessment of vocal function. *Am. J. Speech Lang. Pathol*. 27, 887–905. doi: 10.1044/2018_AJSLP-17-0009

Patel, R. R., Dixon, A., Richmond, A. M., and Donohue, K. D. (2012). Pediatric high speed digital imaging of vocal fold vibration: a normative pilot study of glottal closure and phase closure characteristics. *Int. J. Pediatr. Otorhinolaryngol*. 76, 954–959. doi: 10.1016/j.ijporl.2012.03.004

Peterson, G. E., and Barney, H. L. (1952). Control methods used in a study of the vowels. *J. Acoust. Soc. Am*. 24, 175–184. doi: 10.1121/1.1906875

Pickup, B. A., and Thomson, S. L. (2009). Influence of asymmetric stiffness on the structural and aerodynamic response of synthetic vocal fold models. *J. Biomech*. 42, 2219–2225. doi: 10.1016/j.jbiomech.2009.06.039

Probst, J., Lodermeyer, A., Fattoum, S., Becker, S., Echternach, M., Richter, B., et al. (2019). Acoustic and aerodynamic coupling during phonation in MRI-based vocal tract replicas. *Appl. Sci*. 9:3562. doi: 10.3390/app9173562

Rammage, L. A., Peppard, R. C., and Bless, D. M. (1992). Aerodynamic, laryngoscopic, and perceptual-acoustic characteristics in dysphonic females with posterior glottal chinks: a retrospective study. *J. Voice* 6, 64–78. doi: 10.1016/S0892-1997(05)80010-4

Reichardt, H. (1951). Vollständige Darstellung der turbulenten Geschwindigkeitsverteilung in glatten Leitungen. *Zeitschrift für Angewandte Mathematik und Mechanik* 31, 208–219. doi: 10.1002/zamm.19510310704

Riede, T., Thomson, S. L., Titze, I. R., and Goller, F. (2019). The evolution of the syrinx: An acoustic theory. *PLoS Biol*. 17:e2006507. doi: 10.1371/journal.pbio.2006507

Rogers, D. J., Setlur, J., Raol, N., Maurer, R., and Hartnick, C. J. (2014). Evaluation of true vocal fold growth as a function of age. *Otolaryngol. Head Neck Surg*. 151, 681–686. doi: 10.1177/0194599814547489

Romero, R. G., Colton, M. B., and Thomson, S. L. (2020). 3D-printed synthetic vocal fold models. *Journal of Voice*. doi: 10.1016/j.jvoice.2020.01.030. [Epub ahead of print].

Ruben, R. J. (2000). Redefining the survival of the fittest: communication disorders in the 21st century. *Laryngoscope* 110, 241–241. doi: 10.1097/00005537-200002010-00010

Sadeghi, H. (2019). *Large eddy simulations of phonatory aerodynamics in a 3D-FVM larynx model* (Ph.D. thesis). Friedrich-Alexander University Erlangen, Erlangen, Germany.

Sadeghi, H., Döllinger, M., Kaltenbacher, M., and Kniesburges, S. (2019a). Aerodynamic impact of the ventricular folds in computational larynx models. *J. Acoust. Soc. Am*. 145, 2376–2387. doi: 10.1121/1.5098775

Sadeghi, H., Kniesburges, S., Falk, S., Kaltenbacher, M., Schützenberger, A., and Döllinger, M. (2019b). Towards a clinically applicable computational larynx model. *Appl. Sci*. 9:2288. doi: 10.3390/app9112288

Sadeghi, H., Kniesburges, S., Kaltenbacher, M., Schützenberger, A., and Döllinger, M. (2018). Computational models of laryngeal aerodynamics: potentials and numerical costs. *J. Voice* 33, 385–400. doi: 10.1016/j.jvoice.2018.01.001

Samlan, R. A., and Story, B. H. (2017). Influence of left-right asymmetries on voice quality in simulated paramedian vocal fold paralysis. *J. Speech Lang. Hear. Res*. 60, 306–321. doi: 10.1044/2016_JSLHR-S-16-0076

Samlan, R. A., Story, B. H., Lotto, A. J., and Bunton, K. (2014). Acoustic and perceptual effects of left-right laryngeal asymmetries based on computational modeling. *J. Speech Lang. Hear. Res*. 57, 1619–1637. doi: 10.1044/2014_JSLHR-S-12-0405

Scherer, R. C., Shinwari, D., De Witt, K. J., Zhang, C., Kucinschi, B. R., and Afjeh, A. A. (2001). Intraglottal pressure profiles for a symmetric and oblique glottis with a divergence angle of 10 degrees. *J. Acoust. Soc. Am*. 109, 1616–1630. doi: 10.1121/1.1333420

Schneider, B., and Bigenzahn, W. (2003). Influence of glottal closure configuration on vocal efficacy in young normal-speaking women. *J. Voice* 17, 468–480. doi: 10.1067/S0892-1997(03)00065-1

Schoder, S., Junger, C., Weitz, M., and Kaltenbacher, M. (2019). "Conservative source term interpolation for hybrid aeroacoustic computations," in *25th AIAA/CEAS Aeroacoustics Conference, 2019* (Delft: American Institute of Aeronautics and Astronautics Inc., AIAA). doi: 10.2514/6.2019-2538

Schoder, S., Weitz, M., Maurerlehner, P., Hauser, A., Falk, S., Kaltenbacher, M., et al. (2020). Hybrid aeroacoustic approach for the efficient numerical simulation of human phonation. *J. Acoust. Soc. Am*. 147, 1179–1194. doi: 10.1121/10.0000785

Schuberth, S., Hoppe, U., Döllinger, M., Lohscheller, J., and Eysholdt, U. (2002). High-precision measurement of the vocal fold length and vibratory amplitudes. *Laryngoscope* 112, 1043–1049. doi: 10.1097/00005537-200206000-00020

Sciamarella, D., and Le Quéré, P. (2008). Solving for unsteady airflow in a glottal model with immersed moving boundaries. *Eur. J. Mech*. 27, 42–53. doi: 10.1016/j.euromechflu.2007.06.004

Semmler, M., Döllinger, M., Patel, R. R., Ziethe, A., and Schützenberger, A. (2018). Clinical relevance of endoscopic three-dimensional imaging for quantitative assessment of phonation. *Laryngoscope* 128, 2367–2374. doi: 10.1002/lary.27165

Södersten, M., Hertegard, S., and Hammarberg, B. (1995). Glottal closure, transglottal airflow, and voice quality in healthy middle-aged women. *J. Voice* 9, 182–197. doi: 10.1016/S0892-1997(05)80252-8

Södersten, M., and Lindestad, P. Å. (1990). Glottal closure and perceived breathiness during phonation in normally speaking subjects. *J. Speech Lang. Hear. Res.* 33, 601–611. doi: 10.1044/jshr.3303.601

Story, B. H., Titze, I. R., and Hoffman, E. A. (1996). Vocal tract area functions from magnetic resonance imaging. *J. Acoust. Soc. Am.* 100, 537–554. doi: 10.1121/1.415960

Sundberg, J., Fahlstedt, E., and Morell, A. (2005). Effects on the glottal voice source of vocal loudness variation in untrained female and male voices. *J. Acoust. Soc. Am.* 117, 879–885. doi: 10.1121/1.1841612

Sundberg, J., Titze, I., and Scherer, R. (1993). Phonatory control in male singing: a study of the effects of subglottal pressure, fundamental frequency, and mode of phonation on the voice source. *J. Voice* 7, 15–29. doi: 10.1016/S0892-1997(05)80108-0

Tanaka, S., and Gould, W. J. (1985). Vocal efficiency and aerodynamic aspects in voice disorders. *Ann. Otol. Rhinol. Laryngol.* 94, 29–33. doi: 10.1177/000348948509400107

Tao, C., and Jiang, J. J. (2008). A self-oscillating biophysical computer model of the elongated vocal fold. *Comput. Biol. Med.* 38, 1211–1217. doi: 10.1016/j.compbiomed.2008.10.001

Taylor, C. J., Tarbox, G. J., Bolster, B. D., Bangerter, N. K., and Thomson, S. L. (2019). Magnetic resonance imaging-based measurement of internal deformation of vibrating vocal fold models. *J. Acoust. Soc. Am.* 145, 989–997. doi: 10.1121/1.5091009

Thomson, S. L., Mongeau, L., and Frankel, S. H. (2005). Aerodynamic transfer of energy to the vocal folds. *J. Acoust. Soc. Am.* 118, 1689–1700. doi: 10.1121/1.2000787

Thornton, F., Döllinger, M., Kniesburges, S., Berry, D., Alexiou, C., and Schützenberger, A. (2019). Impact of subharmonic and aperiodic laryngeal dynamics on the phonatory process analyzed in *ex vivo* rabbit models. *Appl. Sci.* 9, 1–18. doi: 10.3390/app9091963

Titze, I. R. (1992). Vocal efficiency. *J. Voice* 6, 135–138. doi: 10.1016/S0892-1997(05)80127-4

Titze, I. R. (2000). *Principles of Voice Production*. Iowa City, IA: National Center of Voice and Speech.

Titze, I. R., Maxfield, L., and Palaparthi, A. (2016). An oral pressure conversion ratio as a predictor of vocal efficiency. *J. Voice* 30, 398–406. doi: 10.1016/j.jvoice.2015.06.002

Vaca, M., Cobeta, I., Mora, E., and Reyes, P. (2017). Clinical assessment of glottal insufficiency in age-related dysphonia. *J. Voice* 31:128.e1–128.e5. doi: 10.1016/j.jvoice.2015.12.010

van den Berg, J., Zantema, J. T., and Doornenbal, P. (1957). On the air resistance and the bernoulli effect of the human larynx. *J. Acoust. Soc. Am.* 29, 626–631. doi: 10.1121/1.1908987

Van Hirtum, A., and Pelorson, X. (2017). High-speed imaging to study an auto-oscillating vocal fold replica for different initial conditions. *Int. J. Appl. Mech.* 9:1750064. doi: 10.1142/S1758825117500648

Xue, Q., Mittal, R., Zheng, X., and Bielamowicz, S. (2010). A computational study of the effect of vocal-fold asymmetry on phonation. *J. Acoust. Soc. Am.* 128, 818–827. doi: 10.1121/1.3458839

Xue, Q., Zheng, X., Mittal, R., and Bielamowicz, S. (2014). Subject-specific computational modeling of human phonation. *J. Acoust. Soc. Am.* 135, 1445–1456. doi: 10.1121/1.4864479

Yamauchi, A., Yokonishi, H., Imagawa, H., Sakakibara, K.-,i., Nito, T., Tayama, N., et al. (2014). Age- and gender-related difference of vocal fold vibration and glottal configuration in normal speakers: analysis with glottal area waveform. *J. Voice* 28, 525–531. doi: 10.1016/j.jvoice.2014.01.016

Yamauchi, A., Yokonishi, H., Imagawa, H., Sakakibara, K. I., Nito, T., Tayama, N., et al. (2016). Quantification of vocal fold vibration in various laryngeal disorders using high-speed digital imaging. *J. Voice* 30, 205–214. doi: 10.1016/j.jvoice.2015.04.016

Zañartu, M., Galindo, G. E., Erath, B. D., Peterson, S. D., Wodicka, G. R., and Hillman, R. E. (2014). Modeling the effects of a posterior glottal opening on vocal fold dynamics with implications for vocal hyperfunction. *J. Acoust. Soc. Am.* 136, 3262–3271. doi: 10.1121/1.4901714

Zhang, Y., Jiang, W., Sun, L., Wang, J., Smith, S., Titze, I. R., et al. (2020). A deep-learning based generalized reduced-order model of glottal flow during normal phonation. *arXiv preprint arXiv:2005.11427*.

Zhang, Z. (2019). Compensation strategies in voice production with glottal insufficiency. *J. Voice* 33, 96–102. doi: 10.1016/j.jvoice.2017.10.002

Zhang, Z. (2020). Estimation of vocal fold physiology from voice acoustics using machine learning. *J. Acoust. Soc. Am.* 147, EL264–EL270. doi: 10.1121/10.0000927

Zhang, Z., Kreiman, J., Gerratt, B. R., and Garellek, M. (2012). Acoustic and perceptual effects of changes in body layer stiffness in symmetric and asymmetric vocal fold models. *J. Acoust. Soc. Am.* 133, 453–462. doi: 10.1121/1.4770235

Zhang, Z., Mongeau, L., Frankel, S. H., Thomson, S., and Park, J. B. (2004). Sound generation by steady flow through glottis-shaped orifices. *J. Acoust. Soc. Am.* 116, 1720–1728. doi: 10.1121/1.1779331

Zhang, Z., and Mongeau, L. G. (2006). Broadband sound generation by confined pulsating jets in a mechanical model of the human larynx. *J. Acoust. Soc. Am.* 119, 3995–4005. doi: 10.1121/1.2195268

Zörner, S., Kaltenbacher, M., and Döllinger, M. (2013). Investigation of prescribed movement in fluid-structure interaction simulation for the human phonation process. *Comput. Fluids* 86, 133–140. doi: 10.1016/j.compfluid.2013.06.031

Zörner, S., Šidlof, P., Hüppe, A., and Kaltenbacher, M. (2016). Flow and acoustic effects in the larynx for varying geometries. *Acta Acust. United Acust.* 102, 257–267. doi: 10.3813/AAA.918942

Conflict of Interest: The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

frontiers in Physiology

OPINION
published: 06 April 2021
doi: 10.3389/fphys.2021.658583

The Clinician's Guide to the Machine Learning Galaxy

Lin Shen[1,2], *Benjamin H. Kann*[3,4], *R. Andrew Taylor*[5] *and Dennis L. Shung*[6]*

[1] Department of Medicine, Brigham and Women's Hospital, Boston, MA, United States, [2] Division of Gastroenterology, Hepatology and Endoscopy, Brigham and Women's Hospital, Boston, MA, United States, [3] Department of Radiation Oncology, Dana-Farber Cancer Institute/Brigham and Women's Hospital and Harvard Medical School, Boston, MA, United States, [4] Artificial Intelligence in Medicine Program, Brigham and Women's Hospital, Boston, MA, United States, [5] Department of Emergency Medicine, Yale School of Medicine, New Haven, CT, United States, [6] Section of Digestive Diseases, Department of Medicine, Yale School of Medicine, New Haven, CT, United States

Keywords: artificial intelligence, delivery of health care, machine learning, clinical decision support systems, health care outcome and process assessment

INTRODUCTION

Machine learning has the potential to enhance the practice of medicine (Rajkomar et al., 2019). However, an "AI chasm" has been described that limit the clinical application of machine learning models (Keane and Topol, 2018). Clinicians are domain experts that can help bridge the gap by becoming active partners in developing and implementing machine learning models for clinical use. The paradigm of collaboration between domain experts and machine learning engineers has been successful in developing expert-augmented machine learning (Gennatas et al., 2020). However, it is challenging for interested clinicians to understand the capabilities of machine learning and how to best contribute their domain expertise in designing a machine learning solution.

This is a guide for the clinician interested in helping to design and deploy machine learning solutions to improve clinical care. We propose an approach that finds an area with potential for benefit, considers machine learning as one of several solutions, then counts the cost of a perfectly performing machine learning algorithm to determine if it is worth the effort (**Figure 1**).

Key Terms

Artificial intelligence (AI): Generally, the ability for a computer to accomplish tasks typically associated with human intelligence.

***Machine learning (ML)*:** a subfield of artificial intelligence, broadly refers to the ability of a computational platform to learn from data and make predictions or recommendations based on this data without being explicitly programmed. In general, there are two major categories of machine learning, supervised and unsupervised. *Supervised learning* is conducted with the concept of "truth" where the model tries to approximate the relationship between inputs and labeled outputs. For example, given images of cats and dogs, where each image has a correct answer, can you train a model that accurately identifies of cats versus dogs? *Unsupervised learning* is performed without data labels and the goal is for the computer to infer inherent structure or patterns in the data. For example, given a set of heart rate, accelerometer, and location data from a wearable fitness monitor, can the computer identify periods of rest versus exercise based on differences in the raw data?

Neural networks (NN): a form of machine learning with a basic architecture consisting of nodes and connections existing in multiple layers, loosely analogous to neurons and synapses in the biological brain. This broad category is inclusive of many kinds of modern machine learning models which are used in tasks such as computer vision, voice recognition, bioinformatics, and among others.

Deep learning: A broad family of neural network architectures that have multiple layers (aka deep).

Edited by:
Gary An,
University of Vermont Larner College of Medicine, United States

Reviewed by:
Joakim Sundnes,
Simula Research Laboratory, Norway
Tuhin K. Roy,
Mayo Clinic, United States

***Correspondence:**
Dennis L. Shung
dennis.shung@yale.edu

Specialty section:
This article was submitted to Computational Physiology and Medicine, a section of the journal Frontiers in Physiology

Received: *26 January 2021*
Accepted: *10 March 2021*
Published: *06 April 2021*

Citation:
Shen L, Kann BH, Taylor RA and Shung DL (2021) The Clinician's Guide to the Machine Learning Galaxy. Front. Physiol. 12:658583. doi: 10.3389/fphys.2021.658583

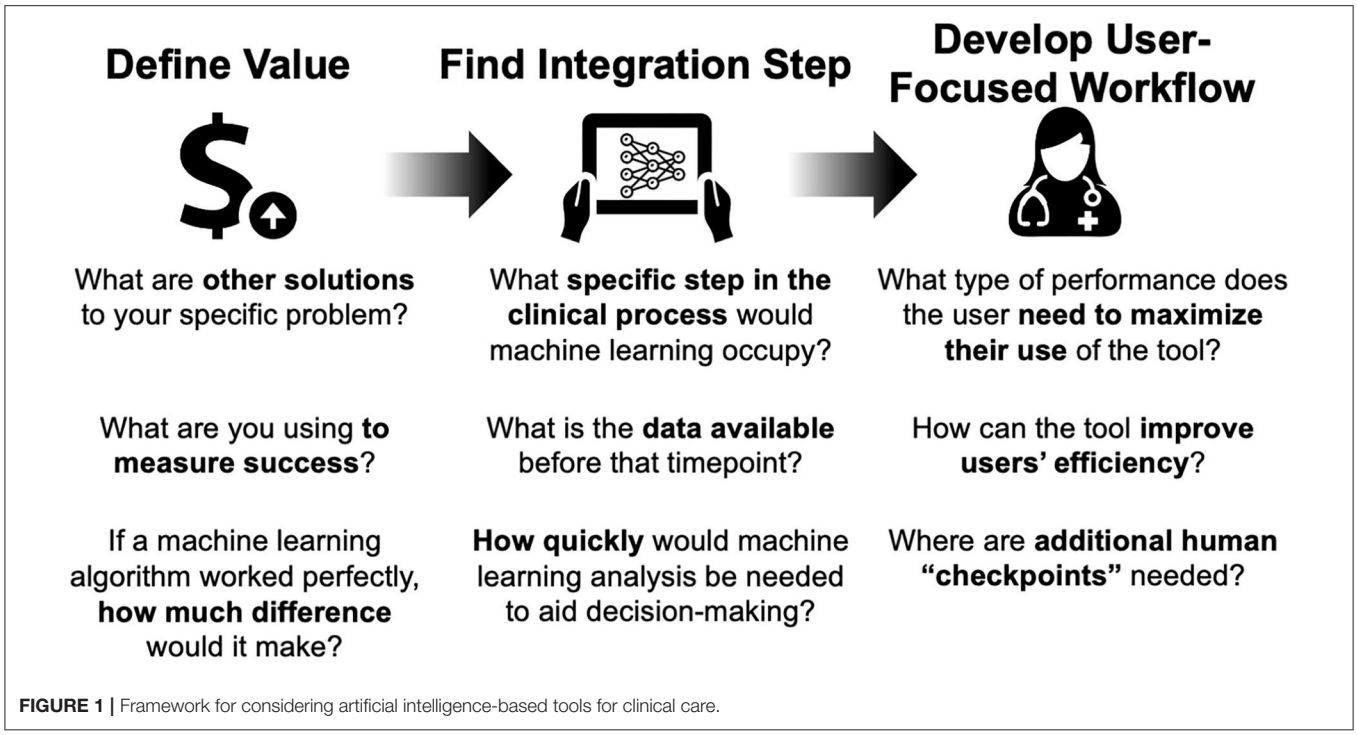


FIGURE 1 | Framework for considering artificial intelligence-based tools for clinical care.

KEY QUESTIONS

For the interested clinician, these following self-assessment questions may help in determining whether a machine learning tool makes sense for your specific scenario.

What Is My Unmet Need?

For machine learning to make a positive impact on patient care, finding the right use case is the place to start. As a practicing clinician, this should draw from your understanding of the clinical workflow and impact on patient care. A proposed paradigm is starting with a larger problem, mapping out the workflow, and identifying areas in need of improvement.

Is Machine Learning Useful for This Need?

Critically consider if machine learning is the best tool to improve that specific area. Consider other solutions involving personnel, workflow, or policy changes. If an information technology solution is the best answer, consider its impact on the workflow in the best case scenario. This depends on what is important for each clinical scenario: accuracy, timeliness, or reliability. If even the best case scenario leads to minimal improvement and significant changes in the workflow (with attendant costs), machine learning may not be the best solution. Consider other solutions involving personnel, rules-based systems, or process redesign.

Are You Asking the Right Question to Put the ML Tool in the Highest Value Use Within the Clinical Workflow?

In order to do this effectively, first find the right use case (e.g., right information to the right person at the right time). Next, figure out where the model fits into the clinical pathway, which includes process mapping to understand types of input data needed and output desired. Finally, consider the workflow and needs of the end-user, including timeliness.

Should Computer Simulation Be Considered in the Development Process for the ML Tool?

Depending on the deployment setting, the ML tool may benefit from data augmentation to improve generalizability, particularly if the tool is to be applied across different radiological, electronic health record, or genomic platforms. This can be achieved with generation of synthetic data or techniques of data transformation. These are methods where data is artificially manufactured rather than the result of real-world measurement. This approach can sometimes be used judiciously to augment real-world data in scenarios where real-world data is sparse or difficult to obtain. Data can be either created de novo based on a set of criteria or by digitally manipulating real-world measured data. This approach should be used cautiously due to multiple tricky considerations including bias and generalizability.

CASE STUDIES

Medical Imaging Perspective

Radiographic medical imaging, whether CT, MRI, Ultrasound, or other modality, is an ever-growing source of big data in healthcare. Medical imaging began as a field in which advanced technology was used to generate visual data that could be analyzed and assessed qualitatively by a clinician. As the field evolved, quantitative imaging metrics were developed to assist

with image interpretation and management decisions (Giger et al., 2008). The advent of computer-aided diagnosis and detection in the 1980's and 1990's brought early machine learning techniques to the medical imaging field with important applications in breast cancer mammography and ultrasound (Jiang et al., 1999; Freer and Ulissey, 2001). Over the past decade, the emergence of deep learning neural networks has generated a tremendous amount of attention in the field of medical image analysis for its transformative potential (Ker et al., 2018). Deep learning utilizes raw pixel or voxel input from images and feeds them through progressively more complex layers of a neural network to generate an output prediction. Through an iterative training process, millions of mathematical parameters of a neural network are optimized such that input images fed into the network generate predictions that best fit the true output. Rather than rely on user input to pre-engineer and determine appropriate features for the machine learning model, deep learning utilizes raw imaging data to "learn" the features that optimize predictive performance. Unlike prior forms of computer-aided analysis, deep learning has the potential to form end-to-end prediction models encompassing multiple parallel or sequential imaging tasks, including object segmentation, detection, and identification. Deep learning has the potential to affect medical imaging in healthcare by (1) improving diagnostic efficiency and achieving cost savings by freeing up limited human resources, (2) augmenting human performance at diagnostic prediction in challenging scenarios, and (3) discerning previously impossible-to-discern patterns and predictions from imaging data.

Case Study: Lung Cancer Screening

Lung cancer is the leading cause of cancer death in the United States. Lung cancer screening with low-dose CT has been shown to reduce mortality and is currently recommended routinely for high-risk individuals (National Lung Screening Trial Research et al., 2011; de Koning et al., 2020). Despite imaging guidelines for lung cancer detection, there remains significant concern surrounding inter-rater variability, and false-positive and negative rates (Field et al., 2016). Additionally, uptake of CT screening, even among high risk populations has remained extremely low, in part, owing to lack of high-volume radiology center resources (Jemal and Fedewa, 2017). Given these challenges, there is a unique opportunity to explore machine learning to improve accuracy of detection and access to screening. In embarking on an investigation of machine learning for lung cancer screening, the following should be considered:

What is the goal and what is the metric of success? The ultimate goal and metric for success may not be the same thing, particularly in initial phases of algorithm development. The ultimate goal should reflect clinically meaningful endpoints: improving patient survival, quality of life, or healthcare costs. The metric for success often begins more narrowly. In the case of lung cancer screening, percent accuracy, sensitivity, specificity, and the area under the ROC curve in identifying a lung nodule as benign or malignant may be appropriate. Ultimately, as study progresses, metrics should move beyond accuracy. Direct measurement of clinical meaningful endpoints, such as survival, morbidity, and quality of life should be incorporated into clinical trials of the application.

What type of machine learning is optimal for the task? The type of machine learning utilized will be driven by the medical imaging task, however, in general, convolutional neural network-based deep learning architectures are the current gold standard for image analysis. In simplistic terms, a convolutional neural network takes images as input data, and applies various filters which manipulate the image to extract features. This is analogous to image filters you can use in photo manipulation software or various social media programs. Some filters may enhance borders or edges, others may detect specific colors or brightness levels. This strategy is used in a neural network where the final output is based on extracting meaningful features from the images and making decisions based on those features. Older methods utilizing pre-engineered radiomic features may be suitable for certain classification problems where the image region of interest is well-defined, but deep learning has the ability to both localize an object (in this case lung nodule) *and* classify it (malignant vs. benign). Deep learning is particularly well-suited for this "end-to-end" task completion. Several studies have shown extremely high accuracy of lung nodule and malignancy prediction using a deep learning based approach to CT diagnosis (Field et al., 2016; Jemal and Fedewa, 2017; Kang et al., 2017; Causey et al., 2018; Ardila et al., 2019).

What type of data is needed? Data collection, curation, and annotation are perhaps the most critical aspects of training a successful machine learning algorithm. As the approach shifts from simpler machine learning models to more complex models such as deep learning neural networks, the quantity and quality of data becomes increasingly important. For lung cancer screening, this means access to thousands of CT scans that have been pre-labeled by human experts. Each nodule should have been identified and should have associated with it a "ground truth" label. For an image localization task, this label itself would be a segmented region of interest encompassing the nodule. For malignancy classification, this label could be binary (malignant or benign) or ordinal (suspicion of malignancy on a scale of 1 through 5), depending on how the labeling was performed. Because many imaging-based ML algorithms are prone to overfitting training data, all models must be validated on external datasets, ideally representative of the target scenario for which the algorithm is being developed. Particular considerations for medical images are type of CT scanner, use of contrast agent, image resolution, and artifact. These parameters must be explored and addressed in preprocessing steps and/or validation datasets prior to implementation of an imaging-based ML application.

What is the role of simulated, or synthetic, data? Successful ML development in medicine requires large, high-quality, annotated, and accessible datasets, which are often lacking (Emanuel and Wachter, 2019). A key strategy to mitigate data limitations is the use of data augmentation techniques to create simulated, or synthetic, data to bolster the training process. By applying image transformations, from simple rotations, flips, or deformations to more advanced ML-driven augmentation, model generalization can be improved dramatically even when

training on relatively small datasets (Goel et al., 2020). This is accomplished by introducing transforms that mimic confounding variations expected of data samples encountered in real-world testing, but that are not themselves features that predict a particular data class.

Where does the model fit into the clinical pathway? The ultimate utility of an ML-based healthcare application like lung cancer screening will not be decided by AUC or accuracy, but by clinically meaningful endpoints, such as decreased mortality, treatment-related morbidity, and healthcare resource burden. To maximize the potential utility of the algorithm, it must be determined how the model can best fit into the clinical pathway by considering timing, physical space, costs, user interface, and responsibility. In the context of lung cancer screening, for example, an algorithm could be executed automatically at the time of scan or by the radiologist during review. The former could improve resource allocation by flagging abnormal scans for expedited review, but the latter would allow for human oversight of the algorithm with less risk of bias. On the other hand, incorporation at the time of radiologist review would necessitate a streamlined user interface that does not compromise efficiency. Simple workflow decisions such as this can also have profound implications for responsibility, trust, and decision-making and raise medico-legal issues. If an algorithm triages patients incorrectly to the reviewing radiologist, who is liable for this error? These subtle implementation characteristics represent significant barriers to entry to real-world clinical use, but are often overlooked in early stages of algorithm development. These factors should be considered (and reconsidered) at each stage of algorithm development, even at model conception.

Ambulatory Provider Perspective

A major advantage of machine learning algorithms is the ability to process large amounts of data in a relatively short amount of time. For an ambulatory provider, this advantage can translate into individualized decision-making by using a model that incorporates relevant variables beyond traditional population-based risk factors. For example, primary care providers often use a clinical decision support tool to recommend initiation of a statin for appropriate patients during routine office visits. Traditional models such as the Atherosclerotic Cardiovascular Disease (ASCVD) risk score uses conventional statistical methods from a population that may not be a good representation for all patients, particularly since risk of disease and treatment guidelines vary among patients of different ethnicities (McCredie et al., 1990; Norwood et al., 2009; Lloyd-Jones et al., 2017; Das et al., 2018; Volgman et al., 2018; Damask et al., 2020).

Case Study: Polygenic Risk Scoring

To better understand differences between individuals of different ethnicities, polygenic risk scoring estimates the predisposition of disease using the presence or absence of known disease-associated genes (Damask et al., 2020). This holds the promise of generating more accurate predictions by using genotypic data in conjunction with other clinical and environmental variables. As a clinician interested in implementing such a model into live practice, what are the important specifics to consider?

Machine learning models can process a large number of variables that are also very different from one another. In order to handle the variety of data, data management is critical during the early stages of planning. Effective data management considers (1) data type, (2) data reliability, and (3) the sample size.

In regards to data types, the inputs used to generate a model can come in various forms. One of the major advantages of newer machine learning models over traditional statistical models is increased flexibility to take different types of inputs. This can range from simple mixing of categorical vs. continuous variables to handling high dimensional complex inputs such as raw imaging, video, audio, or even genome sequencing data. Another advantage of handling multiple data types is that one can imagine a machine learning pipeline that utilizing several layers of processing while appearing seamless to the end user. If a data type is not readily available in modern EHRs but is of critical importance, it should be considered for integration as part of future policy/health IT infrastructure development. For example, in order to fully utilize genomic risk prediction, sequencing data must be available. At present, most genomic sequencing is often done for a specific panel of genes and the results are often saved as a report in the EHR. The actual genomic data is not saved as most mainstream EHRs lack the capability to store this type of data.

When considering data reliability, the electronic health record (EHR) is a rich source of potential data but most clinicians recognize that there is a wide range to the reliability and accuracy of EHR data. Some data types are structured (for example a hemoglobin A1C laboratory value), meaning both the value and the context are discretely defined. Structured data are more easily accepted by machine learning algorithms with less preprocessing needed. These types of data fall on the more reliable end of the spectrum. Diagnosis and billing codes are also structured, in that the value and context are clearly defined, but most clinicians understand that they are limited terms of accurate representation of the patient. Fully unstructured data include data types such as notes. Notes are often considered the most representative of clinical truth in the EHR, but often can still contain errors. As unstructured data, notes are difficult for machine learning models to accept as input without preprocessing.

A challenge for generalizability in polygenic risk scores is the heterogeneity of available data across electronic health record systems that vary across institutions and health systems, and scarcity of fully annotated genomic datasets. Two promising approaches have shown the ability to generate synthetic data by characterizing different data distributions in electronic health record data and genomic datasets using generative adversarial neural networks and ordinary differential equation-based models (Fratello et al., 2015; Baowaly et al., 2019).

A critical question for data scientists is identifying what types of data is important to include into a model. Clinicians can inform data scientists about important concepts to include and point them to the best sources of data to represent those concepts. This often draws on the clinician's medical knowledge and may mirror their own human analytic process when making clinical decisions. For example, clinicians understand that diabetes is an

important risk factor to include in models for cardiovascular disease. Therefore, an important concept to identify is the presence of "diabetes." However, data that could represent diabetes include laboratory values, clinical documentation, billing codes, and among others. The ultimate decision on which to use (including combinations) is best made in conjunction with a clinician who understands the medical considerations, the workflow considerations, and the data considerations as discussed above. Once there is a thoughtful strategy on what are the best sources of data for specific concepts, advanced ML techniques can be employed to obtain more difficult to extract data if necessary.

For example, a common challenge in utilizing the EHR is how to best utilize clinical notes, where information largely exists as free-text. One approach to make use of unstructured free-text data is natural language processing. Natural language processing models can capture specific meaning and interpret intent based on not just terms but context. A combination of ML models optimized for specific tasks can be integrated into a larger model either directly or through a series of preprocessing steps where the output is used in a subsequent ML model. Specific ML models may be optimized for natural language processing of notes, or detect polygenic risk of cardiovascular disease from genomic data. In the prior example, NLP may be used to identify the concept of "poor adherence to insulin" from clinical notes whereas a genomic risk factor model may find mutations that confer risk of developing type II diabetes. These data points can subsequently integrate into a model that accounts for environmental exposures such as smoking and other clinical risk factors like obesity.

The last major consideration is the number of cases or patients with the relevant data available. While most of the current excitement is over deep learning or neural networks, these types of machine learning techniques require large numbers of examples to train. Other forms of machine learning can perform well with smaller training samples, and some approaches handle missing data better than others, which is a frequent occurrence when working with clinical variables. Lastly, understanding the population characteristics can be helpful when selecting good machine learning model candidates. Like all prediction modeling, incidence and prevalence are important considerations when attempting classification tasks. For example, rare events can be more difficult for machine learning to predict, and data scientists often address issues related to class balance when building the model. If a machine learning model tried to predict whether you would win the lottery, and just predicted that nobody would win the lottery ever, it would be right the majority of the time and be very "high performing" from an accuracy perspective.

Proceduralist Perspective

Real-time deep learning-based computer vision can also enhance the performance of the proceduralist by providing visual enhancement of anatomy and pathology. These can be overlaid directly onto images collected during the procedure, whether from laparoscopic surgery or diagnostic endoscopy. The algorithms could provide optical biopsies, map out anatomical boundaries and tissue planes, and identify abnormal areas relevant to the particular operative procedure.

Case Study: Colonoscopy With Computer-Aided Detection

For a gastroenterologist, finding precancerous lesions is the top priority to prevent colorectal cancer. In order to measure the success in preventing colorectal cancer that develops before the recommended next colonoscopy, gastroenterologists have traditionally used adenoma detection rate (ADR) as a proxy indicator for high quality colonoscopy as a 1% increase in ADR in correlated with a 3% decreased risk of interval colorectal cancer. The endoscopist can track their adenoma detection rate, and if it is lower than expected could undergo additional training to improve their ability to detect pre-cancerous lesions. However, adenoma detection rate has wide variation across endoscopists, and a tool that would standardize the performance of endoscopists would help decrease the incidence of preventable colorectal cancer. Recent advances in deep learning through convoluted neural networks have led to high-performing algorithms that hold promise in enhancing endoscopist performance by identifying polyps in real-time colonoscopy videos and detecting adenomas, which can increase the adenoma detection rates for all endoscopists (Misawa et al., 2018; Urban et al., 2018; Wang et al., 2018, 2019, 2020; Gong et al., 2020).

As a clinician interested in developing or implementing deep learning tools to improve the adenoma detection rate, how should you think about the approach?

First, identify the inputs (e.g., images or video) to train the model, which in this case would be deep convolutional neural networks described earlier in section Medical Imaging Perspective. If the model is meant to detect polyps, the ideal input would be colonoscopy videos with labeled images of "polyp present" and where in the frame the polyp is located. This is the rate-limiting step, since labeling is human capital-intensive, deep learning requires numerous examples, and the algorithms learn explicitly from the label of "polyp present" or "polyp absent."

In this particular task, data transformation has been considered due to the relative absence of large annotated image databases of polyps. These approaches have included changing the image dimensions, changing pixel values, and adding in external conditions with the goal to maintain or achieve better generalizability for ML tools to detect polyps (Sánchez-Peralta et al., 2020).

As with all supervised machine learning, labels must be present in the data to train the algorithm, which can sometimes be costly as content experts are needed to create the labels. Furthermore, one key challenge is to make sure the algorithm can perform well in other datasets, referred to as "robustness," such as in real time for a new procedure. In this case, the specific way the data is captured may affect the algorithm performance. For example, if the algorithm is to be used in a practice with high definition endoscopes that have specific image processing settings (e.g., narrow band imaging), the input data should ideally be captured from that specific brand of endoscope and also include images with the specific setting. A clinician is critical in informing

the data scientist the parameters of the data used during the procedure so that adequate data of sufficient quality is collected to train the algorithm.

As the model is developed, the issue of timeliness and workflow is highlighted as a key area for clinician involvement (Shung and Byrne, 2020). This is particularly relevant for endoscopic units in ambulatory surgical centers, where the trend toward lower reimbursement for endoscopy have led to the development of performance metrics to enhance efficiency (Gellad et al., 2013). Proceduralists provide crucial information about the existing clinical process to guide how software should be designed. The user needs of the endoscopist must be considered, particularly the tolerance for false positives and the impact of the software on efficiency (i.e., duration of the procedure). Since a colonoscopy procedure involves diagnosis, assessment, and treatment (find the polyps, assess if they are problematic, and remove them), real-time processing is a prerequisite to any software solution. For high volume ambulatory surgery centers, algorithms must have minimal impact the amount of time to perform procedures. Clinicians' preferences and insight into the workflow of how the deep learning software enhances the user experience and performance are key in optimization, in this case providing real-time recommendations that do not unnecessarily prolong the overall procedure.

Finally, as these algorithms are implemented, clinicians have an important role in providing feedback. User assessments and improvements in the interface for each iteration of the software implementation. If there are clear discrepancies in what the software detects and provider assessment, quality control is crucial to maintain provider confidence in the software recommendations.

Limitations and Additional Considerations

While these scenarios delve into specific ways in which clinicians can inform the development and validation of ML tools in clinical care, the potential applications of AI in healthcare go from individualized recommendations with personalized medicine to informing policy in public health. A discussion of all the potential applications is beyond the scope of this article, but a comprehensive compilation of AI and ML-based medical devices approved by regulatory bodies in the United States and Europe provide a glimpse into the personalization of care (Muehlematter et al., 2021), while another article delves into the ways in which AI and ML can be used across populations to tailor policies to promote health, protect health, and improve the efficiency of services for communities within the greater population (Panch et al., 2019).

Limitations for integrating ML tools into clinical care can broadly fall under data maintenance and real-world deployment.

Bias, heterogeneity, and gaps in data can lead to poor performance or contribute toward perpetuating disparities or harmful discriminatory practices. Indeed, a prominent recent example was the Amazon AI recruitment tool that was deactivated after it showed bias against hiring women (Dastin, 2018). A new concept of algorithmic stewardship addresses the limitations of constantly changing sources and storage of healthcare data by monitoring, correcting, and updating the dataflow to accurately reflect different ways of data capture as well as practice patterns or epidemiological shifts (Eaneff et al., 2020). Data equity and representation is a key limitation that should be actively addressed with the development of any ML tool to ensure that inherent health inequities, such as race correction, will not be perpetuated (Vyas et al., 2020).

Generalizability and interpretability are two key limitations that can hamper real-world deployment of ML tools. For clinicians, the focus is on the individual patient, which requires that the algorithm performs well and does not generate an erroneous result. For deep learning tools in particular, the key limitation of overfitting due to complexity of the network architecture and large number of parameters must be addressed with rigorous validation on multiple datasets representative of real world data. This is analogous to training a robot to play tennis only on a clay court, and then deploying the robot to play on the grass courts of Wimbledon. Since clinicians are experts with advanced training, the need to trust and verify the ML tool output is key to ensure that the ML tools are used in clinical practice. For this, a measure of interpretability is important so that ML tools can complement the professional authority of clinical providers (Kelly et al., 2019).

The use of AI with its dependence on data also introduces additional risks into the healthcare environment with regards to ethical, regulatory, and legal issues. Privacy compliance, the role of the algorithm in shared patient-provider decision making, data access, system failures, computer viruses/malware, and intentional adversarial attacks geared toward machine learning models require additional strategies to mitigate risk for patients when considering the use of ML in medicine (Finlayson et al., 2019). Ethical research methodology, including fairness and equity for both representation in the data used for the algorithms and in sharing the benefits realized by the algorithm, must be practiced when using patient data. Clinician researchers adept in these consideration can help guide data scientists in this regard. We recommend consultation with institutional review boards (IRB) for all projects related to patient data to ensure appropriateness and proper protection of patients. Prior to commercialization and deployment of informatics based tools in patient care, approval from regulatory bodies may be necessary. The regulatory guidelines continue to evolve in the United States with the FDA, the European Union with General Data Protection and Regulation framework, and internationally through the International Medical Device Regulators Forum. For the FDA, ML algorithms have been assessed in a similar fashion to medical devices, although there is now a growing recognition that software-based products are a unique category within that track.

CONCLUSIONS

Machine learning integrated medicine is the future of patient care. Analytic tools to take full advantage of an increasingly information-dense practice environment, but clinicians are critical partners in developing successful ML models that can be integrated into real-world patient care. While data scientists are experts in the technical aspects of machine learning, clinicians are needed to identify the appropriate settings for ML solutions,

the best data to use to help shape model development, the best integration point into a real world workflow environment, and the final usability of the tool.

AUTHOR CONTRIBUTIONS

LS came up with concepts and critically revised the manuscript. BK wrote substantive parts of the manuscript and critically revised the manuscript. RT critically revised the manuscript. DS wrote the manuscript, contributed concepts, and critically revised the manuscript. All authors contributed to the article and approved the submitted version.

FUNDING

DS was funded by NIH T32 DK007017. BK was funded by NIH/NIDCR grant K08 DE030216.

ACKNOWLEDGMENTS

We appreciate feedback from Allen Hsiao, MD, Vice President; Chief Medical Informatics Officer, Yale-New Haven Health and Adam Landman MD, MS, MIS, MHS Vice President; Chief Information Officer; Digital Innovation Officer, Brigham Health for the ideas in the manuscript.

REFERENCES

Ardila, D., Kiraly, A. P., Bharadwaj, S., Choi, B., Reicher, J. J., Peng, L., et al. (2019). End-to-end lung cancer screening with three-dimensional deep learning on low-dose chest computed tomography. *Nat. Med.* 25, 954–961. doi: 10.1038/s41591-019-0447-x

Baowaly, M. K., Lin, C. C., Liu, C. L., and Chen, K. T. (2019). Synthesizing electronic health records using improved generative adversarial networks. *J. Am. Med. Inform. Assoc.* 26, 228–241. doi: 10.1093/jamia/ocy142

Causey, J. L., Zhang, J., Ma, S., Jiang, B., Qualls, J. A., Politte, D. G., et al. (2018). Highly accurate model for prediction of lung nodule malignancy with CT scans. *Sci. Rep.* 8:9286. doi: 10.1038/s41598-018-27569-w

Damask, A., Steg, P. G., Schwartz, G. G., Szarek, M., Hagstrom, E., Badimon, L., et al. (2020). Patients with high genome-wide polygenic risk scores for coronary artery disease may receive greater clinical benefit from alirocumab treatment in the ODYSSEY OUTCOMES Trial. *Circulation* 141, 624–636. doi: 10.1161/CIRCULATIONAHA.119.044434

Das, S. R., Everett, B. M., Birtcher, K. K., Brown, J. M., Cefalu, W. T., Januzzi, J. L. Jr., et al. (2018). 2018 ACC expert consensus decision pathway on novel therapies for cardiovascular risk reduction in patients with type 2 diabetes and atherosclerotic cardiovascular disease: a report of the american college of cardiology task force on expert consensus decision pathways. *J. Am. Coll. Cardiol.* 72, 3200–3223. doi: 10.1016/j.jacc.2018.09.020

Dastin, J. (2018). *Amazon Scraps Secret AI Recruiting Tool That Showed Bias Against Women*. San Francisco, CA: Reuters.

de Koning, H. J., van der Aalst, C. M., de Jong, P. A., Scholten, E. T., Nackaerts, K., Heuvelmans, M. A., et al. (2020). Reduced lung-cancer mortality with volume CT screening in a randomized trial. *N. Engl. J. Med.* 382, 503–513. doi: 10.1056/NEJMoa1911793

Eaneff, S., Obermeyer, Z., and Butte, A. J. (2020). The case for algorithmic stewardship for artificial intelligence and machine learning technologies. *JAMA* 324, 1397–1398. doi: 10.1001/jama.2020.9371

Emanuel, E. J., and Wachter, R. M. (2019). Artificial intelligence in health care: will the value match the hype? *JAMA* 321, 2281–2282. doi: 10.1001/jama.2019.4914

Field, J. K., Duffy, S. W., Baldwin, D. R., Whynes, D. K., Devaraj, A., Brain, K. E., et al. (2016). UK Lung Cancer RCT Pilot Screening Trial: baseline findings from the screening arm provide evidence for the potential implementation of lung cancer screening. *Thorax* 71, 161–170. doi: 10.1136/thoraxjnl-2015-207140

Finlayson, S. G., Bowers, J. D., Ito, J., Zittrain, J. L., Beam, A. L., and Kohane, I. S. (2019). Adversarial attacks on medical machine learning. *Science* 363, 1287–1289. doi: 10.1126/science.aaw4399

Fratello, M., Serra, A., Fortino, V., Raiconi, G., Tagliaferri, R., and Greco, D. (2015). A multi-view genomic data simulator. *BMC Bioinformat.* 16:151. doi: 10.1186/s12859-015-0577-1

Freer, T. W., and Ulissey, M. J. (2001). Screening mammography with computer-aided detection: prospective study of 12,860 patients in a community breast center. *Radiology* 220, 781–786. doi: 10.1148/radiol.2203001282

Gellad, Z. F., Thompson, C. P., and Taheri, J. (2013). Endoscopy unit efficiency: quality redefined. *Clin. Gastroenterol. Hepatol.* 11:1046–9.e1. doi: 10.1016/j.cgh.2013.06.005

Gennatas, E. D., Friedman, J. H., Ungar, L. H., Pirracchio, R., Eaton, E., Reichmann, L. G., et al. (2020). Expert-augmented machine learning. *Proc. Natl. Acad. Sci. U. S. A.* 2020:201906831. doi: 10.1073/pnas.1906831117

Giger, M. L., Chan, H. P., and Boone, J. (2008). Anniversary paper: history and status of CAD and quantitative image analysis: the role of Medical Physics and AAPM. *Med. Phys.* 35, 5799–5820. doi: 10.1118/1.3013555

Goel, K., Gu, A., Li, Y., and Ré, C. (2020). Model patching: closing the subgroup performance gap with data augmentation. *arXiv*. arXiv:2008.06775.

Gong, D., Wu, L., Zhang, J., Mu, G., Shen, L., Liu, J., et al. (2020). Detection of colorectal adenomas with a real-time computer-aided system (ENDOANGEL): a randomised controlled study. *Lancet Gastroenterol. Hepatol.* 5, 352–361. doi: 10.1016/S2468-1253(19)30413-3

Jemal, A., and Fedewa, S. A. (2017). Lung cancer screening with low-dose computed tomography in the United States-2010 to 2015. *JAMA Oncol.* 3, 1278–1281. doi: 10.1001/jamaoncol.2016.6416

Jiang, Y., Nishikawa, R. M., Schmidt, R. A., Metz, C. E., Giger, M. L., and Doi, K. (1999). Improving breast cancer diagnosis with computer-aided diagnosis. *Acad. Radiol.* 6, 22–33. doi: 10.1016/S1076-6332(99)80058-0

Kang, G., Liu, K., Hou, B., and Zhang, N. (2017). 3D multi-view convolutional neural networks for lung nodule classification. *PLoS ONE* 12:e0188290. doi: 10.1371/journal.pone.0188290

Keane, P. A., and Topol, E. J. (2018). With an eye to AI and autonomous diagnosis. *NPJ Digit. Med.* 1:40. doi: 10.1038/s41746-018-0048-y

Kelly, C. J., Karthikesalingam, A., Suleyman, M., Corrado, G., and King, D. (2019). Key challenges for delivering clinical impact with artificial intelligence. *BMC Med.* 17:195. doi: 10.1186/s12916-019-1426-2

Ker, J., Wang, L., Rao, J., and Lim, T. (2018). Deep learning applications in medical image analysis. *IEEE Access.* 6, 9375–9389. doi: 10.1109/ACCESS.2017.2788044

Lloyd-Jones, D. M., Morris, P. B., Ballantyne, C. M., Birtcher, K. K., Daly, D. D. Jr., DePalma, S. M., et al. (2017). 2017 Focused update of the 2016 ACC expert consensus decision pathway on the role of non-statin therapies for LDL-cholesterol lowering in the management of atherosclerotic cardiovascular disease risk: a report of the American College of Cardiology Task Force on Expert Consensus Decision Pathways. *J. Am. Coll. Cardiol.* 70, 1785–1822. doi: 10.1016/j.jacc.2017.07.745

McCredie, M., Coates, M. S., and Ford, J. M. (1990). Cancer incidence in migrants to New South Wales. *Int. J. Cancer.* 46, 228–232. doi: 10.1002/ijc.2910460214

Misawa, M., Kudo, S.-E., Mori, Y., Cho, T., Kataoka, S., Yamauchi, A., et al. (2018). Artificial intelligence-assisted polyp detection for colonoscopy: initial experience. *Gastroenterology* 154, 2027–2029.e3. doi: 10.1053/j.gastro.2018.04.003

Muehlematter, U. J., Daniore, P., and Vokinger, K. N. (2021). Approval of artificial intelligence and machine learning-based medical devices in the USA and Europe (2015–20): a comparative analysis. *Lancet Digit. Health* 3, e195–e203. doi: 10.1016/S2589-7500(20)30292-2

National Lung Screening Trial Research, T., Aberle, D. R., Adams, A. M., Berg, C. D., Black, W. C., Clapp, J. D., et al. (2011). Reduced lung-cancer mortality with low-dose computed tomographic screening. *N. Engl. J. Med.* 365, 395–409. doi: 10.1056/NEJMoa1102873

Norwood, M. G., Mann, C. D., Hemingway, D., and Miller, A. S. (2009). Colorectal cancer: presentation and outcome in British South Asians. *Colorectal. Dis.* 11, 745–749. doi: 10.1111/j.1463-1318.2008.01675.x

Panch, T., Pearson-Stuttard, J., Greaves, F., and Atun, R. (2019). Artificial intelligence: opportunities and risks for public health. *Lancet Digit. Health* 1, e13–e4. doi: 10.1016/S2589-7500(19)30002-0

Rajkomar, A., Dean, J., and Kohane, I. (2019). Machine learning in medicine. *N. Engl. J. Med*. 380, 1347–1358. doi: 10.1056/NEJMra 1814259

Sánchez-Peralta, L. F., Picón, A., Sánchez-Margallo, F. M., and Pagador, J. B. (2020). Unravelling the effect of data augmentation transformations in polyp segmentation. *Int. J. Comput. Assist. Radiol. Surg.* 15, 1975–1988. doi: 10.1007/s11548-020-02262-4

Shung, D. L., and Byrne, M. F. (2020). How artificial intelligence will impact colonoscopy and colorectal screening. *Gastrointest. Endosc. Clin. N Am*. 30, 585–595. doi: 10.1016/j.giec.2020.02.010

Urban, G., Tripathi, P., Alkayali, T., Mittal, M., Jalali, F., Karnes, W., et al. (2018). Deep learning localizes and identifies polyps in real time with 96% accuracy in screening colonoscopy. *Gastroenterology* 155, 1069–1078.e8. doi: 10.1053/j.gastro.2018.06.037

Volgman, A. S., Palaniappan, L. S., Aggarwal, N. T., Gupta, M., Khandelwal, A., Krishnan, A. V., et al. (2018). Atherosclerotic cardiovascular disease in South Asians in the United States: epidemiology, risk factors, and treatments: a scientific statement from the American Heart Association. *Circulation* 138, e1–e34. doi: 10.1161/CIR.0000000000000600

Vyas, D. A., Eisenstein, L. G., and Jones, D. S. (2020). Hidden in plain sight — reconsidering the use of race correction in clinical algorithms. *N. Engl. J. Med*. 383, 874–882. doi: 10.1056/NEJMms 2004740

Wang, P., Berzin, T. M., Glissen Brown, J. R., Bharadwaj, S., Becq, A., Xiao, X., et al. (2019). Real-time automatic detection system increases colonoscopic polyp and adenoma detection rates: a prospective randomised controlled study. *Gut* 68:1813. doi: 10.1136/gutjnl-2018-317500

Wang, P., Liu, X., Berzin, T. M., Glissen Brown, J. R., Liu, P., Zhou, C., et al. (2020). Effect of a deep-learning computer-aided detection system on adenoma detection during colonoscopy (CADe-DB trial): a double-blind randomised study. *Lancet Gastroenterol. Hepatol.* 5, 343–351. doi: 10.1016/S2468-1253(19)30411-X

Wang, P., Xiao, X., Glissen Brown, J. R., Berzin, T. M., Tu, M., Xiong, F., et al. (2018). Development and validation of a deep-learning algorithm for the detection of polyps during colonoscopy. *Nat. Biomed. Eng.* 2, 741–748. doi: 10.1038/s41551-018-0301-3

Conflict of Interest: The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

frontiers in Physiology

ORIGINAL RESEARCH
published: 19 May 2021
doi: 10.3389/fphys.2021.662845

Utilizing the Heterogeneity of Clinical Data for Model Refinement and Rule Discovery Through the Application of Genetic Algorithms to Calibrate a High-Dimensional Agent-Based Model of Systemic Inflammation

Chase Cockrell and Gary An*

Departmen of Surgery, Larner College of Medicine, The University of Vermont, Burlington, VT, United States

OPEN ACCESS

Edited by:
Fernando Soares Schlindwein, University of Leicester, United Kingdom

Reviewed by:
Edward C. Stites, Washington University in St. Louis, United States
Clarissa Lim Velayo, University of the Philippines Manila, Philippines

****Correspondence:***
Chase Cockrell
robert.cockrell@med.uvm.edu

Specialty section:
This article was submitted to Computational Physiology and Medicine, a section of the journal Frontiers in Physiology

Received: *01 February 2021*
Accepted: *27 April 2021*
Published: *19 May 2021*

Citation:
Cockrell C and An G (2021) Utilizing the Heterogeneity of Clinical Data for Model Refinement and Rule Discovery Through the Application of Genetic Algorithms to Calibrate a High-Dimensional Agent-Based Model of Systemic Inflammation. Front. Physiol. 12:662845. doi: 10.3389/fphys.2021.662845

Introduction: Accounting for biological heterogeneity represents one of the greatest challenges in biomedical research. Dynamic computational and mathematical models can be used to enhance the study and understanding of biological systems, but traditional methods for calibration and validation commonly do not account for the heterogeneity of biological data, which may result in overfitting and brittleness of these models. Herein we propose a machine learning approach that utilizes genetic algorithms (GAs) to calibrate and refine an agent-based model (ABM) of acute systemic inflammation, with a focus on accounting for the heterogeneity seen in a clinical data set, thereby avoiding overfitting and increasing the robustness and potential generalizability of the underlying simulation model.

Methods: Agent-based modeling is a frequently used modeling method for multi-scale mechanistic modeling. However, the same properties that make ABMs well suited to representing biological systems also present significant challenges with respect to their construction and calibration, particularly with respect to the selection of potential mechanistic rules and the large number of associated free parameters. We have proposed that machine learning approaches (such as GAs) can be used to more effectively and efficiently deal with rule selection and parameter space characterization; the current work applies GAs to the challenge of calibrating a complex ABM to a specific data set, while preserving biological heterogeneity reflected in the range and variance of the data. This project uses a GA to augment the rule-set for a previously validated ABM of acute systemic inflammation, the Innate Immune Response ABM (IIRABM) to clinical time series data of systemic cytokine levels from a population of burn patients. The genome for the GA is a vector generated from the IIRABM's Model Rule Matrix (MRM), which is a matrix representation of not only the constants/parameters associated with the IIRABM's cytokine interaction rules, but also the existence of rules themselves. Capturing heterogeneity is accomplished by a fitness function that incorporates the sample value range ("error bars") of the clinical data.

Results: The GA-enabled parameter space exploration resulted in a set of putative MRM rules and associated parameterizations which closely match the cytokine time course data used to design the fitness function. The number of non-zero elements in the MRM increases significantly as the model parameterizations evolve toward a fitness function minimum, transitioning from a sparse to a dense matrix. This results in a model structure that more closely resembles (at a superficial level) the structure of data generated by a standard differential gene expression experimental study.

Conclusion: We present an HPC-enabled machine learning/evolutionary computing approach to calibrate a complex ABM to complex clinical data while preserving biological heterogeneity. The integration of machine learning, HPC, and multi-scale mechanistic modeling provides a pathway forward to more effectively representing the heterogeneity of clinical populations and their data.

Keywords: machine learning, agent based modeling, high performance computing, genetic algorithm, biological heterogeneity

INTRODUCTION

Heterogeneity of biological phenotype is an essential characteristic that provides robustness for organisms in variable and ever-changing environments and provides the range of fitness across individuals necessary for natural selection and evolution to function (Csete and Doyle, 2004; Stelling et al., 2006). Accounting for biological heterogeneity, be it in experimental systems or in clinical data, represents one of the most critical challenges to identifying shared and fundamental properties across biological entities (Gough et al., 2017). In addition to the concepts described in Gough et al. (2017), we have previously proposed that multi-scale computational models can serve as focused abstractions of biological systems to enhance the study and understanding of how these systems function; furthermore, enhancing their ability to capture and reflect complex biological heterogeneity can increase their utility as means of generating more robust, generalizable and translatable knowledge (An, 2018). All computational and mathematical models incorporate parameters that help define their behavior; variations of those parameters can be used to represent the heterogeneity seen in the dynamics of the biological systems represented by those models (Cockrell et al., 2020). We have extended this concept to the propose that a "parameter space" that results in recapitulation of bioplausible phenotypes can reflect genetic and epigenetic variation within a population, and assert that the model rule structure, which represents knowledge of the interactions between the components of the biological system, can be optimized to reflect a more accurate interaction network able to capture an increased variation of behavioral phenotypes. Herein we present a method utilizing genetic algorithms (GAs), a machine learning method for complex optimization, to calibrate and refine an agent-based model (ABM) of systemic inflammation to capture the heterogeneity and variability of a clinical data set. This method represents a departure from traditional approaches to calibration and parameterization that generally focus on using "cleaner" data sets with less variation/heterogeneity and/or fitting to a regression that draws a mean through what variation is present in the selected data, a process that can result in over-fit and brittle models. Alternatively, we propose that models (in terms of both parameters and interaction rules) selected for being able to reproduce an entire range of data within a dataset are more robust and generalizable, and therefore able to enhance the translation and applicability of knowledge.

This work focuses on enhancing the utility of ABMs as means of instantiating mechanistic hypotheses (An, 2009). Agent-based modeling is an object-oriented, discrete-event, rule-based, spatially explicit, stochastic modeling method (Bonabeau, 2002). In an ABM, individual agents representing components of the overall system are simulated interacting with each other and with their environment. These interactions are mediated by a pre-defined set of rules, typically derived from the literature and expert knowledge, and often containing stochastic components, to reflect either known probabilistic components in their behavioral rules or epistemic uncertainty regarding how those rules are resolved. As such, ABMs are computational instantiations of mechanistic knowledge regarding the systems being modeled and consequently are often used to simulate complex systems in which the aggregate of individual agent interactions can lead to non-trivial or unintuitive macro-state/system-level behaviors (An et al., 2009). This makes agent-based modeling a powerful technique for representing biological systems; rules are derived from experimentally observed biological behaviors, and the spatially explicit nature of the models give it an inherent ability to capture space/geometry/structure of biological tissue, which facilitates the ability of biomedical researchers to express and represent their hypotheses in an ABM (An, 2009). ABM's have been used to study and model a wide variety of biological systems (Bonabeau, 2002), from general purpose anatomic/cell-for-cell representations of organ systems capable of reproducing multiple independent phenomena (Cockrell et al., 2014, 2015) to platforms for drug development (An et al., 2011;

Cockrell and Axelrod, 2018), and are frequently used to model non-linear dynamical systems such as the human immune system (Baldazzi et al., 2006; Bailey et al., 2007; Cockrell and An, 2017; An, 2018).

In the process of developing an ABM, hypotheses or pieces of existing knowledge are re-framed as *rules* that determine the behavior of the agents when they interact with each and their environment. For example, in the context of a biomedical ABM one of those rules might be the definition of a cytokine signaling pathway, i.e., Tumor Necrosis Factor α (TNFα), a pro-inflammatory cytokine, upregulates Interleukin-10 (IL-10), an anti-inflammatory cytokine. The quantification of the effect that TNFα has on IL-10 in this hypothetical rule is determined by adjusting the parameters associated with that rule during model calibration, a critical step in the development and refinement of an ABM (Bonabeau, 2002; Rogers and Von Tessin, 2004; Bianchi et al., 2007; Windrum et al., 2007; Liu et al., 2017).

Parameter Space as a Means of Capturing Genetic/Epigenetic/Intrapopulation Variability

All computational models incorporate parameters within the rules/equations that make up the model. In dynamic mechanistic models, like ABMs, those rules often represent cellular functions and molecular events, such as receptor binding, signaling, gene activation, protein synthesis or secretion (**Figure 1**). However, the vast majority of mechanism-based computational models do not explicitly represent every component of every step present in the cell; in practice this is nearly functionally impossible at the current time because the sum total of interactions between components, or even the total set of components, is not known. Therefore, essentially all computational models that utilize rules to govern cellular behavior use some degree of abstraction and developer choice in what entities and functions are represented; these choices are often termed the *variables* of the model. These models invariably incorporate sets of parameters/coefficients that reflect the contribution/effect of a particular biological entity/mediator explicitly represented within a model's rules; these are the *parameters* that modify the variables within a stated rule. We assert that for rules of this type/form the parameters/coefficients represent a concatenation of various mediators, pathways and genes *not explicitly represented* that affect the interaction process represented in the rule (**Figure 1**), and therefore provide a means of capturing "hidden" control factors (known and unknown) that provide variation across a population of biological entities.

Note that these parameters are an aggregation of a whole series of factors: i.e., the effect of other health factors, such as co-morbidities or age, on the represented rules/functions, unknown mediators or genes, essentially any potential factor than can affect the functional output of the represented rule. Cast in this fashion, the multi-dimensional space of parameters can encompass a range of genetic/epigenetic/functional variability of the type present in a heterogeneous clinical population. We propose that characterizing this parameter space and its associated ensemble of model forms enhances the applicability and generalizability of a model's rule structure and can avoid "overfitting" and the generation of brittle models. Given the high-dimensional nature of this type of model parameter space we propose to use a machine learning/evolutionary computing optimization method, GAs, in order to generate an *ensemble of parameterizations* able to recapitulate a heterogeneous clinical data set. We would like to emphasize that while GA is an optimization method that will converge to an "optimal" solution, we do not suppose that the optimized solution is necessarily more plausible that the rest of the sufficient parameterizations within the ensemble. Rather, we are utilizing the convergence process of the GA to identify a set of parameterizations sufficient to represent the range of heterogenous clinical data; this ensemble of parameterizations then forms the bioplausible manifestations of simulation model, which can then be used for further studies on disease forecasting (Larie et al., 2020) or therapeutic control discovery (Cockrell and An, 2018; Petersen et al., 2019). Our proposed method is related to how parameter spaces are used to define the behavior of ordinary differential equation (ODE) models, where different fits are used to match different values within a range in a time series of data. However, we believe that the use of ABMs provides an extension of the representational capability of ODE parameter space characterization by the stochastic properties of the ABMs, which reflect intrinsic biological stochasticity, to generate population distributions for individual parameterizations (as opposed to unique deterministic trajectories seen in an ODE).

We also note our attempt to avoid the use of the term "fitting" for this process, a term that brings to mind the way that statistical models are adjusted to match data (though often applied to the calibration of ODE models). Rather than trying to precisely and restrictively identify "fitted" parameterizations, which commonly requires a lossy process by which the heterogeneity of the data is compressed into a mean, we aim to find *sufficient* parameterizations that are able to recapitulate the range of data present. Given how we have defined the role of the parameters in the model (**Figure 1**) there is no supposition that a "single" parameterization exists within the clinical population, but rather that a population is represented by an ensemble of parameterizations. However, given the epistemic uncertainty associated with all the potential factors that might affect the behavior of the model, it is currently impossible to specify what the distribution across a real population of those parameterizations; the only means we have of determining their plausibility is via the existing data. This strategy is specifically designed to avoid "overfitting," which we interpret as a failure of generalizability of a particular model when it is exposed to new, additional data; our intent is to preserve and refine the expressiveness of a model's rule structure with a focus on recapitulating the heterogeneity seen in biological data.

In the sections below we present a method and results that uses the convergence process of GAs to identify an ensemble of parameterizations for an ABM of acute systemic inflammation sufficient to recapitulate the heterogeneity of a clinical data set from burn patients.

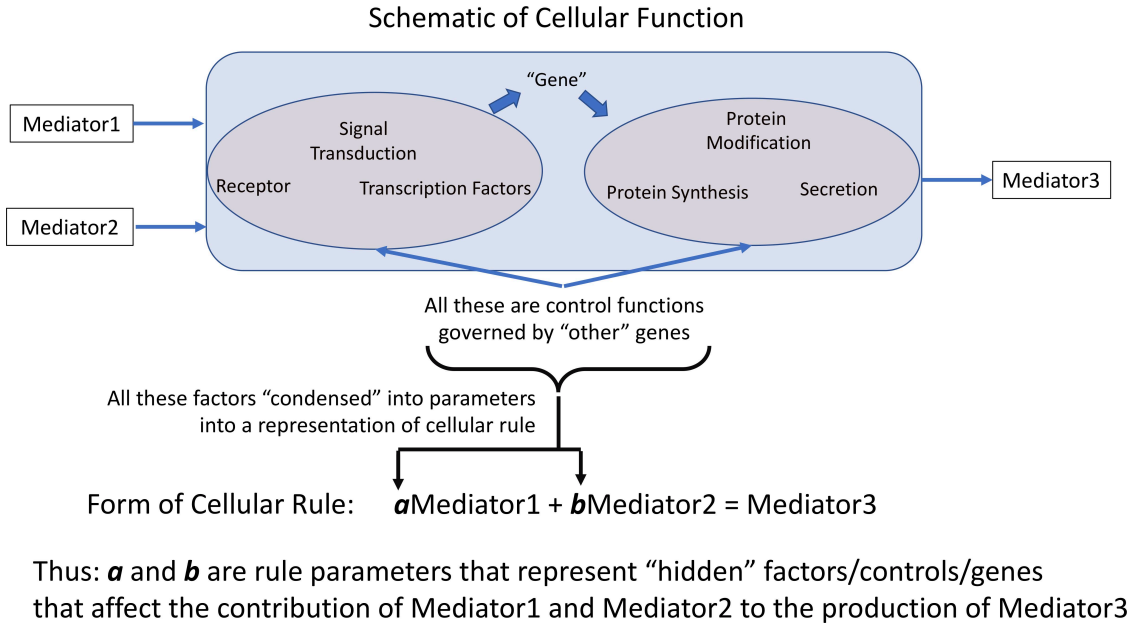


FIGURE 1 | Depiction of how representation of cellular behavioral rules governing the effect and generation of various mediators is accomplished by rule parameters. Cellular rules are presented as input-output relationships for specific cell types; in practice not every mechanistic step is represented in such a rule. The weight of each contributing mediator to the overall function of the cell is represented by the parameters associated within each rule. We pose that these parameters essentially aggregates the influence of non-represented or unknown cellular components on the represented rule. a and b are rule parameters that represent "hidden" factors/controls/genes that affect the contribution of Mediator 1 and Mediator 2 to the production of Mediator 3.

MATERIALS AND METHODS

The Model Rule Matrix

In our ABMs the rules and a set of coefficients that quantify the effect of the rules (see **Figure 1**) are stored in an object which we refer to as the Model Rule Matrix (MRM). In this scheme, specific rules are represented by rows in the matrix; each computationally relevant entity in the model is then represented by the matrix columns. As a simple example, the system of model rule equations for a single cell:

$$M1_{t+1} = M1_t + M2_t$$

$$M2_{t+1} = -M1_t + M3_t$$

Would be represented by the matrix:

$$\begin{bmatrix} 1 & 1 & 0 \\ -1 & 0 & 1 \end{bmatrix}$$

Where the first column holds the rule coefficients for Mediator 1 (M1), the second column holds the rule coefficients for Mediator 2 (M2), and the third column holds the rule coefficients for Mediator 3 (M3). We note that this is a simplified rule for the purpose of illustration. The matrix is readily decomposable into a one-dimensional vector, upon which we can operate using GAs. The number of rows in the matrix then is equal to the number of rules that it represents, and the number of columns is equal to the number of entities that could potentially contribute to the decision made by their associated rule. Note that if a particular interaction between model components is not represented then the corresponding position within the MRM contains a "0." Therefore, the MRM presents a compact mathematical representation of the interaction rules present in an ABM.

The resulting product of this work is an ensemble of biologically/clinically plausible model parameterizations, representing a genetically/epigenetically/functionally diverse cohort of *in silico* patients, able to represent a range of heterogeneous experimental or clinical data. In this sense, elements of this work are similar to traditional sensitivity analysis techniques (Cukier et al., 1978; Saltelli et al., 2004, 2008); the primary distinction lies in the fact that these algorithms consider alternate rule configurations (as represented by the conversion of zero to non-zero elements in the MRM), which can change model-parameter sensitivities (Cockrell et al., 2020).

The Reference Model: IIRABM

In this work, we utilize a previously developed an ABM of systemic inflammation, the Innate Immune Response ABM (IIRABM). Though the IIRABM has been calibrated to simulate blunt trauma and infectious insult, it is an abstract and generalizable (An, 2004; Cockrell and An, 2017) model of human response to injury. Cytokine time series and systemic response varies significantly between both blunt trauma/infectious insult and severe/large surface area burns. In this work, we demonstrate the changes necessary to recalibrate the model from simulating an infectious injury to a caustic and sterile injury. The IIRABM is a two-dimensional abstract representation of the human endothelial-blood interface. This abstraction is designed to model

the endothelial-blood interface for a traumatic (in the medical sense) injury and does so by representing this interface as the unwrapped internal vascular surface of a 2D projection of the terminus for a branch of the arterial vascular network. The closed circulatory surface can be represented as a torus, and this two-dimensional surface defines the interaction space simulated by the model. The spatial geometry of the circulatory system and associated organ interfaces are not directly mapped using this scheme. This abstraction reproduces the circulatory topology accessible by the innate immune system and presents a unified means of representing interaction between leukocytes and endothelial surfaces across multiple tissue and organ types. The IIRABM utilizes this abstraction to simulate the human inflammatory signaling network response to injury; the model has been calibrated such that it reproduces the general clinical trajectories of sepsis. The IIRABM operates by simulating multiple cell types and their interactions, including endothelial cells, macrophages, neutrophils, T-lymphocyte subtypes (TH0, TH1, and TH2 cells) as well as their associated precursor cells. Intrinsic biological stochasticity, such as the spatial distribution of cells at initialization or movement direction not governed by chemotaxis and the manifestation of switches governing cellular actions, is represented by the introduction of randomness into the IIRABM; this allows the IIRABM to generate a population distribution of different trajectories from an identical parameterization/initial conditions. The simulated system dies when total damage (defined as aggregate endothelial cell damage) exceeds 80%; this threshold represents the ability of current medical technologies to keep patients alive (i.e., through mechanical organ support) in conditions that previously would have been lethal. The IIRABM is initiated using five parameters representing the size and nature of the injury/infection as well as a metric of the host's resilience: (1) initial injury size, (2) microbial invasiveness (rate at which infection spreads), (3) microbial toxigenesis (rate at which infection damages tissue), (4) environmental toxicity (amount of spontaneous infectious exposure in the environment, such as an Intensive Care Unit), and (5) host resilience (the rate at which damaged but not dead tissue recovers). These five parameters clearly have correlates in the real world, and yet are nearly inherently un-quantifiable. Therefore, they are treated as dimension-less coordinate axes in which the behavior of the IIRABM exists.

The IIRABM characterizes the human innate immune response through the simulated generation of a suite of biomarkers, including the pro-inflammatory and anti-inflammatory cytokines represented in the IIRABM. At each time step, the IIRABM outputs the total amount of cytokine present for all mediators in the model across the entire simulation. The ordered set of these cytokine values creates a high-dimensional trajectory through cytokine space that lasts throughout the duration of the simulation (until the *in silico* patient heals completely or dies). We note that stochastic effects can play a significant role in simulation dynamics. Model parameterizations used in this work lead to a simulated mortality rate of 50%; in these simulations, identical injuries and initial conditions are given to the model and over time, the trajectories diverge to the point that half of the simulated cohort heals completely and half dies. The fact that the initial conditions are exactly identical means that it is indeed stochasticity, not chaos, that leads to the diverging trajectories. A detailed discussion of this can be found in Cockrell and An (2017).

While the IIRABM successfully simulates the human immune response to injury at a high, overall system level (outcome proportions, time to outcome, etc.), it may not always replicate specific cytokine time series. A cytokine time series is not a single sequence of numerical values; rather, it is a sequence of ranges, indicating significant heterogeneity clinical response to severe burns, within which the cytokine measurements fall for a given patient in the cohort that generated the time series. This heterogeneity is challenging because the magnitude of these ranges is not temporally constant. In order for a computational model to be biologically realistic, it must be able to generate any physiological state which can experienced by the biology that is being simulated and do so with the appropriate frequency. We have previously characterized the shapes of the probabilistic "clouds" of multi-dimensional state space of the IIRABM (Cockrell and An, 2017); these distributions, which are more akin to the range of variable behavior generated by biological systems, are too complex to be represented by a small/simple set of stochastic differential equations with an analytically defined "noise" term. This prompts the need to execute the ABM at large scale in order to more effectively capture the population dynamics structure present in a clinical data set.

Application of Genetic Algorithms

In this work, we use GA to operate on the IIRABM's rule set such that it can accurately simulate the cytokine time course and final outcomes for a serious burn injury. As noted in the Introduction, we are employing GA is a non-standard fashion, where rather than seeking a specific optimal parameterization of the MRM we are using the process of convergence of the GA to identify an ensemble of valid parameterizations. Cytokine time series were extracted via inspection from Bergquist et al. (2019). In Bergquist et al. (2019) provide a variety of blood cytokine levels over 15 time points and 22 days for patients which exhibited severe burns over 50% of the surface area of their bodies. The authors observed a mortality rate of 50% for this category of injury.

A GA (Goldberg and Holland, 1988; Fonseca and Fleming, 1993; Haupt and Ellen Haupt, 2004) is a population-based optimization algorithm that is inspired by biological evolution. In a GA, a candidate solution is represented by a synthetic "genome," which, for an individual, is typically a one-dimensional vector containing numerical values. Each individual in a GA can undergo computational analogs to the biological processes of reproduction, mutation, and natural selection. In order to reproduce, two individual vectors are combined in a crossover operation, which combines the genetic information from two parents into their progeny.

Using this scheme, cytokines produced by a given cell type are held fixed, while the stimuli that lead to the production of that specific cytokine are allowed to vary. This maintains a distinction between the cell and tissue types represented in the model throughout the MRM evolution from the GA.

The candidate genomes which comprise the rule set are then tested against a fitness function which is simply the sum of cytokine range differences between the experimental data and the computational model:

$$F = \sum_{i,t} \left| \max\left(C^e_{i,t}\right) - \max\left(C^m_{i,t}\right) \right| + k\left|R_e - R_m\right|,$$

where $C^e_{i,t}$ represents the normalized blood serum level of cytokine i at time point t from the experimental data, $C^m_{i,t}$ represents the normalized blood serum level of cytokine i at time point t from the IIRABM, R_e represents the experimentally observed mortality rate, R_m represents the model-generated mortality rate, and k is an adjustable parameter to govern the importance of the mortality rate contribution to the fitness function. For the purposes of this work, we consider an optimal solution to be one that minimizes the above fitness function. In order to avoid issues of over-fitting, we held the time points at t = 48 h post-burn and t = 8 days post-burn back from the evaluation of candidate fitness. Despite this, these time points were well-matched between the *in silico* and *in vivo* experiments.

We note that 50 stochastic replicate simulations of the IIRABM were used to generate simulated ranges, while only 20 patients comprised the clinical data set. The reasoning for this is that the simulated range was not stable using only 20 stochastic replicates; we found that when we ran 50 replicates per parameterization, the simulated cytokine ranges varied only by a few percent. Additionally, we did not have access to individual data points, or distributions at different time points; we only had the maximum and minimum values, and thus were unable to evaluate the effect that additional clinical patients would have had on the observed clinical data range.

Candidate genomes are then selected against each other in a tournament fashion, with a tournament size of 2 [28, 29]. The tournament winners make up the breeding pool, and progenitor genomes are randomely selected and paired. We implement a variant of elitism in that, at the completion of the tournament, the least fit 10% of the candidate progenitors are replaced with the fittest 10% of candidate genomes from the precious generation. Progeny genomes are defined with a uniform crossover operation using a standard continuous formulation (Haupt and Haupt, 2004):

$$C_{1,i} = \beta P_{1,i} + (1-\beta)P_{2,i}$$

$$C_{2,i} = \beta P_{2,i} + (1-\beta)P_{1,i}$$

Where $C_{1,i}$ is the value for gene i in child 1, P is the value for gene i in parent 1, and β is a random floating-point number between 0 and 1. After breeding, each child is subject to a random chance of mutation which begins at 1% and increases with each generation.

We employ an elitist strategy by replacing the least fit 10% of the breeding population with the most fit parameterizations. This ensures that our best solutions are not lost due to mutation. Additionally, we utilize two non-standard additions to the GA: the *non-viability criterion* and the *ensemble retainment criterion*. As noted above, the potential parameter space is astronomically large, and the vast majority of those putative parameterizations are in no way biologically viable or plausible; it is therefore desirable to filter these regions of parameter space early in this process. The non-viability criterion immediately rejects any parameterization which leads the model to die before the first clinical time point (3 h post-injury); these are replaced with fitter candidates. In our experience with this model, this non-viability criterion is only activated in the first few generations, as the algorithm quickly finds a focus on viable regions of parameter space. Further, we recognize that any putative parameterization which generates cytokine trajectories that always lie within the cilnically observad range cannot be invalidated by the data, and are therefore biologically plausible; thus, these parameterizations should be retained for inclusion into the final ensemble. As the goal of the fitness function is to obtain maximum coverage over the clinical data range, some of these viable parameterizations may be lost as the population evolves.

The IIRABM was optimized for 250 generations with a starting population size of 1024 candidate parameterizations. The IIRABM was implemented in C++ and the GA was implemented in Python 3; and simulations were performed on the Cori Cray XC40 Supercomputer at the National Energy Research Scientific Computing Center and at the Vermont Advanced Computing Center. Codes can be found at https://github.com/An-Cockrell/IIRABM_MRM_GA. Pseudocode for this procedure is given below:

(1) Initialize starting population, P, where each $P_i \in P$, is represented by a matrix with elements randomly assigned in the range $[-2,2]$
(2) REPEAT-UNTIL stopping condition is met (maximum generations or minimum fitness)
(a) BROADCAST candidate parameterizations to available processes
(b) On each process, CALL IIRABM simulation
(c) Determine Fitness, F_i
(i) NON-VIABILITY CRITERION: IF $F_i > F_c$ THEN
(1) Discard P_i
(2) Replace with $P_{j\neq i}$, where $F_j < F_c$
(d) ENSEMBLE RETAINMENT: Determine Bioplausibility
(i) IF all simulated cytokine values are contained within the range of clinical data, then retain parameterization for inclusion into the ensemble, E
(e) GATHER fitnesses to root process
(f) Tournament Selection
(i) Randomly select pairs of parameterizations
(ii) Select fitter parameterization for inclusion into breeding pool B
(g) Breeding
(i) Randomly select pairs of parameterizations from B
(ii) Generate two progeny parameterizations, where matrix elements are combined using the standard continuous formulation.
(h) Mutation
(i) Set mutation probability, $r_m = 0.01 + 0.002 * g_n$, where g_n is the number of generations completed by the GA
(ii) Generate random number r

(iii) IF $r \leq r_m$ THEN randomly select matrix element to mutate, and assign a random value in the range [−2,2]
(i) Check if any fitness has reached the minimum value (0, indicating a single parameterization matches the data perfectly) or the maximum number of generations has been reached.

We note that we ran the algorithm 10 times, all with random seeding parameterizations, and found that, though the initial populations were completely random, the GA converged to the same region of parameter space each time we ran it. This does not preclude the existence of alternate regions, but indicates that, if they exist, their hypervolumes are significantly smaller than the region of parameter space represented by our ensemble population, which is contiguous at the level of resolution that we have used to examine it. Additionally, the simulation never reached a fitness of 0, indicating that a single parameterization of our model cannot explain all the data.

RESULTS

For the initial attempt with the GA the contributions of each of the five cytokines were weighted equally. This generated an ensemble of sufficient forms of the MRM that produced excellent results for four out of five of the comparison cytokines. However, the GA could not converge well enough to produce MRMs able to generate IL-10 concentrations which matched the literature, with peaking occurring at 6 h post-insult rather than 5 days post-insult, as was seen clinically **Figure 2A**). As a potential explanation for this inability to replicate IL-10 data we note that in comparison to the other cytokine time series IL-10 showed spikes at t = 5 days but is near zero everywhere else, suggesting that a poor fit is more likely when using a fitness function that weights the contributions of each cytokine equally. A candidate MRM parameterization that minimizes IL-10 production over the entire time course would thus contribute less to the overall fitness (in this case, we seek to minimize the fitness function) than a hypothetical parameterization that was 10% off on TNF levels for every time step. In order to address this, we both doubled and tripled the weight of the coefficient to the portion of the fitness function that incorporated IL-10 contribution. Both of these modifications showed similar improvements over the initial fitness function, but neither was significantly better than the other. This leads us to expect that a doubling of the IL-10 contribution to the fitness is sufficient. We display this difference in **Figure 2**.

A plot of cytokine ranges for 5 cytokines which existed in the clinical data set and were already present in the model at the start of this work (GCSF, TNF-α, IL-4, IL-10, and IFN-γ) is shown in **Figure 3**. Ranges for the original model, described in Cockrell and An (2017); An (2018), are shown in black; ranges for the published data (Bergquist et al., 2019) are shown in red; and results from the optimized ensemble model are shown in green.

The temporal cytokine dynamics expressed by the optimized IIRABM are significantly modified from its original incarnation. We note that the ensemble models are optimized to match four out of five of the cytokines used in the fitness function to be nearly indistinguishable from the clinical data. We note a slight under-expression of IL-10 at t = 5 days post-injury. This discrepancy identifies a weakness in our model when it is being used to simulate burns, namely, that the cellular production of IL-10 is not well enough defined, in that its production is limited to activated macrophages and TH2 helper cells. Given that the IIRABM was developed to represent the innate immune response to traumatic injury, we consider this recalibration to burn injuries to be a success.

In **Figure 4**, we depict the MRM as a heat map of the values (**Figures 4A,B**) before and at the end of the GA runs. Numerical values for these matrices can be found in the **Supplementary Material**. **Figure 4A** shows the MRM values of the original implementation of the IIRABM prior to training; the sparseness of the matrix reflects the necessary abstracting modeling choices made in terms of which rules to represent. **Figure 4B** shows the "optimized" MRM at the end of the GA runs, noting that while this MRM is the one that most closely matches the *range* of data seen clinically it is representative of the ensemble of MRM able to generate data matching the ranges seen in the clinical data. The optimized matrix has a much more connected structure, and is a dense matrix, as opposed to the sparse original rule matrix. There are not any matrix elements with a value of 0 in the optimized matrix, though there are many elements with comparatively small values. This is an intuitive result and is the intended output based on how the MRM is defined in terms of **Figure 1**; as all mechanism-based computational models represent a limited and reduced representation of biological reality it is not surprising that there are additional connections needed in order for the model to recapitulate real-world data. As such, this structure of the optimized MRM is similar to what is seen in experimental bioinformatic studies; all of the cytokines in this network appear to be connected to each other, at least to a small degree, while a smaller number of strong connections (which could also be considered correlations) provide the majority of the influence on the system dynamics. The original rule matrix, formatted and with complete labeling, can be found in the **Supplementary Material**.

We note that while the process of the GA will lead to convergence to an "optimal" MRM that most closely matches the *range* of data observed clinically, any parameterization which generates a range of data that is encompassed by the clinical data is retained in the ensemble of valid parameterizations. It is this ensemble that is the intended output of the GA process. In **Figure 5** we depict the ranges of values of the MRM in the valid ensemble, both as a 2-dimensional heatmap and the same data shown as a 3-dimensional bar graph to aid in visualization of the range of MRM values within the ensemble.

In **Figure 6**, we present the time evolution of the diversity of the simulated population. We define the total diversity of a population to be the sum of the ranges of each matrix element. In **Figure 6A**, matrix element ranges are ordered from low to high. In the first several generations, diversity is maximized over the entire matrix. As the system evolves toward an optimum parameterization, diversity decreases, and the matrix begins to converge to a single value. In order to

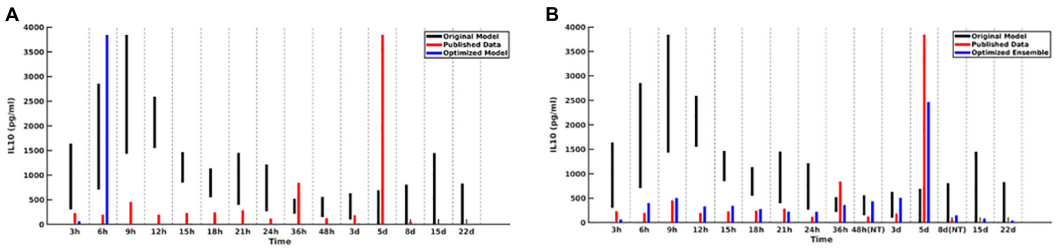


FIGURE 2 | Cytokine ranges are shown for IL-10 for the original model (black), published data (red), and optimized ensemble model (blue). On the left **(A)**, the IL-10 contribution to the fitness function is weighted equally to the other cytokines, with the result that simulated IL-10 levels after 6 h are essentially 0; on the right **(B)**, the IL-10 contribution to the fitness function was doubled.

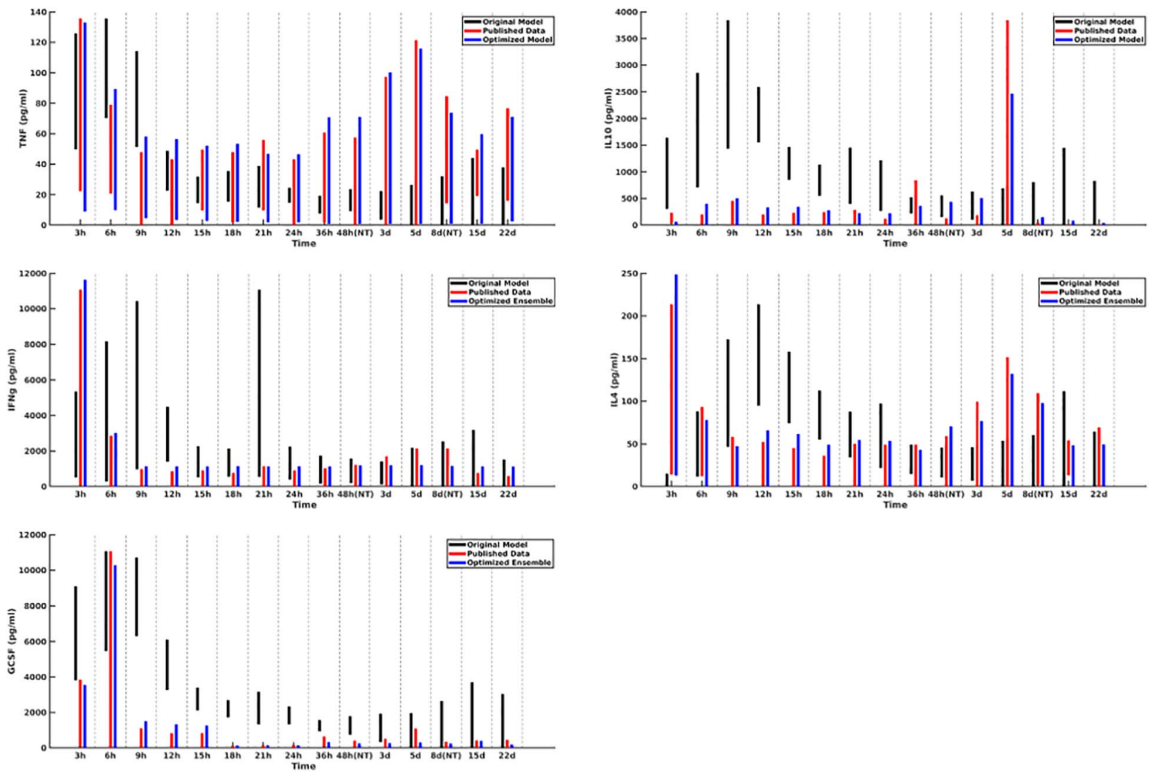


FIGURE 3 | Cytokine ranges are shown for the original model (black), published data (red), and optimized ensemble model (blue) for TNFα (top-left), IL-10 (top-right), IFNγ (center-left), IL-4 (center-right) and GCSF (bottom-left). Ranges for the computational models were generated using 50 stochastic replicates.

combat this, we use a mutation rate that increases as a function of the generation number, which begins to reintroduce diversity into the population. This is seen in **Figure 6A**, as the matrix element ranges begin to return to a diverse configuration, and more globally in **Figure 6B**, which plots the total diversity metric as a function of generation number.

DISCUSSION

The IIRABM rule set utilized in this work contained 432 free and continuous parameters, many of which had highly non-linear or conditional effects on the model-generated cytokine trajectories and outcomes. This high-dimensional parameter space provides an astronomically large set of possible behaviors, of which only a subset are bioplausible. Concurrently, biological objects manifest population-level individual heterogeneity, which means that "bioplausibility" is not a particular trajectory (or mean of trajectories) but rather a set of behaviors and outputs producible by the biological system. Our only guide to this set of behaviors is the range of outputs captured within a data set. The task, then, is to establish a concordance between the range of behaviors represented by a subset of the parameter space of the computational model and the range of outputs seen in the data set and to bound the putative bioplausible parameter space using the data available. The subject of this paper is to present an

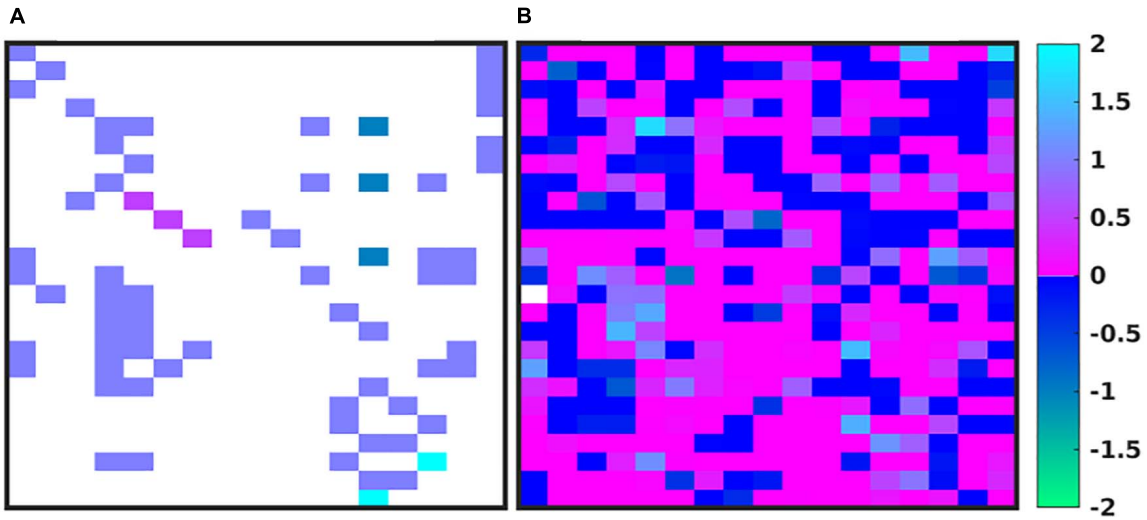


FIGURE 4 | Depictions of the MRM A heatmap of the original rule matrix is shown in panel **(A)**, the optimized matrix representative of the valid ensemble is shown in panel **(B)**. In panels **(A,B)**, the white blocks represent a matrix element with a value of 0 (e.g., no connection); the dark blue to green represents a negative matrix element; the pink to light blue represents a positive matrix element. The optimization process vastly increases the connectivity of the ABM elements (as would be expected in the true biological system).

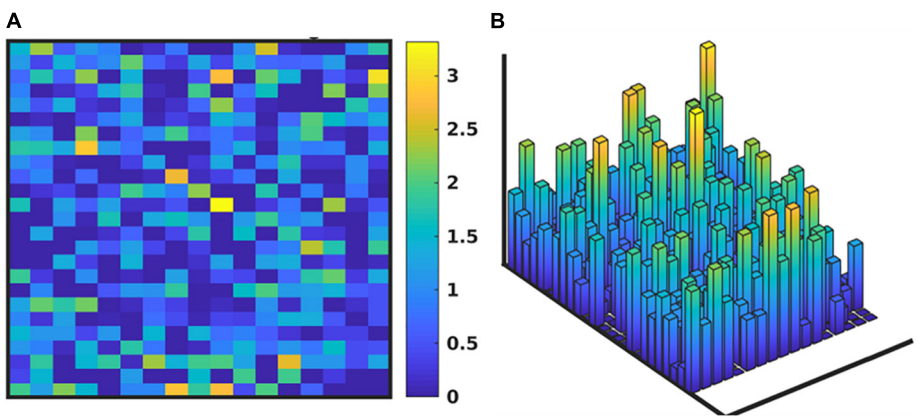


FIGURE 5 | Depiction of the range of values of the MRM for the valid ensemble able to produce data consistent with the clinical data. Panel **(A)** shows the ranges of the MRM values as a heatmap, where dark blue is a range of 0 and yellow indicates a range of 3.42, with a maximal range of 4.0. Panel **(B)** shows this same data as a 3-dimensional bar graph, where the height of each cell reflects the range of the values for each matrix element.

alternative means of calibrating a computational model to a data set with an emphasis on maintaining the capability to represent the heterogeneity of the data, thereby potentially reflecting critical biological processes that account for the ubiquitous inter-individual variability seen in biological systems.

There are the critical and intertwined issues regarding definition of the fitness function, overfitting, and choice of algorithm. Our utilization of GA was non-standard: while the algorithm sought to optimize the results of the simulation to minimize a fitness function, the discovery of the optimum parameterization was not the actual goal of the work. As our GA traversed the parameter space toward its optimum destination, it gathered all model parameterizations that *were not invalidated* by the available data into the final ensemble. The fitness function was designed such that an optimal solution would minimize the difference between the range of data generated by the model and the range of data observed clinically, but with the explicit aim of defining this bioplausible set rather than finding "a" particular optimal solution.

The design of the fitness function is intimately connected to the concept of overfitting, and some might interpret transition from a sparse rule matrix to a dense rule matrix as the parameter set is optimized as an indication of potential overfitting. This concern stems from the concept of overfitting of statistical models, where the addition of new terms can lead to spurious relationships that may not be present in new data and therefore lead to decreased performance (e.g., failure of generalizability). To some degree this is not the case for mechanism-based dynamic

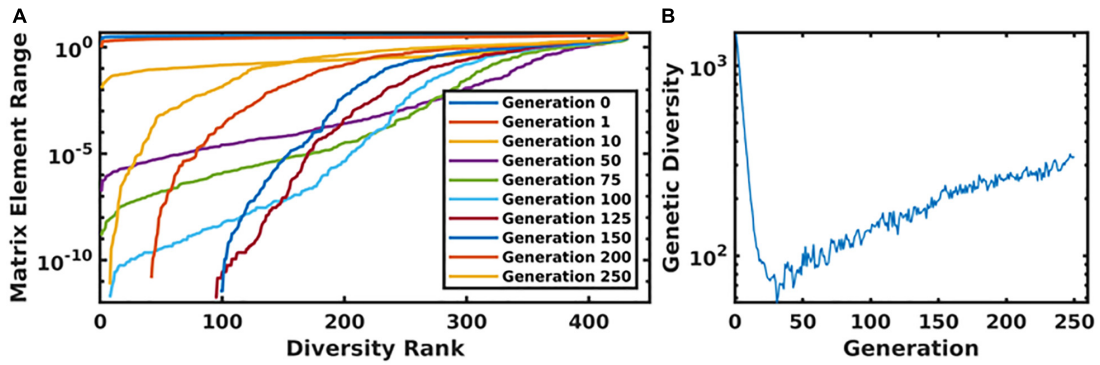


FIGURE 6 | Panel **(A)** displays the ordered matrix element ranges for a variety of time points throughout the genetic algorithm. In this plot, the most diverse generations are represented by a nearly horizontal line at the top of the plot. As the system evolved, this diversity begins to collapse until the increasing mutation rate compensates for the algorithm's convergence. This is displayed in panel **(B)**, which shows the total diversity of the population as a function of generation number.

models, where the putative additions to the model represent additional knowledge that (1) has a scientific justification for its addition, (2) theoretically increases the expressiveness (e.g., increased generalizability) of the data and (3) are actually present in the real-world biological object. In addition, from a methodological standpoint, we contend that the traditional concern of overfitting (e.g., failure to generalize) should not be an issue for this approach, according to the following logic:

(1) The primary danger of overfitting is the introduction of spurious elements to the model which would lead to the model's failure to generalize to new data outside of the data used to train it, ultimately resulting in an invalidation of the model. The primary goal of this work is to generate a diverse population of model parameterizations which are encompassed by the clinical data; when taken in aggregate, and due to the fact that each parameterization generates a range of behavior, this population of parameterizations fills out the range of data observed clinically. While one could claim that a particular added component may not be necessary in order to replicate the data (violation of the concept of parsimony), the addition of such a term cannot be invalidated in comparison to the data.

(2) The introduction of new data *cannot invalidate* individual parameterizations in our ensemble because the introduction of new data can take only two forms: (1) it is either encompassed within the range of the existing data, in which case the previously valid parameterizations are still valid, or (2) new data can be outside the existing range, which does not invalidate any of the previously validated parameterization, but rather suggests an insufficiency in the expressiveness of the previously defined parameter space. In this case an additional search of the parameter space is needed because the current ensemble is insufficiently expressive to explain the heterogeneity of the clinical data and therefore parameterizations that were formerly considered invalid would now be seen to be biologically plausible.

We note that by setting the fitness function to match the published data as exactly as possible we are limiting the targeted degree of heterogeneity to that presented by the relatively small cohort of clinical patients.; the true range of biologically plausible blood cytokine concentrations in undoubtedly larger than what is seen in a small cohort of 20 individuals. In order to obtain a more generalizable model, we propose two alternative approaches to the above presented work: (1) that the fitness function should be configured to over-encompass the available data, with cytokine range boundaries determined by the probability density function (pdf) which governs the experimental data; or (2) synthesize multiple datasets in order to design a fitness with maximum cytokine rage coverage that is still supported by experimental data. Incorporating the shape of the probability density function into the fitness function can be difficult purely as a matter of practicality–often the raw data for human cytokine levels isn't available, and only the absolute range can be extracted from published manuscripts, and it is also common to see a cohort size that is too small to definitively propose a single pdf which adequately describes the data.

Our approach also involves addressing the limited representation inherent in all computational models. As essentially all mathematical/computational models of biological processes represent some degree of abstraction and are therefore necessarily incomplete, we recognize that the task of model "validation" is more often one of determining the conditions in which a model is "valid" and at what point the model is insufficient. While the employment of the MRM refinement is a means of "encompassing" the uncertainties and "missing" components of the ABM rules, there are still cases where the constraints placed by the choice of rules in the model preclude fitting to particular data points; it is at this point that the model is recognized to be falsified (in the Popperian sense). However, being able to specify where the model fails is extremely useful. In this case, the difficulties in being able to reproduce the trajectories of IL-10 help point to where the IIRABM is insufficient as a representation of the systemic response to burn injury, specifically with respect to the level of representation of anti-inflammatory components. This insight points to the need to incorporate other known anti-inflammatory components into future iterations of the IIRABM.

In future work, we will utilize this method to generate diverse *in silico* cohorts as part of our machine-learning therapeutic discovery workflow (Cockrell and An, 2018;

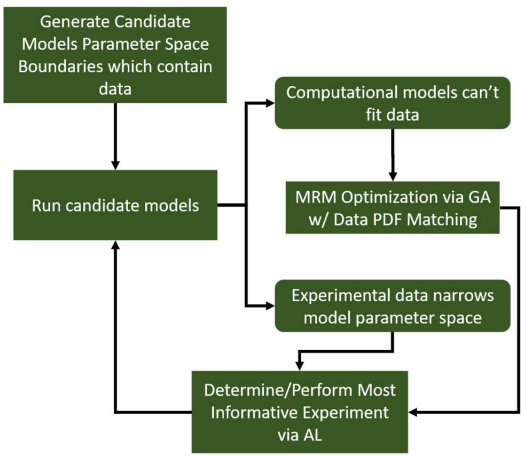


FIGURE 7 | A diagram indicating a hybrid experimental/computational workflow for the automated calibration and validation of ABMs using the MRM scheme. In this workflow, a computational model containing all mechanistic knowledge hypothesized to be relevant to the biological system in question is developed. The range of output for a comprehensive set of viable model parameterizations is determined and compared to biological data. At this point, experimental data can be used to eliminate some of the formerly viable model parameterizations or invalidate the model. In the event the model is invalidated, it can be redesigned/reconfigured to address its shortcomings. After that, the remaining set of putative model parameterizations is investigated to determine which specific parameters contribute the most variability to the model output. These are then the parameters that are selected for further characterization in subsequent biological experiments.

Petersen et al., 2019). We note the importance of *in silico* genetic diversity for therapeutic discovery in Cockrell and An (2018); in this work, we developed a multi-cytokine/multi-time-point therapeutic regimen which decreased the mortality rate from ∼80 to ∼20% for a severe simulated injury. The therapy was discovered using GAs on a single model internal parameterization. When we examined the non-responders, we noted that hyperactivity in specific pathways could manifest negatively, specifically, excess Granulocyte Colony Stimulating Factor activity lead to excess neutrophil recruitment, which instigated a state of perpetual inflammation. Brittle policies/solutions (i.e., those that are not applicable outside of the very specific circumstances used to train them) have long been recognized as a weakness of machine learning research (Holland J.H.(ed.), 1983). In order to overcome this obstacle, data used to train machine-learning algorithms should be sourced as broadly as possible. A useful analogy would be to compare the machine learning experiment to an *in vivo* biological experiment: performing a biological experiment on a set of genetically identical animals will yield less generalizable information than an experiment performed on a set of genetically heterogenous animals.

Further, we note that, while we generated a diverse *in silico* patient cohort which generates cytokine trajectories that match clinical data, the diversity is limited by the algorithm. We recognize that by using GA to find a path through parameter space toward some optimum of the fitness function, even though we collect viable parameterizations as the algorithm progresses, they are sampled from a limited region of parameter space. Many of the genes in each individual parameterization end up tightly constrained by the algorithm, while others have a larger range. These latter parameters are those about which the model is most uncertain. Future work will seek to more comprehensively explore the entire parameter space using active learning, similar to Cockrell et al. (2020). Active Learning is a sampling technique used in machine learning in which sampled data is chosen based on how much information it can apply to the machine learning model. A similar approach can be taken in this case. In order to most efficiently update and refine the computational model, experiments should be designed to query the model features that are most uncertain. This approach is illustrated in **Figure 7**. In this way, GA can play an integral role in the iterative cycle of model refinement and experimentation necessary to construct a high-fidelity generalizable computational model.

DATA AVAILABILITY STATEMENT

The original contributions presented in the study are included in the article/**Supplementary Material**, further inquiries can be directed to the corresponding author/s.

AUTHOR CONTRIBUTIONS

CC designed the machine-learning workflow, ran simulations, performed data analysis, and contributed to the manuscript. GA designed the initial IIRABM simulation and assisted in the design of the machine-learning workflow and contributed to the manuscript. Both authors contributed to the article and approved the submitted version.

FUNDING

This work was supported by National Institutes of Health grant U01EB025825. Additionally, this research used high performance computing resources of the National Energy Research Scientific Computing Center, a DOE Office of Science User Facility supported by the Office of Science of the United States Department of Energy under Contract No. DE-AC02-05CH11231, as well as resources provided by the Vermont Advanced Computing Core (VACC).

ACKNOWLEDGMENTS

This manuscript has been released as a Pre-Print at https://www.biorxiv.org/content/10.1101/790394v2.

SUPPLEMENTARY MATERIAL

The Supplementary Material for this article can be found online at: https://www.frontiersin.org/articles/10.3389/fphys.2021.662845/full#supplementary-material

REFERENCES

An, G. (2004). In silico experiments of existing and hypothetical cytokine-directed clinical trials using agent-based modeling. *Crit. Care Med.* 32, 2050–2060.

An, G. (2009). Dynamic knowledge representation using agent-based modeling: ontology instantiation and verification of conceptual models. *Methods Mol. Biol.* 500, 445–468. doi: 10.1007/978-1-59745-525-1_15

An, G. (2018). The crisis of reproducibility, the denominator problem and the scientific role of multi-scale modeling. *Bull. Mathematical Biol.* 80, 3071–3080. doi: 10.1007/s11538-018-0497-0

An, G., Bartels, J., and Vodovotz, Y. (2011). In silico augmentation of the drug development pipeline: examples from the study of acute inflammation. *Drug. Dev. Res.* 72, 187–200. doi: 10.1002/ddr.20415

An, G., Mi, Q., Dutta-Moscato, J., and Vodovotz, Y. (2009). Agent-based models in translational systems biology. *Wiley Int. Rev. Syst. Biol. Med.* 1, 159–171. doi: 10.1002/wsbm.45

Bailey, A. M., Thorne, B. C., and Peirce, S. M. (2007). Multi-cell agent-based simulation of the microvasculature to study the dynamics of circulating inflammatory cell trafficking. *Ann. Biomed. Eng.* 35, 916–936. doi: 10.1007/s10439-007-9266-1

Baldazzi, V., Castiglione, F., and Bernaschi, M. (2006). An enhanced agent based model of the immune system response. *Cell Immunol.* 244, 77–79. doi: 10.1016/j.cellimm.2006.12.006

Bergquist, M., Hastbacka, J., Glaumann, C., Freden, F., Huss, F., and Lipcsey, M. (2019). The time-course of the inflammatory response to major burn injury and its relation to organ failure and outcome. *Burns* 45, 354–363. doi: 10.1016/j.burns.2018.09.001

Bianchi, C., Cirillo, P., Gallegati, M., and Vagliasindi, P. A. (2007). Validating and calibrating agent-based models: a case study. *Comput. Econo.* 30, 245–264. doi: 10.1007/s10614-007-9097-z

Bonabeau, E. (2002). Agent-based modeling: methods and techniques for simulating human systems. *Proc. Natl. Acad. Sci. U.S.A.* 99(Suppl. 3), 7280–7287. doi: 10.1073/pnas.082080899

Cockrell, C., and An, G. (2017). Sepsis reconsidered: identifying novel metrics for behavioral landscape characterization with a high-performance computing implementation of an agent-based model. *J. Theor. Biol.* 430, 157–168. doi: 10.1016/j.jtbi.2017.07.016

Cockrell, C., and Axelrod, D. (2018). Optimization of dose schedules for chemotherapy of early colon cancer determined by high performance computer simulations. *Cancer Inform* 18:1176935118822804.

Cockrell, C., Christley, S., and An, G. (2014). Investigation of inflammation and tissue patterning in the gut using a spatially explicit general-purpose model of enteric tissue (SEGMEnT). *PLos Comput. Biol.* 10:e1003507. doi: 10.1371/journal.pcbi.1003507

Cockrell, C., Ozik, J., Collier, N., and An, G. (2020). Nested active learning for efficient model contextualization and parameterization: pathway to generating simulated populations using multi-scale computational models. *Simulation* 97:0037549720975075.

Cockrell, R. C., and An, G. (2018). Examining the controllability of sepsis using genetic algorithms on an agent-based model of systemic inflammation. *PLos Comput. Biol.* 14:e1005876. doi: 10.1371/journal.pcbi.1005876

Cockrell, R. C., Christley, S., Chang, E., and An, G. (2015). Towards anatomic scale agent-based modeling with a massively parallel spatially explicit general-purpose model of enteric tissue (SEGMEnT_HPC). *PLoS One* 10:e0122192. doi: 10.1371/journal.pone.0122192

Csete, M., and Doyle, J. (2004). Bow ties, metabolism and disease. *Trends Biotechnol.* 22, 446–450.

Cukier, R., Levine, H., and Shuler, K. (1978). Nonlinear sensitivity analysis of multiparameter model systems. *J. Comput. Phys.* 26, 1–42.

Fonseca, C. M., and Fleming, P. J. (1993). Genetic algorithms for multiobjective optimization: formulationdiscussion and generalization. *Icga* 93, 416–423.

Goldberg, D. E., and Holland, J. H. (1988). Genetic algorithms and machine learning. *Machine Learn.* 3, 95–99.

Gough, A., Stern, A. M., Maier, J., Lezon, T., Shun, T.-Y., Chennubhotla, C., et al. (2017). Biologically relevant heterogeneity: metrics and practical insights. *Slas Dis. Adv. Life Sci. R. D.* 22, 213–237.

Haupt, R. L., and Ellen Haupt, S. (2004). *Practical Genetic Algorithms*. Hoboken, NJ: John Wiley & Sons.

Haupt, R. L., and Haupt, S. E. (2004). *Practical Genetic Algorithms*. Hoboken, NJ: John Wiley & Sons.

Holland, J. H. (ed.) (1983). "Escaping brittleness," in *Proceedings of the Second International Workshop on Machine Learning*, (Citeseer).

Larie, D., An, G., and Cockrell, C. (2020). Artificial neural networks for disease trajectory prediction in the context of sepsis. *arXiv [preprint]* arXiv:200714542,

Liu, Z., Rexachs, D., Epelde, F., and Luque, E. (2017). A simulation and optimization based method for calibrating agent-based emergency department models under data scarcity. *Comput. Indus. Eng.* 103, 300–309.

Petersen, B. K., Yang, J., Grathwohl, W. S., Cockrell, C., Santiago, C., An, G., et al. (2019). Deep reinforcement learning and simulation as a path toward precision medicine. *J. Comput. Biol.* 26, 597–604.

Rogers, A., and Von Tessin, P. (2004). "Multi-objective calibration for agent-based models," in *Proceeding of the Agent-Based Simulation 5*.

Saltelli, A., Ratto, M., Andres, T., Campolongo, F., Cariboni, J., Gatelli, D., et al. (2008). *Global Sensitivity Analysis: the Primer*. Hoboken, NJ: John Wiley & Sons.

Saltelli, A., Tarantola, S., Campolongo, F., and Ratto, M. (2004). *Sensitivity Analysis in Practice: a Guide to Assessing Scientific Models*. England: John Wiley & Sons.

Stelling, J., Sauer, U., Iii, F., and Doyle, J. (2006). "Complexity and robustness of cellular systems," in *System Modeling in Cellular Biology*, eds Z. Szallasi, V. Periwal, and J. Stelling, (Elsevier), 3–18.

Windrum, P., Fagiolo, G., and Moneta, A. (2007). Empirical validation of agent-based models: alternatives and prospects. *J. Artif. Soc. Soc. Simulat.* 10:8.

Conflict of Interest: The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

frontiers in Physiology

METHODS
published: 25 May 2021
doi: 10.3389/fphys.2021.662314

BayCANN: Streamlining Bayesian Calibration With Artificial Neural Network Metamodeling

Hawre Jalal[1]*, *Thomas A. Trikalinos*[2] *and Fernando Alarid-Escudero*[3]

[1] Department of Health Policy and Management, University of Pittsburgh, Graduate School of Public Health, Pittsburgh, PA, United States, [2] Departments of Health Services, Policy & Practice and Biostatistics, Brown University, Providence, RI, United States, [3] Division of Public Administration, Center for Research and Teaching in Economics (CIDE), Aguascalientes, Mexico

Purpose: Bayesian calibration is generally superior to standard direct-search algorithms in that it estimates the full joint posterior distribution of the calibrated parameters. However, there are many barriers to using Bayesian calibration in health decision sciences stemming from the need to program complex models in probabilistic programming languages and the associated computational burden of applying Bayesian calibration. In this paper, we propose to use artificial neural networks (ANN) as one practical solution to these challenges.

Methods: Bayesian Calibration using Artificial Neural Networks (BayCANN) involves (1) training an ANN metamodel on a sample of model inputs and outputs, and (2) then calibrating the trained ANN metamodel instead of the full model in a probabilistic programming language to obtain the posterior joint distribution of the calibrated parameters. We illustrate BayCANN using a colorectal cancer natural history model. We conduct a confirmatory simulation analysis by first obtaining parameter estimates from the literature and then using them to generate adenoma prevalence and cancer incidence targets. We compare the performance of BayCANN in recovering these "true" parameter values against performing a Bayesian calibration directly on the simulation model using an incremental mixture importance sampling (IMIS) algorithm.

Results: We were able to apply BayCANN using only a dataset of the model inputs and outputs and minor modification of BayCANN's code. In this example, BayCANN was slightly more accurate in recovering the true posterior parameter estimates compared to IMIS. Obtaining the dataset of samples, and running BayCANN took 15 min compared to the IMIS which took 80 min. In applications involving computationally more expensive simulations (e.g., microsimulations), BayCANN may offer higher relative speed gains.

Conclusions: BayCANN only uses a dataset of model inputs and outputs to obtain the calibrated joint parameter distributions. Thus, it can be adapted to models of various levels of complexity with minor or no change to its structure. In addition, BayCANN's efficiency can be especially useful in computationally expensive models. To facilitate BayCANN's wider adoption, we provide BayCANN's open-source implementation in R and Stan.

Keywords: Bayesian calibration, machine learning, mechanistic models, artificial neural networks, emulators, surrogate models, metamodels

OPEN ACCESS

Edited by:
Michael Döllinger, University Hospital Erlangen, Germany

Reviewed by:
Koen Degeling, The University of Melbourne, Australia
Dominic G. Whittaker, University of Nottingham, United Kingdom

***Correspondence:**
Hawre Jalal
hjalal@pitt.edu

Specialty section:
This article was submitted to Computational Physiology and Medicine, a section of the journal Frontiers in Physiology

Received: *01 February 2021*
Accepted: *20 April 2021*
Published: *25 May 2021*

Citation:
Jalal H, Trikalinos TA and Alarid-Escudero F (2021) BayCANN: Streamlining Bayesian Calibration With Artificial Neural Network Metamodeling. Front. Physiol. 12:662314. doi: 10.3389/fphys.2021.662314

1. BACKGROUND

Modelers and decision-makers often use mathematical simulation models to simplify real-life complexity and inform decisions, particularly those for which uncertainty is inherent. However, some of the model parameters might be either unobserved or unobservable due to various financial, practical or ethical reasons. For example, a model that simulates the natural history of cancer progression may lack an estimate for the rate at which an individual transitions from a pre-symptomatic cancer state to becoming symptomatic. Although this rate might not be directly observable, it may be estimated using a technique commonly referred to as calibration (Alarid-Escudero et al., 2018; Vanni et al., 2011; Rutter et al., 2009). Thus, calibration involves modifying the model input parameters until the desired output is obtained.

Calibration has the potential for improving model inference, and recent guidelines recommend that model calibration of unknown parameters should be performed where data on outputs exist (Weinstein et al., 2003; Briggs et al., 2012). Modelers are also encouraged to report the uncertainty around calibrated parameters and use these uncertainties in both deterministic and probabilistic sensitivity analyses (Briggs et al., 2012).

There are several calibration techniques with various levels of complexity. For example, Nelder-Mead is a direct-search algorithm commonly used to calibrate models in health and medicine. Nelder-Mead is a deterministic approach that searches the parameter space for good-fitting parameter values (Nelder and Mead, 1965). Although Nelder-Mead is generally effective, it cannot estimate parameter distributions or directly inform on the correlations among the calibrated parameters. It is also not guaranteed to find a global optimal value because it might converge to a local optimum. Unlike the direct-search algorithms, Bayesian methods are naturally suited for calibration because they estimate the input parameter's posterior joint and marginal distributions (Menzies et al., 2017). However, Bayesian methods are difficult to implement due to the complexity of the models used and the computational challenges of applying these methods. Bayesian calibration often requires tens or hundreds of thousands of simulation runs and a model written in a probabilistic programming language, such as Stan (Carpenter et al., 2017) or Bayesian inference Using Gibbs Sampling (BUGS) (Lunn et al., 2009). We argue that the complexity of these tasks and their potential computational demand have prevented a wider adoption of Bayesian calibration methods in health decision science models.

In this manuscript, we use artificial neural network (ANN) metamodels as a practical approach to streamlining Bayesian calibration in complex simulation models. Metamodels have increasingly been used to overcome the computational burden of Bayesian calibration. A metamodel is a surrogate model that can be used to approximate the model's input-output relationship (Jalal et al., 2013). Metamodels can provide an approximation to the simulation model in a fraction of the time. While ANN metamodels are not fully probabilistic, they are flexible functions that can map highly non-linear relationships in large data. We use an ANN metamodel as an emulator to substitute the simulation model in the Bayesian calibration analysis. Thus, the ANN acts as a low computational cost proxy of the simulation model. In addition, analysts do not need to program their simulation model in a probabilistic language because coding the ANN in probabilistic languages (e.g., Stan) is relatively straight-forward, and analysts can reuse the provided Stan code with little or no modification.

We refer to our approach as Bayesian calibration via artificial neural networks, or BayCANN for short. We demonstrate BayCANN by calibrating a realistic model of the natural history of colorectal cancer (CRC). We compare this approach's results to an approximate Bayesian calibration of the original model using an incremental mixture importance sampling (IMIS) algorithm. We provide the code in R and Stan for our application that researchers can adapt to calibrate their models.

2. METHODS

We start this exposition by reviewing elements of Bayesian calibration. We describe the computational burden of using Bayes theorem in most realistic models, and how deep ANNs can streamline Bayesian calibration methods to calibrate these models. We illustrate this approach by calibrating a natural history model of CRC. We also compare BayCANN's performance to a Bayesian calibration using IMIS directly on a simulation model.

2.1. Bayesian Calibration

The Bayes theorem states that

$$p(\theta|\text{data}) = \frac{p(\text{data}|\theta)p(\theta)}{p(data)}, \tag{1}$$

where θ is a set of model parameters, data is the observed data, and $p(\text{data}|\theta)$ is the same as the likelihood $l(\theta|\text{data})$. Because the denominator is not a function of θ, we can rewrite Equation (1) as

$$p(\theta|\text{data}) \propto l(\theta|\text{data})p(\theta). \tag{2}$$

Table 1 shows how each term in Equation (2) can be mapped to a component in a calibration exercise. The prior distribution, $p(\theta)$, represents our uncertainty about the distribution of the model parameters before calibrating the model. Modelers often use various distributions to describe this uncertainty, including beta or logit-normal distribution for probabilities, gamma for rates, or log-normal distributions for rates or hazard ratios. Thus, we can think of a prior distribution as the uncertainty of the pre-calibrated model input parameters. For example, we can represent a vague distribution by a uniform distribution where all the values are equally likely within a defined range.

Bayesian calibration will update the prior distribution based on the observed target data. The term $p(\theta|data)$ is called the posterior distribution, representing the updated distribution of θ after observing some data. The posterior distribution is equivalent to the calibrated parameter distribution when the data are the calibration targets.

The likelihood function, $l(\theta|data)$, denotes how likely the observed data arise from a given data generation mechanism

TABLE 1 | The Bayes formula in a calibration context.

Term	Bayesian context	Calibration context
$p(\theta)$	Prior distribution of the model input parameters θ	Pre-calibrated model input parameters
$p(\theta\|\text{data})$	Posterior distribution of the model parameters θ given observed data	Calibrated model parameters to target data
$l(\theta\|\text{data})$	Probability of the data given model parameters θ (model likelihood)	Objective function or goodness-of-fit measure; how well the model output fits the target data given a particular value of θ

with a parameter set values θ. From a simulation modeling perspective, $l(\theta|data)$ is equivalent to measuring the goodness of the model output fit to the calibration targets given a simulation model's input parameter set θ.

Thus, we can map all components of Bayes theorem to calibration components and use Bayesian inference to obtain the calibrated parameter distributions (a.k.a. the posterior distributions).

Bayesian calibration is often challenging to adopt in practice in health decision science models. The main challenge lies in the complexity of applying Equation (2). Specifically, an analytical solution for $p(\theta|\text{data})$ is unlikely to exist for most realistic simulation models. Thus, specialized algorithms, such as Markov-Chain Monte-Carlo (MCMC) might be necessary at the expense of being practically challenging to implement for complex models and computationally expensive.

2.2. Metamodels

To overcome the computational and practical challenges of Bayesian calibration, we propose to use artificial neural network (ANN) metamodels. As described above, a metamodel is a surrogate model that approximates the relationship between the simulation model's inputs and outputs (i.e., a metamodel is a model of the model) (Blanning, 1974; Kleijnen, 1975; Kleijnen et al., 2005; Kleijnen, 2015). Metamodels range from simple models, such as linear regressions, to complex non-linear models, such as artificial neural networks (ANN). Although linear regression models are the most common form of metamodels (Barton and Meckesheimer, 2006; Barton, 2009; Sacks et al., 1989; Fu, 1994; Weiser Friedman, 1996; Banks, 1998; Kleijnen and Sargent, 2000; Jalal et al., 2013, 2015), in this paper we focus on ANN because they are more flexible while still being relatively simple to implement in Stan or BUGS.

Metamodels are often used because they generally offer a vast reduction in computation time (Kleijnen, 1979; Friedman and Pressman, 1988; Barton, 1992; Weiser Friedman, 1996; O'Hagan et al., 1999; Barton and Meckesheimer, 2006; Santos and Santos, 2007; Reis dos Santos and Reis dos Santos, 2009; Khuri and Mukhopadhyay, 2010). For example, a model that takes several hours or even days to run can be approximated with a metamodel that may only take a few milliseconds. This feature has been an attractive attribute of metamodels for many decades in engineering and computer science. Examples of metamodels in health decision sciences involve revealing model uncertainty using linear regression mdetamodeling (Jalal et al., 2013), and speeding up computationally expensive microsimulation models using Gaussian processes metamodeling (Stevenson et al., 2004; de Carvalho et al., 2019).

An additional benefit of using metamodels for Bayesian calibration is that one can reuse the same metamodel structure to calibrate very different simulation models. The same BayCANN code can be adapted to other problems with no or minimal change.

2.2.1. ANN Metamodels

Artificial neural networks (ANNs) are networks of non-linear regressions that were initially developed to mimic the neural signal processing in the brain and to model how the nervous system processes complex information (Másson and Wang, 1990; Michie et al., 1994; Rojas, 1996; Jain et al., 1996; Olden et al., 2008). ANNs have recently witnessed significant advances for applications in machine learning, artificial intelligence, and pattern recognition (Ravì et al., 2016).

Figure 1 illustrates the basic structure of a four-layer neural network with two hidden layers with I neurons (nodes) in the input layer, J hidden nodes in the first hidden layer, K hidden nodes in the second hidden layer, and O output nodes in the output layer. The ANNs with more than one hidden layer are often referred to as deep ANNs. The following sets of equations represent the structure of this ANN

$$\begin{aligned} z^{(1)} &= W^{(1)}\theta + b^{(1)} \\ h^{(1)} &= f^{(1)}\left(z^{(1)}\right) \\ z^{(2)} &= W^{(2)}h^{(1)} + b^{(2)} \\ h^{(2)} &= f^{(2)}\left(z^{(2)}\right) \\ z^{(3)} &= W^{(3)}h^{(2)} + b^{(3)} \\ Y &= f^{(3)}\left(z^{(3)}\right), \end{aligned} \tag{3}$$

where θ is the simulation model inputs, Y is the model outputs to be compared to the calibrated targets, and $(W, b) = \left(W^{(1)}, b^{(1)}, W^{(2)}, b^{(2)}, W^{(3)}, b^{(3)}\right)$ are the ANN coefficients. $W^{(1)}$ are the weights connecting the inputs θ with the nodes in the first hidden layer, $W^{(2)}$ are the weights connecting the nodes in the first and second hidden layers, and $W^{(3)}$ are the weights connecting the nodes in the second hidden layer with the output Y. The terms $b^{(1)}$, $b^{(2)}$ and $b^{(3)}$ are corresponding bias (intercept) terms. $f^{(1)}$ is the activation function, commonly, a sigmoid function such as a hyperbolic tangent function

$$f^{(1)}\left(z^{(1)}\right) = \frac{2}{1 + e^{-2z^{(1)}}} - 1. \tag{4}$$

The function $f^{(3)}$ is called a transfer function that transforms the results from the last hidden layer's nodes into a working output.

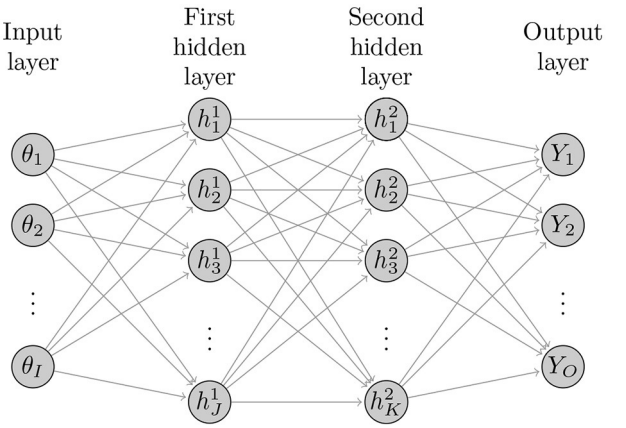


FIGURE 1 | Diagram of general structure of a deep neural network with *I* inputs, two hidden layers with *J* and *K* hidden nodes and *O* outputs.

The transfer function can also be a sigmoid function or a simple linear function. Thus, the $z^{(1)}, z^{(2)}$ and $z^{(3)}$ are the weighted sum of inputs from the input layer and the first and second hidden layers, respectively. ANNs can be made more flexible by increasing the number of hidden layers and/or the number of nodes in these layers.

2.3. BayCANN Algorithm

We implement BayCANN with TensorFlow to fit the ANN and Stan to obtain the parameter's posterior distributions. We use the package `keras` in R to create ANN metamodels that approximate the relationship between our model's input parameters and outputs and estimate the coefficients b and W (R Core Team, 2018; Jalal et al., 2017). We built the ANN from a set of probabilistic samples using a Latin hypercube sampling (LHS) design of experiment (DoE) to efficiently sample the input parameter space. Once we obtain the ANN coefficients, we perform the Bayesian calibration using the ANN rather than the simulation model.

We implemented the deep ANN in Stan (Carpenter et al., 2017) which uses a guided MCMC using gradient descent, referred to as Hamiltonian Monte-Carlo. Similarly, the R package `rstan`.

Both TensorFlow and Stan utilize multithreading; thus, it is essential to ensure sufficient memory is available for all threads to run efficiently.

Below we outline the steps to conduct BayCANN.

1. Structure the simulation model such that it produces outputs corresponding to the calibration targets. For example, if calibration targets are disease incidence or prevalence, ensure the model generates these outputs.
2. Generate two datasets of input parameter sets—one for training the ANN (training dataset) and the second for validating it (validation dataset). The analyst could use an LHS to efficiently sample the model inputs' prior distributions.
3. Run the simulation model using both training and validation datasets to generate their corresponding simulation model outputs.
4. Train an ANN using the training dataset, and validate it using the validation dataset. Obtaining a high-fidelity ANN is crucial to ensure getting accurate and reliable results from BayCANN (Degeling et al., 2020). Adjust the ANN's structure to obtain an accurate metamodel before proceeding.
5. Perform the Bayesian calibration by passing the ANN coefficients W and b, the prior input parameter samples, and the calibration targets to the programmed ANN framework in Stan. Stan then returns the joint posterior distribution of the calibrated parameters.

The code for implementing BayCANN is available on GitHub at https://github.com/hjalal/BayCANN. In the case study below, we use BayCANN to calibrate a colorectal cancer natural history model.

2.4. Case Study: Natural History Model of Colorectal Cancer

We use BayCANN to calibrate a state-transition model (STM) of the natural history of colorectal cancer (CRC) implemented in R (Jalal et al., 2017). We refer to our model as CRCModR. CRCModR is a discrete-time STM based on a model structure originally proposed by (Wu et al., 2006) that has previously been used for testing other methods (Alarid-Escudero et al., 2018; Heath et al., 2020). Briefly, CRCModR has 9 different health states that include absence of the disease, small and large precancerous lesions (i.e., adenomatous polyps), and early and late preclinical and clinical cancer states by stage. **Figure 2** shows the state-transition diagram of the model. The progression between health states follows a continuous-time age-dependent Markov process. There are two age-dependent transition intensities (i.e., transition rates), $\lambda_1(a)$ and $\mu(a)$, that govern the age of onset of adenomas and all-cause mortality, respectively. Following Wu's original specification (Wu et al., 2006), we specify $\lambda_1(a)$ as a Weibull hazard such that

$$\lambda_1(a) = l\gamma a^{\gamma-1}, \tag{5}$$

where l and γ are the scale and shape parameters of the Weibull hazard model, respectively. The model simulates two adenoma categories: small (adenoma smaller than or equal to 1 cm in size) and large (an adenoma larger than 1 cm in size). All adenomas start small and can transition to the large size category at a constant annual rate λ_2. Large adenomas may become preclinical CRC at a constant annual rate λ_3. Both small and large adenomas may progress to preclinical CRC, although most will not in an individual's lifetime. Early preclinical cancers progress to late stages at a constant annual rate λ_4 and could become symptomatic at a constant annual rate λ_5. Late preclinical cancer could become symptomatic at a constant annual rate λ_6. After clinical detection, the model simulates the survival time to death from early and late CRC using time-homogeneous mortality rates, λ_7 and λ_8, respectively. In total, the model has nine health states: normal, small adenoma, large adenoma, early preclinical CRC, late

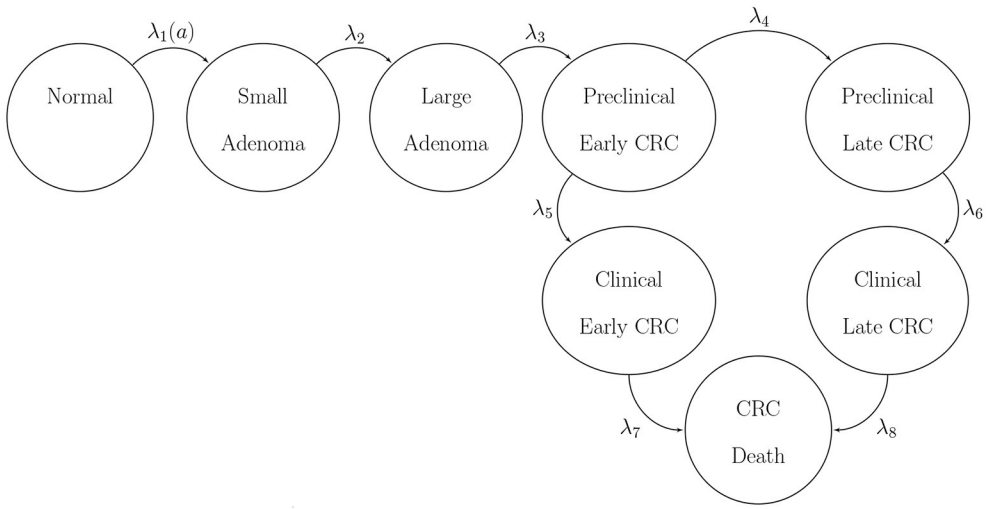


FIGURE 2 | State-transition diagram of the natural history model of colorectal cancer. Ovals represent health states and arrows represent transitions. All states can transition to death from causes other than CRC with rate $\mu(a)$. CRC, colorectal cancer.

TABLE 2 | The parameters of the natural history model of colorectal cancer (CRC).

Parameter	Description	Base value	Calibrate?	Source	Prior range
l	Scale parameter of Weibull hazard	2.86e-06	Yes	Wu et al., 2006	$[2 \times 10^{-6}, 2 \times 10^{-5}]$
g	Shape parameter of Weibull hazard	2.78	Yes	Wu et al., 2006	[2.00, 4.00]
λ_2	Small adenoma to large adenoma	0.0346	Yes	Wu et al., 2006	[0.01, 0.10]
λ_3	Large adenoma to preclinical early CRC	0.0215	Yes	Wu et al., 2006	[0.01, 0.04]
λ_4	Preclinical early to preclinical late CRC	0.3697	Yes	Wu et al., 2006	[0.20, 0.50]
λ_5	Preclinical early to clinical early CRC	0.2382	Yes	Wu et al., 2006	[0.20, 0.30]
λ_6	Preclinical late to clinical late CRC	0.4852	Yes	Wu et al., 2006	[0.30, 0.70]
λ_7	CRC mortality in early stage	0.0302	No	Wu et al., 2006	-
λ_8	CRC mortality in late stage	0.2099	No	Wu et al., 2006	-
p_{adeno}	Prevalence of adenoma at age 50	0.27	Yes	Rutter et al., 2007	[0.25, 0.35]
p_{small}	Proportion of small adenomas at age 50	0.71	Yes	Wu et al., 2006	[0.38, 0.95]

The base values are used to generate the calibration targets and the ranges of the uniform distribution used as priors for the Bayesian calibration.

preclinical CRC, CRC death, and other causes of death. The state-transition diagram of the model is shown in **Figure 2**. The model simulates the natural history of CRC of a hypothetical cohort of 50-year-old women in the U.S. over a lifetime. The cohort starts the simulation with a prevalence of adenoma of p_{adeno}. A proportion, p_{small}, corresponds to small adenomas and prevalence of preclinical early and late CRC of 0.12 and 0.08, respectively. The simulated cohort in any state is at risk of all-cause mortality $\mu(a)$ obtained from the U.S. life tables Arias (2014). Similar models to CRCmodR have been used to inform population-level screening guidelines in the U.S. (Knudsen et al., 2016).

CRCModR has 11 parameters summarized in **Table 2** (Alarid-Escudero et al., 2018). Mortality rates from early and late stages of CRC (λ_7, λ_8]) could be obtained from cancer population registries (e.g., SEER in the U.S.). Thus, we calibrate the model to the remaining nine parameters (p_{adeno}, p_{small}, l,γ, λ_2, λ_3, λ_4, λ_5 and λ_6).

2.4.1. Confirmatory Analysis

We conducted a confirmatory analysis to compare BayCANN vs. IMIS. To obtain the "truth" that we could compare BayCANN and IMIS against, we *generated* the synthetic targets using the base-case values in **Table 2**. We generated four age-specific targets, including adenoma prevalence, the proportion of small adenomas, and CRC incidence for early and late stages which represent commonly used calibration targets for this type of model (Kuntz et al., 2011). To generate the calibration targets, we ran CRCModR as a microsimulation (Krijkamp et al., 2018) 100 times to produce different adenoma-related and cancer incidence outputs. We then aggregated the results across all 100 outputs to compute their mean and standard errors (SE). Different calibration targets could have different levels of uncertainty given the amount of data to compute their summary measures. Therefore, to account for different variations in the amount of data on different calibration targets, we simulated different numbers of individuals for adenoma targets ($N = 500$) and

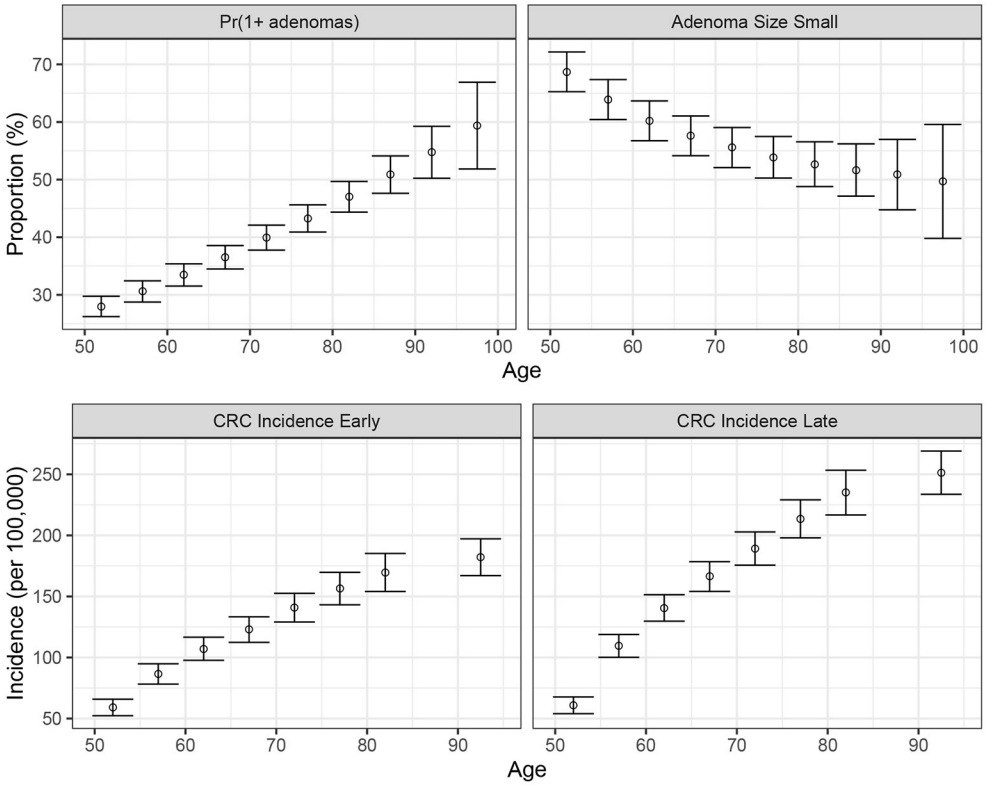


FIGURE 3 | Generated calibration targets and its 95% credible interval of a cohort of 500 and 100,000 simulated individuals for adenoma-related targets cancer incidence targets, respectively plotted against age in years on the x-axis. These distributions are from 100 different runs using the same parameter set values in each set of runs.

cancer incidence targets ($N = 100,000$). **Figure 3** shows the generated adenoma-related and cancer incidence calibration targets aggregated over 100 different runs using the parameter set in **Table 2**.

To create a deep ANN metamodel, we generated a DOE by sampling each of the nine parameters from the ranges of the uniform distributions using an LHS design as shown in **Table 2**. Specifically, we created two LHS input datasets of sizes 8,000 samples and 2,000 samples for training and validating the ANN, respectively. We then ran the natural history model and generated adenoma prevalence and CRC incidence for each parameter set.

We define an ANN with two hidden layers and 100 nodes per hidden layer. Then, we evaluated the ANN's performance by validating the predicted values for the 36 outcomes against the observed values from the validation dataset. The likelihood function for BayCANN was constructed by assuming that the targets, y_{t_i}, are normally distributed with mean ϕ_{t_i} and standard deviation σ_{t_i}, where $\phi_{t_i} = M[\theta]$ is the model-predicted output for each type of target t and age group i at parameter set θ. We defined uniform prior distributions for all θ_u based on previous knowledge or nature of the parameters (**Table 2**).

We compare BayCANN against a full Bayesian calibration of the natural history model using the incremental mixture importance sampling (IMIS) algorithm. The IMIS algorithm has been described elsewhere (Raftery and Bao, 2010), but briefly, this algorithm reduces the computational burden of Bayesian calibration by incrementally building a better importance sampling function based on Gaussian mixtures.

3. RESULTS

We present the ANN's performance in approximating the output of the simulation model and compare the generated joint posterior distribution of the simulation model parameters produced from BayCANN against the full joint posterior from the IMIS approach. We compare both BayCANN and IMIS results recovering the "true" values—the parameter values we used to generate the calibration targets in the confirmatory analysis.

3.1. Validation

Figure 4 illustrates the ANN's performance in predicting the model outputs using the validation dataset. Each plot represents one of the model outputs, where we compare the ANN's prediction on the y-axis against the model's output on the x-axis. Each red dot represents one of the 2,000 DOE validation samples not used to train the ANN. The ANN had a high prediction performance in approximating the model outputs

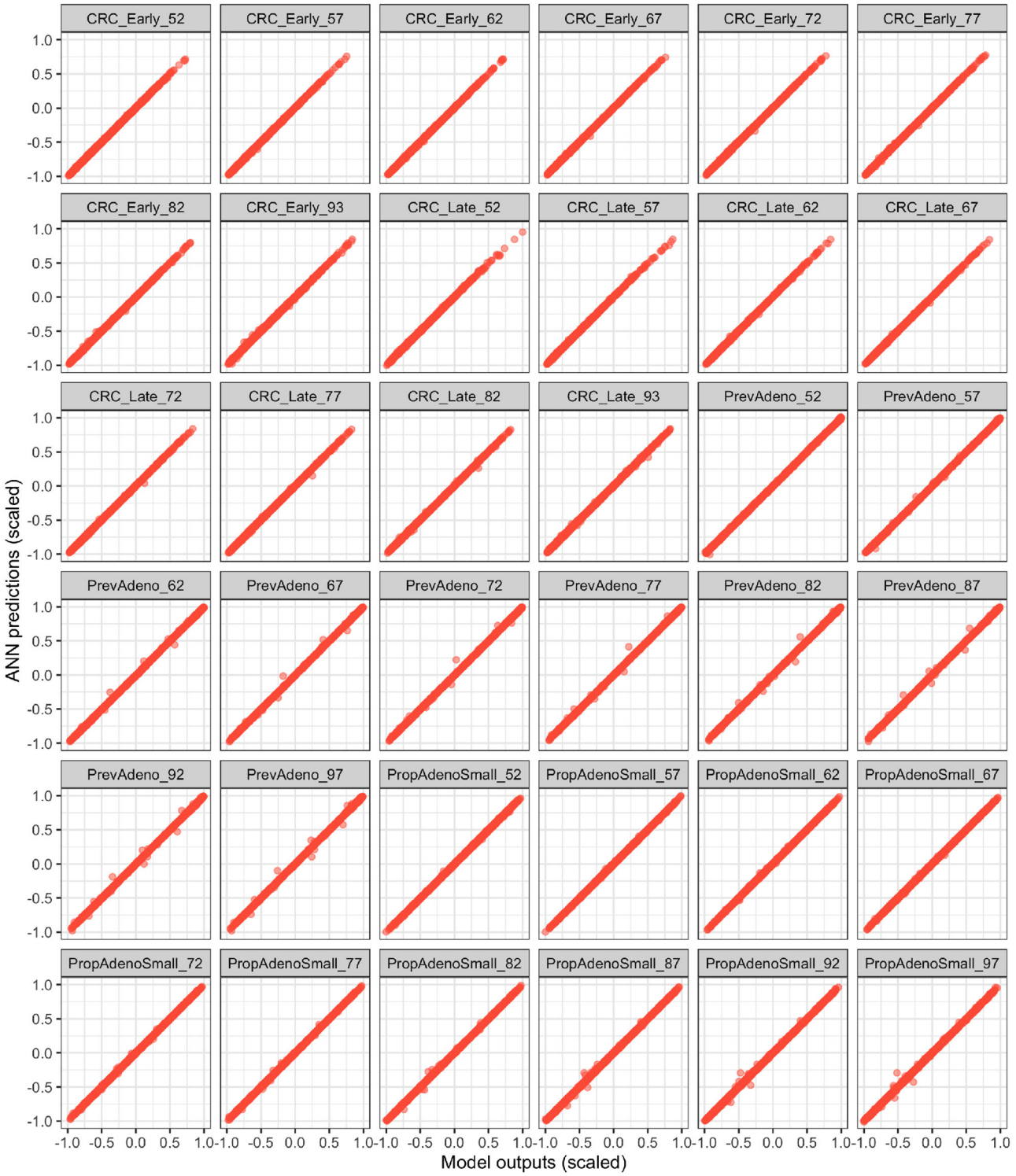


FIGURE 4 | Validation of the fitted ANN on the validation Latin hyper cube sample (LHS) dataset. The x and y axes represent the scaled model outputs and scaled ANN predictions, respectively.

($R^2 > 99.9\%$), indicating that the deep ANN is a high fidelity metamodel of the simulation model within the parameter ranges we evaluated.

3.2. Comparing BayCANN and IMIS

Figure 5 compares BayCANN against IMIS in recovering the true parameter values used to generate the targets. The 95% credible

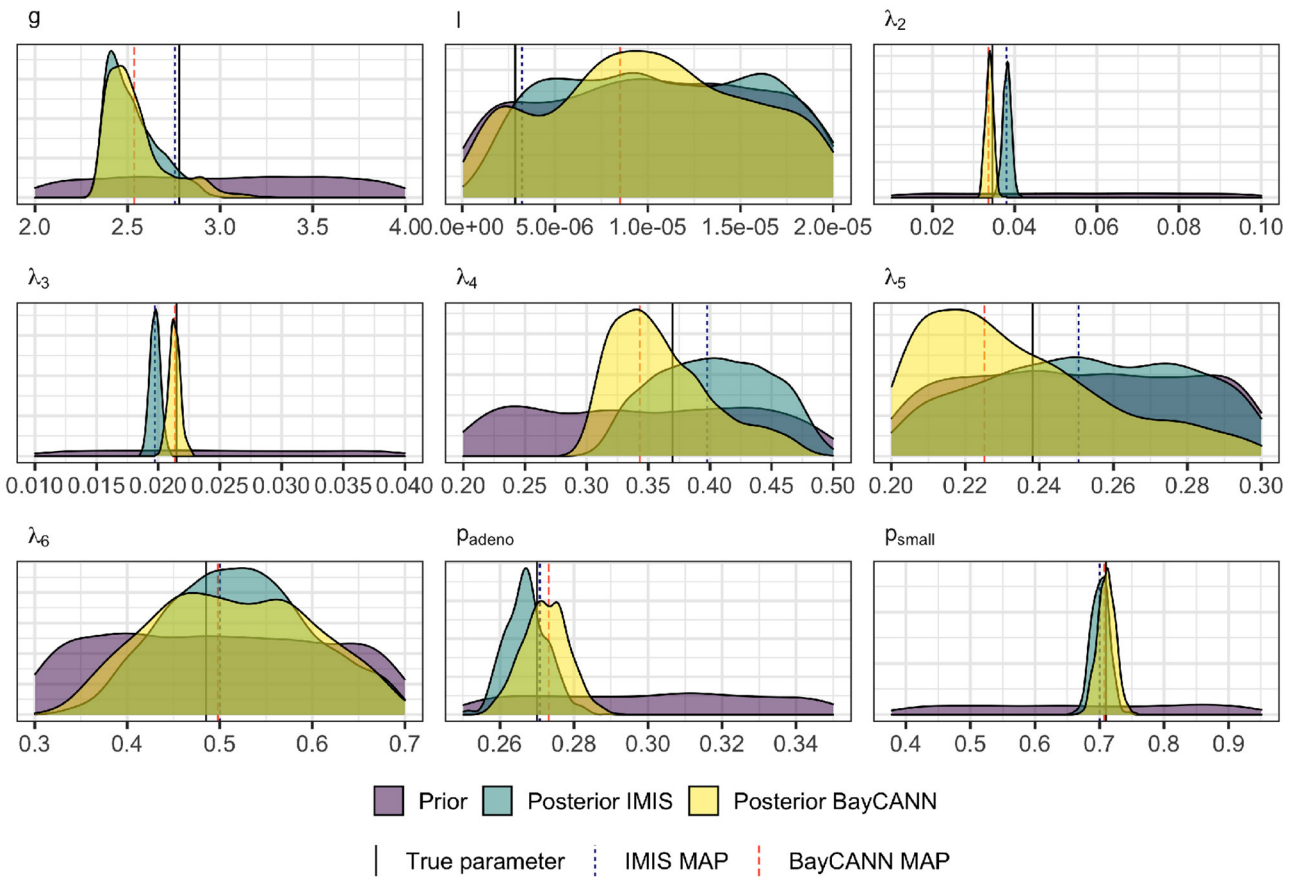


FIGURE 5 | Prior and Marginal posterior distributions of the calibrated parameters from the IMIS and BayCANN methods. The vertical solid lines indicate the "true" parameter values (i.e., the value of the parameters used to generate the calibration targets in the confirmatory analysis). The vertical dashed lines represent the maximum a posteriori (MAP) for BayCANN and the incremental mixture importance sampling (IMIS).

intervals (CrI) of each parameter distribution obtained from BayCANN cover all nine true parameters. For IMIS, the 95% CrI did not cover the true parameters for λ_2 and λ_3. This figure also shows the maximum a posteriori (MAP) estimate for both BayCANN and IMIS. The MAP is the sample associated with the highest log-posterior and indicates the posterior parameter set that best fits the target data.

Figure 6 compares the results of BayCANN against all the calibration targets for the probability of developing multiple adenomas, the proportion of small adenomas, and early and late clinical CRC incidence. Upon visual inspection, BayCANN fits all calibration targets well, indicating that the joint posterior distribution from BayCANN can produce targets in the desired ranges. The results here represent the model-predictive mean and the credible interval of using 10,000 posterior samples from BayCANN. We also present the results of using BayCANN's MAP estimates which closely follow the model-predicted posterior mean from the 10,000 posterior samples.

In this example, BayCANN was five times faster than the IMIS. The IMIS algorithm took 80 min to run in a MacBook Pro Retina, 15-inch, Late 2013 with a 2.6 GHz Intel Core i7 processor with 4 cores and 16 gigabytes of RAM. BayCANN took only 15 min on the same computer; 5 min to produce 10,000 samples for both LHS DOE dataset generations and about 10 min to fit the ANN in TensorFlow and produce the joint posterior distributions in Stan. The computational gain of BayCANN was modest given that our case study model was efficient and deterministic.

Figure 7 presents the joint distribution of all pairwise parameters in BayCANN, and along the diagonal, the marginal distributions of each parameter. This figure reveals insightful information about this calibration exercise. In practice, many calibrated parameters are correlated as shown in this figure. The absolute value of these correlations range from 0.013 to 0.963. The strength of the correlation reflects the level of non-identifiability between that pair of parameters. The stronger the correlation the higher the non-identifiability and the greater need to add additional target data or modify the model structure to *separate* the parameters in question (Alarid-Escudero et al., 2018).

4. DISCUSSION

In this study, we propose BayCANN as a feasible and practical solution to Bayesian calibration challenges in complex health

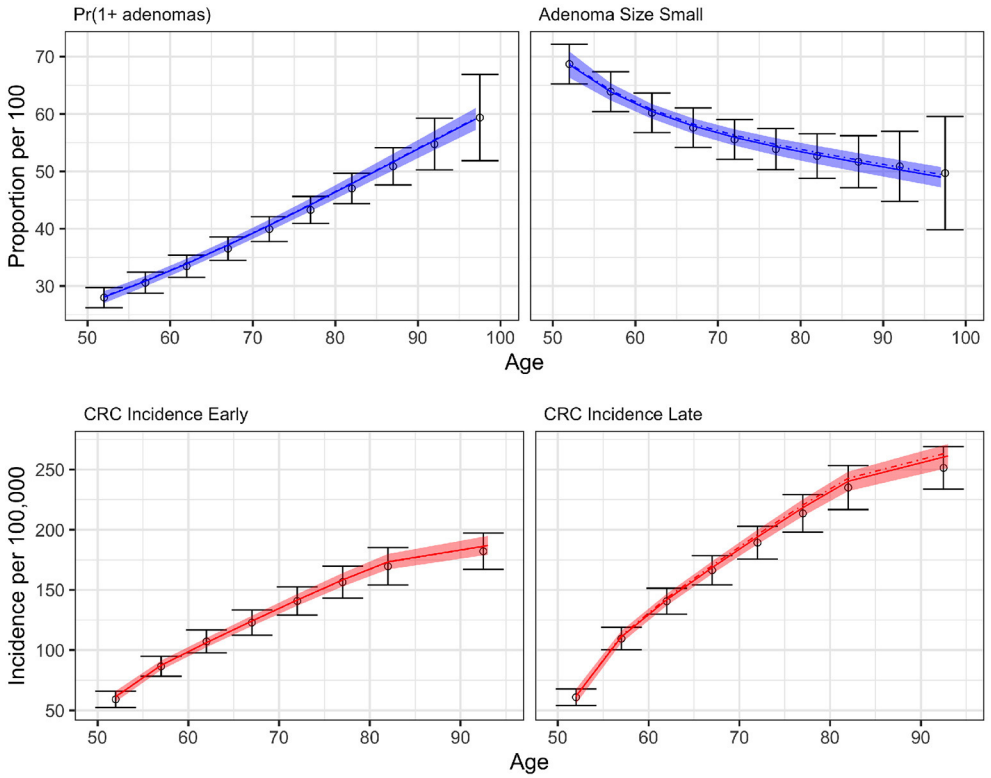


FIGURE 6 | BayCANN calibration results by age in years on the x-axis. The upper panels show adenoma targets and lower panels show cancer incidence targets by stage. Calibration targets with their 95% confidence intervals are shown in black. The colored curves show the posterior model-predicted mean, and the shaded area shows the corresponding 95% posterior model-predicted credible interval of the outcomes. The dashed-dotted lines represent the output using the maximum a posteriori (MAP) estimate from BayCANN.

decision science models. The distinct advantage of using BayCANN is that it represents the model on a functional basis as an ANN. Then, the ANN can become a high-fidelity representation of the model. Thus, those interested in implementing BayCANN can do so without the need to code their models in a probabilistic programming language. Given the high computational efficiency of the ANN, BayCANN can also provide a computational advantages over other Bayesian calibration methods.

BayCANN uses ANNs specifically to streamline Bayesian calibration. ANNs have also been used as metamodels of both stochastic and deterministic responses, mainly for their computational efficiency (Barton, 2009; Badiru and Sieger, 1998; Hurrion, 1997; Chambers and Mount-Campbell, 2002; Zobel and Keeling, 2008). One of the first implementations of ANN as metamodels was in 1992 for a scheduling simulation model (Pierreval et al., 1992; Pierreval and Huntsinger, 1992). Since then, ANNs have been successfully implemented as emulators of all sorts of discrete-event and continuous simulation models in a wide variety of fields (Kilmer, 1996; Sabuncuoglu and Touhami, 2002; Fonseca et al., 2003; El Tabach et al., 2007). ANNs have also been proposed as proxies for non-linear and simulation models (Paiva et al., 2010; Mareš and Kučerová, 2012; Pichler et al., 2003). An example of ANNs as metamodels is estimating the mean and variance of patient time in emergency department visits (Kilmer, 1994; Kilmer et al., 1997). Nowadays, ANNs are widely popular as machine learning tools in artificial intelligence (Schmidhuber, 2015). Deep learning using ANNs are used for visual recognition in self-driving cars (Ndikumana et al., 2020) and in classifying galaxies (Folkes et al., 1996). ANNs have been used for calibration of computationally expensive models, such as general circulation and rainfall-runoff models in climate science (Khu et al., 2004; Hauser et al., 2012), and other complex global optimization techniques such as genetic algorithms (Wang, 2005).

The superior performance of BayCANN relative to IMIS may pertain to the bias of the ANN in BayCANN being relatively lower than that of the Bayesian approximation of IMIS. BayCANN uses ANNs as high-fidelity metamodels of the simulator and conducts full Bayesian calibration. However, IMIS is an approximation of Bayesian inference that directly uses the simulator itself. Thus, visual examination of the ANN's performance similar to **Figure 4** is an important step to ensure obtaining high-fidelity ANN for BayCANN.

Bayesian calibration provides other practical advantages over direct-search algorithms because the samples from the joint posterior distribution can be used directly as inputs to probabilistic sensitivity analyses (PSA) which are now required for cost-effectiveness analyses (Neumann et al., 2016; Rutter

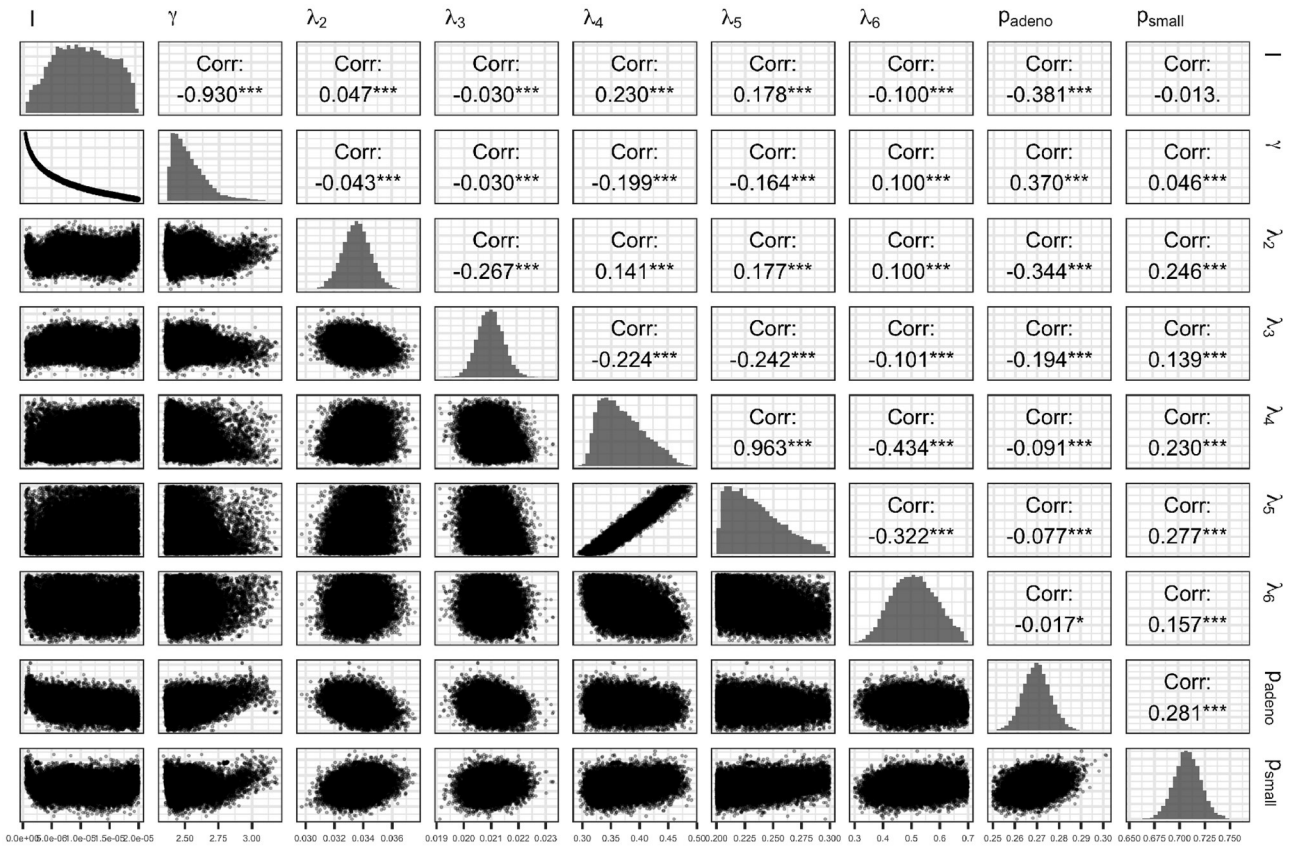


FIGURE 7 | Joint posterior distribution of the calibrated parameters of the case study using the ANN method.

et al., 2019). This joint posterior distribution is also informative in non-identifiable calibration problems where calibration targets are not sufficient to provide a unique solution to the calibrated parameters. Non-identifiability is often overlooked using standard non-Bayesian calibration approaches (Alarid-Escudero et al., 2018).

In our case study, BayCANN was both faster and overall more accurate in recovering the true parameter values than the IMIS algorithm. We developed BayCANN to be generalizable to models of various complexities, and we provide the open-source implementation in R and Stan to facilitate its wider adoption.

BayCANN may have an additional advantage for representing models with first-order Monte-Carlo noise from individual-based state-transition models (iSTM). Traditionally, calibrating these models has been challenging because of (1) the stochasticity of each simulation due to the simulator's output varying given the same set of input parameter values, and (2) the extra computational burden involved in calibrating iSTM. Because BayCANN averages over a set of simulations, it can account for the first-order Monte-Carlo noise. Further research is needed to study BayCANN's performance in stochastic models.

We chose ANNs over other metamodeling techniques because of their flexibility, efficiency, and ability to accept a large number of inputs. The use of metamodels in Bayesian calibration has been mostly limited to Gaussian processes (GP) (Kennedy and O'Hagan, 2001; Gramacy, 2020). GPs are attractive because they can be specified fully as Bayesian models (Kennedy and O'Hagan, 2001). However, GPs are not without limitations, the main one being that they are themselves relatively computationally expensive. In practice, computational challenges limit training GPs to datasets in the low thousands limiting their applicability to health decision sciences models (Gramacy, 2020).

Our approach has some limitations. First, ANNs are not fully probabilistic, thus, the joint posterior distribution produced from the Bayesian calibration is an approximation of the true distribution. Other metamodels, such as GPs are fully probabilistic and can produce the full joint posterior distribution (Gramacy, 2020). However, applying GPs in complex models can be computationally infeasible (Gramacy, 2020). Second, accuracy—Because ANNs (and GPs) are metamodels, they may rarely achieve 100% precision compared to using the simulation model itself. In our example, with a relatively simple ANN (only two hidden layers with 100 hidden nodes each), we were able to achieve 99.9% accuracy. However, for other application, the accuracy of the ANN might be lower especially if the model outputs are not continuous or smooth in certain region of the parameter space. In addition, over-fitting can be a serious problem with any metamodel especially when the purpose of

the metamodel is as sensitive as calibration. To reduce the chance of over-fitting, we validated the model against a subset of simulation runs. We visually inspected the degree of fit for the simulation output against those predicted by the ANN (**Figure 4**). Third, similar to any Bayesian model, the choice of priors can be important. Fortunately, in health decision sciences' models, analysts often make careful choices of their priors when designing their models and running PSA analyses. Additionally, the best-fitting parameters may be outside the simulated ranges. Notably, the joint posterior distribution can give insights into the parameter ranges. For example, if a parameter is skewed heavily without a clear peak, that may indicate that the parameter range needs to be shifted to cover values that may fit better. This process is usually iterative and may involve multiple steps or redefining the parameter ranges and recalibrating the model. Finally, there is no strict guideline for choosing the number of hidden ANN layers or the number of nodes per layer. In this study, we chose an ANN with two hidden layers and 100 nodes per layer. Adjusting these parameters and additional parameters of the Bayesian calibration process can improve the calibration results and can be easily changed in BayCANN. While determining these values apriori can be challenging, we recommend modelers who wish to use BayCANN to start with simple settings and gradually increase the complexity of the ANN to accommodate their particular needs. We provide flexible code in R and Stan to simplify these tasks.

In summary, Bayesian calibration can reveal important insights into model parameter values and produce outcomes that match observed data. BayCANN is one effort to target the computational and technical challenges of Bayesian calibration for complex models.

DATA AVAILABILITY STATEMENT

The original contributions presented in the study are included in the article and the code for BayCANN is available at https://github.com/hjalal/BayCANN, further inquiries can be directed to the corresponding author.

AUTHOR CONTRIBUTIONS

HJ and FA-E conceived the study. HJ, FA-E, and TT conducted the analyses, contributed to interpreting the results, and writing the manuscript. All authors contributed to the article and approved the submitted version.

FUNDING

HJ was supported by the Centers for Disease Control and Prevention Contract No. 34150, and a grant from the National Institute on Drug Abuse of the National Institute of Health under award no. K01DA048985. FA-E was supported by two grants from the National Cancer Institute (U01-CA-199335 and U01-CA-253913) as part of the Cancer Intervention and Surveillance Modeling Network (CISNET), the Gordon and Betty Moore Foundation, and Open Society Foundations (OSF). The funding agencies had no role in the design of the study, interpretation of results, or writing of the manuscript. The content is solely the responsibility of the authors and does not necessarily represent the official views of the Centers for Disease Control and Prevention, National Cancer Institute or National Institutes of Health.

REFERENCES

Alarid-Escudero, F., MacLehose, R. F., Peralta, Y., Kuntz, K. M., and Enns, E. A. (2018). Nonidentifiability in model calibration and implications for medical decision making. *Med. Decis. Mak.* 38, 810–821. doi: 10.1177/0272989X18792283

Arias, E. (2014). *United States Life Tables, 2014*. National Vital Statistics Reports.

Badiru, A. B., and Sieger, D. B. (1998). Neural network as a simulation metamodel in economic analysis of risky projects. *Eur. J. Operat. Res.* 105, 130–142. doi: 10.1016/S0377-2217(97)00029-5

Banks, J. (1998). *Handbook of Simulation: Principles, Methodology, Advances, Applications, and Practice*. Hoboken, NJ: John Wiley & Sons, Inc. doi: 10.1002/9780470172445

Barton, R. R. (1992). "Metamodels for simulation input-output relations," in *Winter Simulation Conference*, Vol. 9, eds J. Swain, D. Goldsman, R. Crain, and J. Wilson (Arlington, VA), 289–299. doi: 10.1145/167293.167352

Barton, R. R. (2009). "Simulation optimization using metamodels," in *Winter Simulation Conference (WSC)* (Austin, TX), 230–238. doi: 10.1109/WSC.2009.5429328

Barton, R. R., and Meckesheimer, M. (2006). "Chapter 18: Metamodel-based simulation optimization," in *Handbooks in Operations Research and Management Science*, eds S. G. Henderson and B. L. Nelson (Amsterdam: North-Holland), 535–574. doi: 10.1016/S0927-0507(06)13018-2

Blanning, R. W. (1974). The sources and uses of sensitivity information. *Interfaces* 4, 32–38. doi: 10.1287/inte.4.4.32

Briggs, A. H., Weinstein, M. C., Fenwick, E. A. L., Karnon, J., Sculpher, M. J., and Paltiel, A. D. (2012). Model parameter estimation and uncertainty analysis: a report of the ISPOR-SMDM modeling good research practices task force working group-6. *Med. Decis. Mak.* 32, 722–732. doi: 10.1177/0272989X12458348

Carpenter, B., Gelman, A., Hoffman, M. D., Lee, D., Goodrich, B., Betancourt, M., et al. (2017). Stan: a probabilistic programming language. *J. Stat. Softw.* 76, 1–32. doi: 10.18637/jss.v076.i01

Chambers, M., and Mount-Campbell, C. A. (2002). Process optimization via neural network metamodeling. *Int. J. Prod. Econ.* 79, 93–100. doi: 10.1016/S0925-5273(00)00188-2

de Carvalho, T. M., Heijnsdijk, E. A., Coffeng, L., and de Koning, H. J. (2019). Evaluating parameter uncertainty in a simulation model of cancer using emulators. *Med. Decis. Mak.* 39, 405–413. doi: 10.1177/0272989X19837631

Degeling, K., IJzerman, M. J., Lavieri, M. S., Strong, M., and Koffijberg, H. (2020). Introduction to metamodeling for reducing computational burden of advanced analyses with health economic models: a structured overview of metamodeling methods in a 6-step application process. *Med. Decis. Mak.* 40, 348–363. doi: 10.1177/0272989X20912233

El Tabach, E., Lancelot, L., Shahrour, I., and Najjar, Y. (2007). Use of artificial neural network simulation metamodelling to assess groundwater contamination in a road project. *Math. Comput. Modell.* 45, 766–776. doi: 10.1016/j.mcm.2006.07.020

Folkes, S., Lahav, O., and Maddox, S. (1996). An artificial neural network approach to the classification of galaxy spectra. *Mnthly Notices R. Astron. Soc.* 283, 651–665. doi: 10.1093/mnras/283.2.651

Fonseca, D., Navaresse, D., and Moynihan, G. (2003). Simulation metamodeling through artificial neural networks. *Eng. Appl. Artif. Intell.* 16, 177–183. doi: 10.1016/S0952-1976(03)00043-5

Friedman, L. W., and Pressman, I. (1988). The metamodel in simulation analysis: can it be trusted? *J. Operat. Res. Soc*. 39, 939–948. doi: 10.1057/jors.1988.160

Fu, M. C. (1994). "A tutorial review of techniques for simulation optimization," in *Proceedings of the 1994 Winter Simulation Conference*, eds J. D. Tew, S. Manivannan, D. A. Sadowski, and A. F. Seila (Lake Buena Vista, FL: IEEE), 8. doi: 10.1109/WSC.1994.717096

Gramacy, R. B. (2020). *Surrogates: Gaussian Process Modeling, Design, and Optimization for the Applied Sciences*. Boca Raton, FL: CRC Press, 333–376.

Hauser, T., Keats, A., and Tarasov, L. (2012). Artificial neural network assisted Bayesian calibration of climate models. *Clim. Dyn*. 39, 137–154. doi: 10.1007/s00382-011-1168-0

Heath, A., Kunst, N., Jackson, C., Strong, M., Alarid-Escudero, F., Goldhaber-Fiebert, J. D., et al. (2020). Calculating the expected value of sample information in practice: considerations from 3 case studies. *Med. Decis. Mak*. 40, 314–326. doi: 10.1177/0272989X20912402

Hurrion, R. (1997). An example of simulation optimisation using a neural network metamodel: Finding the optimum number of kanbans in a manufacturing system. *J. Operat. Res. Soc*. 48, 1105–1112. doi: 10.1038/sj.jors.2600468

Jain, A. K., Mao, J., and Mohiuddin, K. (1996). Artifical neural networks: a tutorial. *Computer* 29, 31–44. doi: 10.1109/2.485891

Jalal, H., Boudreaux, M., Dowd, B., and Kuntz, K. M. (2015). "Measuring decision sensitivity with Monte Carlo simulation and multinomial logistic regression metamodeling," in *The Society for Medical Decision Making Conference* (St. Louis, MO).

Jalal, H., Dowd, B., Sainfort, F., and Kuntz, K. M. (2013). Linear regression metamodeling as a tool to summarize and present simulation model results. *Med. Decis. Mak*. 33, 880–890. doi: 10.1177/0272989X13492014

Jalal, H., Pechlivanoglou, P., Krijkamp, E., Alarid-Escudero, F., Enns, E. A., and Hunink, M. G. M. (2017). An overview of R in health decision sciences. *Med. Decis. Mak*. 37, 735–746. doi: 10.1177/0272989X16686559

Kennedy, M. C., and O'Hagan, A. (2001). Bayesian calibration of computer models. *J. R. Stat. Soc. Ser. B Stat. Methodol*. 63, 425–464. doi: 10.1111/1467-9868.00294

Khu, S.-T., Savic, D., Liu, Y., Madsen, H., and Science, C. (2004). "A fast evolutionary-based meta-modelling approach for the calibration of a rainfall-runoff model," in *Trans. 2nd Biennial Meeting of the International Environmental Modelling and Software Society, iEMSs* (Osnabruck), 1–6.

Khuri, A. I., and Mukhopadhyay, S. (2010). Response surface methodology. *Wiley Interdiscipl. Rev. Comput. Stat*. 2, 128–149. doi: 10.1002/wics.73

Kilmer, R. A. (1994). *Artificial neural network metamodels of stochastic computer simulations* (Ph.D. thesis). Pittsburgh University, Pittsburgh, PA, United States.

Kilmer, R. A. (1996). Applications of artificial neural networks to combat simulations. *Math. Comput. Modell*. 23, 91–99. doi: 10.1016/0895-7177(95)00220-0

Kilmer, R. A., Smith, A. E., and Shuman, L. J. (1997). An emergency department simulation and a neural network metamodel. *J. Soc. Health Syst*. 5, 63–79.

Kleijnen, J. (1975). A comment on Blanning's "metamodel for sensitivity analysis: the regression metamodel in simulation". *Interfaces* 5, 21–23. doi: 10.1287/inte.5.3.21

Kleijnen, J. P., and Sargent, R. G. (2000). A methodology for fitting and validating metamodels in simulation. *Eur. J. Operat. Res*. 120, 14–29. doi: 10.1016/S0377-2217(98)00392-0

Kleijnen, J. P. C. (1979). Regression metamodels for generalizing simulation results. *IEEE Trans. Syst. Man Cybernet*. 9, 93–96. doi: 10.1109/TSMC.1979.4310155

Kleijnen, J. P. C. (2015). *Design and Analysis of Simulation Experiments, 2nd Edn*. New York, NY: Springer US. doi: 10.1007/978-3-319-18087-8

Kleijnen, J. P. C., Sanchez, S. M., Lucas, T. W., and Cioppa, T. M. (2005). State-of-the-art review: a user's guide to the brave new world of designing simulation experiments. *INFORMS J. Comput*. 17, 263–289. doi: 10.1287/ijoc.1050.0136

Knudsen, A. B., Zauber, A. G., Rutter, C. M., Naber, S. K., Doria-Rose, V. P., Pabiniak, C., et al. (2016). Estimation of benefits, burden, and harms of colorectal cancer screening strategies: modeling study for the US preventive services task force. *JAMA* 10017, 2595–2609. doi: 10.1001/jama.2016.6828

Krijkamp, E. M., Alarid-Escudero, F., Enns, E. A., Jalal, H. J., Hunink, M. M., and Pechlivanoglou, P. (2018). Microsimulation modeling for health decision sciences using R: a tutorial. *Med. Decis. Mak*. 38, 400–422. doi: 10.1177/0272989X18754513

Kuntz, K. M., Lansdorp-Vogelaar, I., Rutter, C. M., Knudsen, A. B., van Ballegooijen, M., Savarino, J. E., et al. (2011). A systematic comparison of microsimulation models of colorectal cancer: the role of assumptions about adenoma progression. *Med. Decis. Mak*. 31, 530–539. doi: 10.1177/0272989X11408730

Lunn, D., Spiegelhalter, D., Thomas, A., and Best, N. (2009). The bugs project: evolution, critique and future directions. *Stat. Med*. 28, 3049–3067. doi: 10.1002/sim.3680

Mareš, T. and Kučerová, A. (2012). "Artificial neural networks in calibration of nonlinear models," in *Life-Cycle and Sustainability of Civil Infrastructure Systems-Proceedings of the Third International Symposium on Life-Cycle Civil Engineering (IALCCE'12)* (Vienna: CRC Press Stirlingshire), 2225–2232.

Másson, E., and Wang, Y.-J. (1990). Introduction to computation and learning in artificial neural networks. *Eur. J. Operat. Res*. 47, 1–28. doi: 10.1016/0377-2217(90)90085-P

Menzies, N. A., Soeteman, D. I., Pandya, A., and Kim, J. J. (2017). Bayesian methods for calibrating health policy models: a tutorial. *PharmacoEconomics* 35, 613–624. doi: 10.1007/s40273-017-0494-4

Michie, E. D., Spiegelhalter, D. J., and Taylor, C. C. (1994). *Machine Learning, Neural and Statistical Classification*. River, NJ: Ellis Horwood, 84–106.

Ndikumana, A., Tran, N. H., Kim, K. T., and Hong, C. S. (2020). Deep learning based caching for self-driving cars in multi-access edge computing. *IEEE Trans. Intell. Transport. Syst*. doi: 10.1109/TITS.2020.2976572

Nelder, J., and Mead, R. (1965). A simplex method for function minimization. *Computer J*. 7, 308–313. doi: 10.1093/comjnl/7.4.308

Neumann, P. J., Sanders, G. D., Russell, L. B., Siegel, J. E., and Ganiats, T. G. (2016). *Cost-Effectiveness in Health and Medicine*. New York, NY: Oxford University Press. doi: 10.1093/acprof:oso/9780190492939.001.0001

O'Hagan, A., Kennedy, M. C., and Oakley, J. E. (1999). Uncertainty analysis and other inference tools for complex computer codes. *Bayesian Staist*. 6, 503–524.

Olden, J. D., Lawler, J. J., and Poff, N. L. (2008). Machine learning methods without tears: a primer for ecologists. *Q. Rev. Biol*. 83, 171–193. doi: 10.1086/587826

Paiva, R. M., D., Carvalho, A. R., Crawford, C., and Suleman, A. (2010). Comparison of surrogate models in a multidisciplinary optimization framework for wing design. *AIAA J*. 48, 995–1006. doi: 10.2514/1.45790

Pichler, B., Lackner, R., and Mang, H., a. (2003). Back analysis of model parameters in geotechnical engineering by means of soft computing. *Int. J. Num. Methods Eng*. 57, 1943–1978. doi: 10.1002/nme.740

Pierreval, H., Bernard, U. C., Novembre, B., and Cedex, V. (1992). Training a neural network by simulation for dispatching problems. *Proc. Third Int. Conf. Comput. Integr. Manufact*. 1992, 332–336. doi: 10.1109/CIM.1992.639120

Pierreval, H., and Huntsinger, R. C. (1992). "An investigation on neural network capabilities as simulation metamodels," in *Proceedings of the 1992 Summer Computer Simulation Conference* (Troy, NY), 413–417.

R Core Team (2018). *R: A Language and Environment for Statistical Computing*. Vienna: R Foundation for Statistical Computing. Available online at: https://www.R-project.org/

Raftery, A. E., and Bao, L. (2010). Estimating and projecting trends in HIV/AIDS generalized epidemics using incremental mixture importance sampling. *Biometrics* 66, 1162–1173. doi: 10.1111/j.1541-0420.2010.01399.x

Ravì, D., Wong, C., Deligianni, F., Berthelot, M., Andreu-Perez, J., Lo, B., et al. (2016). Deep learning for health informatics. *IEEE J. Biomed. Health Inform*. 21, 4–21. doi: 10.1109/JBHI.2016.2636665

Reis dos Santos, P. M., and Reis dos Santos, M. I. (2009). Using subsystem linear regression metamodels in stochastic simulation. *Eur. J. Operat. Res*. 196, 1031–1040. doi: 10.1016/j.ejor.2008.05.005

Rojas, R. (1996). "Statistics and neural networks," in *Neural Networks* (Berlin: Springer) 229–264. doi: 10.1007/978-3-642-61068-4_9

Rutter, C. M., Miglioretti, D. L., and Savarino, J. E. (2009). Bayesian calibration of microsimulation models. *J. Am. Stat. Assoc*. 104, 1338–1350. doi: 10.1198/jasa.2009.ap07466

Rutter, C. M., Ozik, J., DeYoreo, M., and Collier, N. (2019). Microsimulation model calibration using incremental mixture approximate bayesian computation. *Ann. Appl. Stat*. 13, 2189–2212. doi: 10.1214/19-AOAS1279

Rutter, C. M., Yu, O., and Miglioretti, D. L. (2007). A hierarchical non-homogenous Poisson model for meta-analysis of adenoma counts. *Stat. Med.* 26, 98–109. doi: 10.1002/sim.2460

Sabuncuoglu, I., and Touhami, S. (2002). Simulation metamodelling with neural networks: an experimental investigation. *Int. J. Product. Res.* 40, 2483–2505. doi: 10.1080/00207540210135596

Sacks, J., Welch, W. J., Mitchell, T. J., and Wynn, H. P. (1989). Design and analysis of computer experiments. *Stat. Sci.* 4, 409–423. doi: 10.1214/ss/1177012420

Santos, I. R., and Santos, P. R. (2007). "Simulation metamodels for modeling output distribution parameters," in *Winter Simulation Conference* (Washington, DC: IEEE), 910–918. doi: 10.1109/WSC.2007.4419687

Schmidhuber, J. (2015). Deep learning in neural networks: an overview. *Neural Netw.* 61, 85–117. doi: 10.1016/j.neunet.2014.09.003

Stevenson, M. D., Oakley, J., and Chilcott, J. B. (2004). Gaussian process modeling in conjunction with individual patient simulation modeling: a case study describing the calculation of cost-effectiveness ratios for the treatment of established osteoporosis. *Med. Decis. Mak.* 24, 89–100. doi: 10.1177/0272989X03261561

Vanni, T., Karnon, J., Madan, J., White, R. G., Edmunds, W. J., Foss, A. M., et al. (2011). Calibrating models in economic evaluation: a seven-step approach. *PharmacoEconomics* 29, 35–49. doi: 10.2165/11584600-000000000-00000

Wang, L. (2005). A hybrid genetic algorithm-neural network strategy for simulation optimization. *Appl. Math. Comput.* 170, 1329–1343. doi: 10.1016/j.amc.2005.01.024

Weinstein, M. C., O'Brien, B., Hornberger, J., Jackson, J., Johannesson, M., McCabe, C., et al. (2003). Principles of good practice for decision analytic modeling in health-care evaluation: report of the ISPOR Task Force on Good Research Practices-Modeling Studies. *Value Health* 6, 9–17. doi: 10.1046/j.1524-4733.2003.00234.x

Weiser Friedman, L. (1996). *The Simulation Metamodel*. Norwell, MA: Kluwer Academic Publishers. doi: 10.1007/978-1-4613-1299-4

Wu, G. H.-M., Wang, Y.-M., Yen, A. M.-F., Wong, J.-M., Lai, H.-C., Warwick, J., et al. (2006). Cost-effectiveness analysis of colorectal cancer screening with stool DNA testing in intermediate-incidence countries. *BMC Cancer* 6:136. doi: 10.1186/1471-2407-6-136

Zobel, C. W., and Keeling, K. B. (2008). Neural network-based simulation metamodels for predicting probability distributions. *Comput. Indus. Eng.* 54, 879–888. doi: 10.1016/j.cie.2007.08.012

Conflict of Interest: The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

frontiers in Physiology

ORIGINAL RESEARCH
published: 24 June 2021
doi: 10.3389/fphys.2021.667828

Deep Learning Approaches to Surrogates for Solving the Diffusion Equation for Mechanistic Real-World Simulations

J. Quetzalcóatl Toledo-Marín[1,2]*, *Geoffrey Fox*[2,3], *James P. Sluka*[1,2] and *James A. Glazier*[1,2]

[1] Biocomplexity Institute, Indiana University, Bloomington, IN, United States, [2] Luddy School of Informatics, Computing and Engineering, Bloomington, IN, United States, [3] Digital Science Center, Bloomington, IN, United States

In many mechanistic medical, biological, physical, and engineered spatiotemporal dynamic models the numerical solution of partial differential equations (PDEs), especially for diffusion, fluid flow and mechanical relaxation, can make simulations impractically slow. Biological models of tissues and organs often require the simultaneous calculation of the spatial variation of concentration of dozens of diffusing chemical species. One clinical example where rapid calculation of a diffusing field is of use is the estimation of oxygen gradients in the retina, based on imaging of the retinal vasculature, to guide surgical interventions in diabetic retinopathy. Furthermore, the ability to predict blood perfusion and oxygenation may one day guide clinical interventions in diverse settings, i.e., from stent placement in treating heart disease to BOLD fMRI interpretation in evaluating cognitive function (Xie et al., 2019; Lee et al., 2020). Since the quasi-steady-state solutions required for fast-diffusing chemical species like oxygen are particularly computationally costly, we consider the use of a neural network to provide an approximate solution to the steady-state diffusion equation. Machine learning surrogates, neural networks trained to provide approximate solutions to such complicated numerical problems, can often provide speed-ups of several orders of magnitude compared to direct calculation. Surrogates of PDEs could enable use of larger and more detailed models than are possible with direct calculation and can make including such simulations in real-time or near-real time workflows practical. Creating a surrogate requires running the direct calculation tens of thousands of times to generate training data and then training the neural network, both of which are computationally expensive. Often the practical applications of such models require thousands to millions of replica simulations, for example for parameter identification and uncertainty quantification, each of which gains speed from surrogate use and rapidly recovers the up-front costs of surrogate generation. We use a Convolutional Neural Network to approximate the stationary solution to the diffusion equation in the case of two equal-diameter, circular, constant-value sources located at random positions in a two-dimensional square domain with absorbing boundary conditions. Such a configuration caricatures the chemical concentration field of a fast-diffusing species like oxygen in a tissue with two parallel blood vessels in a cross section perpendicular to the

OPEN ACCESS

Edited by:
Gary An,
University of Vermont, United States

Reviewed by:
Nikolaos Tsoukias,
Florida International University, United States
Hermann Frieboes,
University of Louisville, United States

***Correspondence:**
J. Quetzalcóatl Toledo-Marín
j.toledo.mx@gmail.com

Specialty section:
This article was submitted to Computational Physiology and Medicine, a section of the journal Frontiers in Physiology

Received: *14 February 2021*
Accepted: *25 May 2021*
Published: *24 June 2021*

Citation:
Toledo-Marín JQ, Fox G, Sluka JP and Glazier JA (2021) Deep Learning Approaches to Surrogates for Solving the Diffusion Equation for Mechanistic Real-World Simulations. Front. Physiol. 12:667828. doi: 10.3389/fphys.2021.667828

two blood vessels. To improve convergence during training, we apply a training approach that uses roll-back to reject stochastic changes to the network that increase the loss function. The trained neural network approximation is about 1000 times faster than the direct calculation for individual replicas. Because different applications will have different criteria for acceptable approximation accuracy, we discuss a variety of loss functions and accuracy estimators that can help select the best network for a particular application. We briefly discuss some of the issues we encountered with overfitting, mismapping of the field values and the geometrical conditions that lead to large absolute and relative errors in the approximate solution.

Keywords: diffusion surrogate, machine learning, virtual tissue, mechanistic modeling, Julia

1. INTRODUCTION

Diffusion is ubiquitous in physical, biological, and engineered systems. In mechanistic computer simulations of the dynamics of such systems, solving the steady state and time-varying diffusion equations with multiple sources and sinks is often the most computationally expensive part of the calculation, especially in cases with multiple diffusing species with diffusion constants differing by multiple orders of magnitude. Examples in biology include cells secreting and responding to diffusible chemical signals during embryonic development, blood vessels secreting oxygen which cells in tissues absorb during normal tissue function, tumors secreting growth factors promoting neoangiogenesis in cancer progression, or viruses spreading from their host cells to infect other cells in tissues. In these situations the natural diffusion constants can range from $\sim 10^3 \mu m^2/s$ for oxygen to $\sim 0.1 - 10^2 \mu m^2/s$ for a typical protein (Phillips, 2018). Dynamic simulations of biological tissues and organs may require the independent calculation of the time-varying concentrations of dozens of chemical species in three dimensions, and in the presence of a complex field of cells and extracellular matrix. As the number of species increases, solving these diffusion equations dominates the computational cost of the simulation. Numerous approaches attempt to reduce the cost of solving the diffusion equation including implicit, particle-based, frequency-domain and finite-element methods, multithreaded, and MPI-based parallelization and GPUs, but all have significant limitations. HPC methods that do not require Deep Learning (DL) can certainly accelerate solution of problems including diffusion equations, e.g., Secomb's Green's function method leverages GPUs to accelerate solution of 3D advection-diffusion in microvessels with time-dependent sinks and sources (Secomb, 2016). Such methods could greatly reduce the time required to generate training sets for DL-assisted approaches. Machine learning has also been applied to solve a growing list of PDE problems (Farimani et al., 2017; Sharma et al., 2018; Edalatifar et al., 2020; He and Pathak, 2020; Li A. et al., 2020; Li Z. et al., 2020; Cai et al., 2021). See Fox and Jha (2019) for a thorough review. Machine learning has also been applied to the *inverse problem*, i.e., attempting to infer the underlying mechanistic equations governing a complex system from experimental data. These methods can potentially lead to the discovery of new physics (Champion et al., 2019). Similarly, Neural-ODEs is a highly active and exciting field where neural networks are embedded into a differential equation. Modeling a process via an ODE typically consists in equating the change rate of a *quantity* (e.g., concentration) to an operator applied to that same *quantity* plus some other *quantities* which are called inhomogeneities or *external fields*, then solving the ODE and comparing it with experimental data of that *thing* to validate the model or fit parameters. The operator in the ODE is selected *a priori* based on the *symmetries* of the process. Neural-ODEs replaces the operator with a neural network. The neural network is trained by solving the Neural-ODE and comparing it with the experimental data (Chen et al., 2018; Rackauckas et al., 2019). Moreover, *Physics Informed Neural Networks* tackle forward and inverse problems by embedding physical information into the neural network. Embedding physical information into the neural network means embedding the ODE, the initial conditions and the boundary conditions into the loss function used to train the neural network (Raissi et al., 2019). In the case of multiscale modeling, the complexity of the system includes different characteristic length and time scales differing by orders of magnitude. Multiscale modeling using standard computational approaches, such as Monte Carlo methods and/or molecular dynamics is time consuming. AI-based surrogates using deep learning methods can accelerate computation by replacing specific classical solvers, while preserving the overall interpretability of mechanistic models. In real-world problems, the number of sources and sinks, their shape, boundary fluxes, and positions differ from instance to instance and may change in time. Boundary conditions may also be complicated and diffusion constants may be anisotropic or vary in space. The resulting lack of symmetry means that many high-speed implicit and frequency-domain diffusion-solver approaches do not work effectively, requiring the use of simpler but slower forward solvers (Schiesser, 2012). Deep learning[1] surrogates to solve either the steady-state field or the time-dependent field for a given set of sources and sinks subject to diffusion could potentially increase the speed of such simulations by several orders of magnitude compared to the use of direct numerical solvers.

[1]We use the terms *deep learning* and *machine learning* interchangeably. We also use *neural network* and *deep neural network* interchangeably.

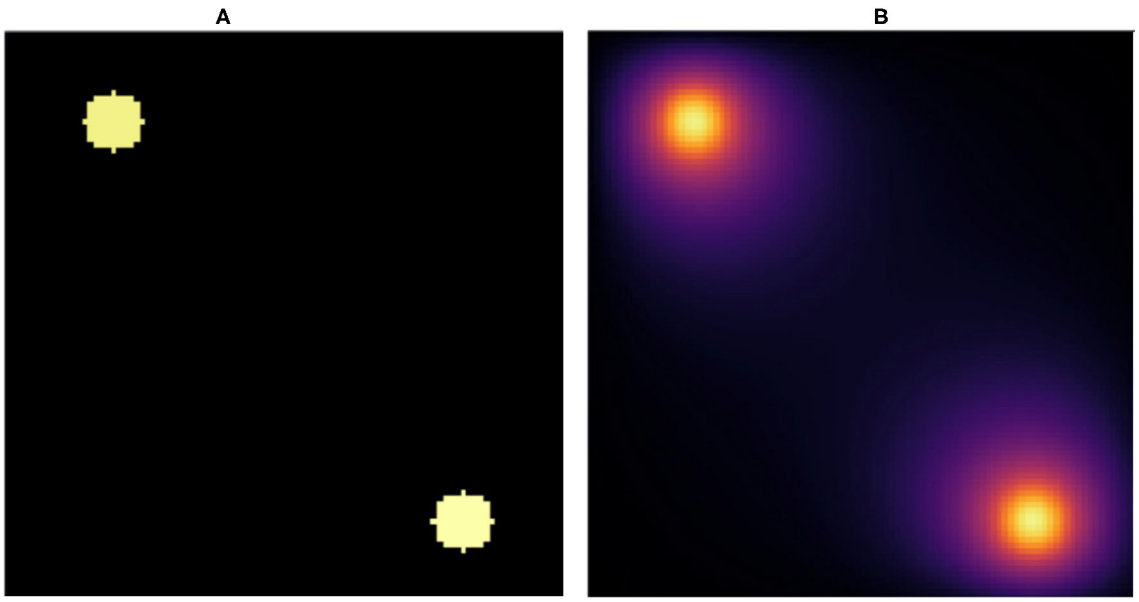


FIGURE 1 | Snapshot of **(A)** initial condition and **(B)** stationary state solution. **(A)** We placed two random value sources of radius 5 *voxels* in random positions fully within a 100 × 100*pixel* lattice and used this configuration as the input to the NN. **(B)** Stationary solution to the diffusion equation with absorbing boundary conditions for the initial conditions in **(A)**. The stationary solution **(B)** is the target for the NN. We fixed the diffusion constant to $D = 1$ *voxels*$^2/s$ and the decay rate to $\gamma = 1/400s^{-1}$, which yields a diffusion length equal to $\sqrt{D/\gamma}$ *voxels* $= 20$*voxels*.

One challenge in developing effective deep neural network (NN) diffusion-solver surrogates is that the dimensionality of the problem specification is potentially very high, with an arbitrary pattern of sources and sinks, with different boundary conditions for each source and sink, and spatially variable or anisotropic diffusivities. As a proof-of-principle we will start with a NN surrogate for a simple version of the problem that we can gradually generalize to a full surrogate in future work. In a two-dimensional square domain represented as $N \times N$*pixels* and with absorbing boundary conditions, we place two circular sources of equal diameters at random positions, with the constraint that the sources do not overlap and are fully contained within the domain. Each source imposes a constant value on the diffusing field within the source and at its boundary. We select the value for one of the sources equal to 1 while the value for the other source is randomly selected from a uniform distribution between $(0, 1]$ (see **Figure 1A**). Outside the sources the field diffuses with a constant diffusion constant (D) and linearly decays with a constant decay rate (γ). This simple geometry could represent the diffusion and uptake of oxygen in a volume of tissue between two parallel blood vessels of different diameters. Although reflecting or periodic boundary conditions might better represent a potion of a larger tissue, we use the simpler absorbing boundary conditions here. In this case, the steady-state field depends critically on the distance between the sources, and between the sources and the boundary, both relative to the diffusion length ($l_D = (D/\gamma)^{1/2}$) and on the sources' field strengths.

In practice then, the solution of the steady state diffusion equation maps an image consisting of $N \times N$ pixels with 0 value outside the sources and constant values between 0 and 1 inside the sources to a second image of the same size, which has the same values inside the sources but values between 0 and 1 elsewhere (see **Figure 1B**). We evaluate the ability of a NN trained on the explicit numerical solutions of the steady-state diffusion field for 20,000 two-source examples to approximate the steady state field for configurations of sources that it had not previously encountered.

Notice that the diffusion kernel convolution used in the direct solution of the time-dependent diffusion equation (e.g., finite-element methods) is a type of convolutional neural network (Schiesser, 2012). Therefore we chose deep convolutional NN as the architecture. However, there are multiple types of convolutional NN. Here we considered two of these. A deep convolutional neural network and an autoencoder (Baur et al., 2020). In addition, because it was possible that these two types would do better at replicating specific aspects of the overall solution, we also evaluated a superposition of the two. Time series surrogates often use recurrent NN (Zhang and Xiao, 2000; Dubois et al., 2020). Similarly, deep generative models have been shown to be useful to sample from high dimensional space, as in the case of molecular dynamics and chemical reaction modeling (Chen and Ferguson, 2018; Noé et al., 2019, 2020; Zhang et al., 2019; Gkeka et al., 2020; Kasim et al., 2020). Since our main interest is the stationary solution, we did not consider these approaches.

2. MODEL

Figure 2 shows the data flow through the NN. We denote by $|x\rangle$ and $|\hat{y}\rangle$ the input and output images, that is the initial condition

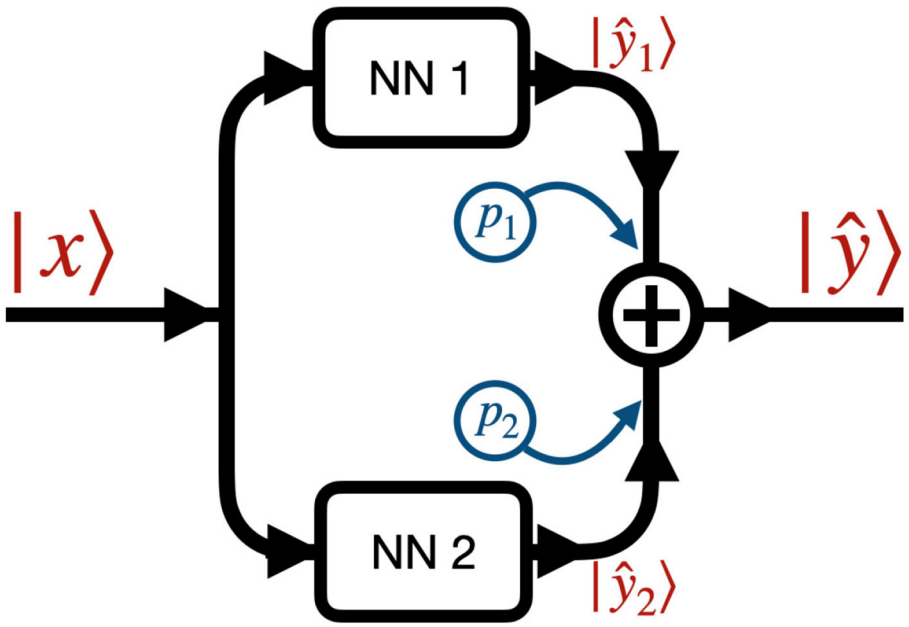


FIGURE 2 | Network architecture: the input image $|x\rangle$ passes through NN 1 (see **Figure 3A**) and NN 2 (see **Figure 3B**), generating the two outputs $\hat{y}_1\rangle$ and $|\hat{y}_2\rangle$. The final output $|\hat{y}\rangle$ is the sum of the outputs of the two NNs weighted by coefficients p_1 and p_2, i.e., $|\hat{y}\rangle = p_1|\hat{y}_1\rangle + p_2|\hat{y}_2\rangle$. p_i are fixed Boolean hyperparameters for the model and fixed for each model we trained. This means that when a given model has $p_i = 0$ ($p_i = 1$) then *NNi* is turned off (on).

layout of the source cells and the predicted stationary solution of the diffusion equation, respectively. The input $|x\rangle$ passes to two different neural networks (NNs) denoted *NN 1* (**Figure 3A**) and *NN 2* (**Figure 3B**) which output $|\hat{y}_1\rangle$ and $|\hat{y}_2\rangle$, respectively. The output $|\hat{y}\rangle$ is a weighted sum of the outputs of the two NNs, $|\hat{y}\rangle = p_1|\hat{y}_1\rangle + p_2|\hat{y}_2\rangle$, where p_1 and p_2 are fixed hyperparameters, i.e., these hyperparameters are fixed during training. In our code (Toledo-Marin, 2020) p_i are real numbers, however, in this paper we only consider the Boolean case where they each take values of 0 or 1. *NN 1* is a deep convolutional neural network that maintains the height and width of the input image through each of 6 convolutional layers. The first layer outputs a 4-channel image, the second layer outputs an 8-channel image, the third layer outputs a 16-channel image, the fourth layer outputs an 8-channel image, the fifth layer outputs a 4-channel image and the sixth layer outputs a 1-channel image. *NN 2* is an autoencoder (Chen et al., 2017) where the first six layers perform a meanpool operation that reduces height and width in half after each layer following the sequence $\{100^2, 50^2, 25^2, 12^2, 6^2, 3^2, 1^2\}$ while adding channels after each layer following the sequence $\{1, 64, 128, 256, 512, 1, 024, 2, 048\}$. Then, the following six layers consist on reducing the number of channels following the sequence $\{1, 024, 512, 256, 128, 64, 1\}$ while increasing the height and width following the sequence $\{1^2, 3^2, 7^2, 13^2, 25^2, 51^2, 100^2\}$. **Figure 3** sketches the architectures of the two NNs, while **Table 1** provides their parameters. We will find that NN 1 will capture the sources whereas NN 2 will capture the field. In **Table 1**, we specify each neural network by specifying for each layer the kind of layer, the activation function and the output shape.

To generate representative two-source initial conditions and paired steady-state diffusion fields, we considered a two-dimensional lattice of size $100 \times 100 units^2$. We generated 20 k configurations with two sources, each with a radius of 5 *units*. One source has a constant source value equal to 1, while the other source has a constant source value between 0 and 1 randomly assigned using a uniform distribution. Everywhere else the field value is 0. We placed the sources in randomly uniform positions in the lattice. This image served as the input for the NN $|x\rangle$. Then we calculated the stationary solution to the diffusion equation with absorbing boundary conditions for each initial condition using the *Differential Equation* package in Julia (Rackauckas and Nie, 2017). The Julia-calculated stationary solution is the target or ground truth image for the NN $|y\rangle$. In **Figures 1A,B**, we show an initial condition and the stationary solution, respectively. We have set the diffusion constant to $D = 1 units^2/s$ and the decay rate $\gamma = 1/400 s^{-1}$, which yield a diffusion length $l_D = \sqrt{D/\gamma} = 20$ *units*. Notice that this length is 4 times the radius of the sources and 1/5 the lattice linear dimension. As γ increases and as D decreases, this length decreases. As this length decreases, the field gradient also decreases (Tikhonov and Samarskii, 2013). The source code to generate the data and train the NN can be found in Toledo-Marin (2020).

We trained the CNN setting the number of epochs to 800 using the deep learning library in Julia called Flux (Innes, 2018). We varied the dropout values between 0.0 and 0.6 in steps of 0.1 (see **Table 2**). We used ADAM as the optimizer (Kingma and Ba, 2014). Deciding on a loss function is a critical choice in the creation of the surrogate. The loss function determines the

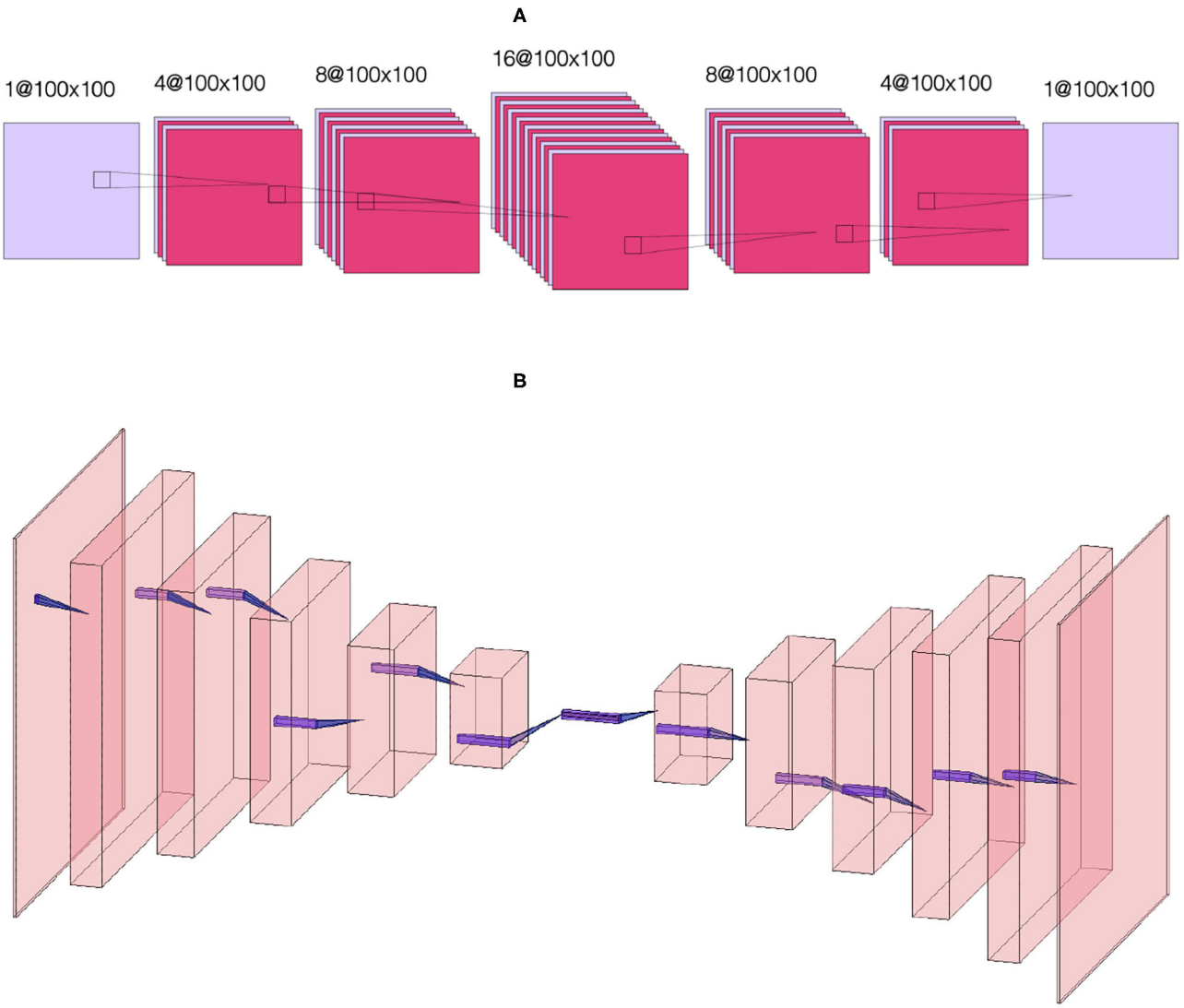


FIGURE 3 | Sketch of **(A)** convolutional NN 1. The first layer takes as input a single-channel $N \times N$ image and applies four 3×3 convolutions to generate four $N \times N$ images, the second layer applies eight 3×3 convolutions to generate eight $N \times N$ images, the third layer applies 16 3×3 convolutions to generate sixteen $N \times N$ images, the fourth layer applies eight 3×3 convolutions to generate eight $N \times N$ images, the fifth layer applies four 3×3 convolutions to generate four $N \times N$ images and the sixth layer applies a 3×3 convolution to generate a single $N \times N$ image. Sketch of **(B)** autoencoder NN 2. The first six layers perform a meanpool operation that reduces image height and width by half after each layer, with the image dimensions following the sequence $\{100^2, 50^2, 25^2, 12^2, 6^2, 3^2, 1^2\}$ while adding channels after each layer following the sequence $\{1, 64, 128, 256, 512, 1024, 2048\}$. Then, the following six layers reverse the process, reducing the number of channels following the sequence $\{1024, 512, 256, 128, 64, 1\}$ while increasing the height and width following the sequence $\{1^2, 3^2, 7^2, 13^2, 25^2, 51^2, 100^2\}$. This sketch only defines the kinds of layers used. For details about the activation functions used in each layer (see **Table 1**).

types of error the surrogate's approximation will make compared to the direct calculation and the acceptability of these errors will depend on the specific application. The mean squared error (*MSE*) error is a standard choice. However, it is more sensitive to larger absolute errors and therefore tolerates large relative errors at pixels with small values. A loss function calculated on the log of the values would be equally sensitive to relative error no matter what the absolute value. In most biological contexts we want to have a small absolute error for small values and a small relative error for large values. We explored the use of both functions, *MAE* and *MSE*, as described in **Table 2**. We used 80 and 20% of the dataset for training and test sets, respectively. We trained each model once. The highest and lowest values in the input and output images are 1 and 0, respectively. The former only occurs in sources and their vicinity. Given the configurations of the sources, the fraction of pixels in the image with values near 1 is $\sim 2\pi R^2/L^2 \approx 2\%$. Thus, pixels with small values are much more common than pixels with large values, and because the loss function is an average over the field, high field values tend to get washed out. To account for this unbalance

TABLE 1 | Convolutional neural network architectures.

Operation	Act	Output shape
Conv 3 × 3	LReLU	4 × 100 × 100
Dropout 1 (D_1)	–	–
BatchNorm	Identity	–
Conv 3 × 3	LReLU	8 × 100 × 100
BatchNorm	Identity	–
Conv 3 × 3	LReLU	16 × 100 × 100
BatchNorm	Identity	–
Conv 3 × 3	LReLU	8 × 100 × 100
BatchNorm	Identity	–
Conv 3 × 3	LReLU	4 × 100 × 100
BatchNorm	Identity	–
Conv 3 × 3	ReLU	1 × 100 × 100
Dropout 2 (D_2)	–	–
BatchNorm	Identity	–
Conv 3 × 3	LReLU	64 × 100 × 100
BatchNorm	Identity	–
Dropout 3 (D_3)	–	–
Meanpool	Identity	64 × 50 × 50
Conv 3 × 3	LReLU	128 × 50 × 50
Meanpool	Identity	128 × 25 × 25
Conv 3 × 3	LReLU	256 × 25 × 25
Meanpool	Identity	256 × 12 × 12
Conv 3 × 3	LReLU	512 × 12 × 12
Meanpool	Identity	512 × 6 × 6
Conv 3 × 3	LReLU	1,024 × 6 × 6
Meanpool	Identity	1,024 × 3 × 3
Conv 3 × 3	LReLU	2,048 × 1 × 1
ConvT 3 × 3	LReLU	1,024 × 3 × 3
ConvT 3 × 3	LReLU	512 × 7 × 7
ConvT 3 × 3	LReLU	256 × 13 × 13
ConvT 3 × 3	LReLU	128 × 25 × 25
ConvT 3 × 3	LReLU	64 × 51 × 51
Dropout 4 (D_4)	–	–
ConvT 4 × 4	ReLU	1 × 100 × 100
BatchNorm	Identity	–

Left panel corresponds to the successive operations of NN 1 while the right panel corresponds to the successive operations NN 2. Act stands for activation function. Conv, ConvT, and (L)ReLU stand for convolution, convolution transpose, and (leaky) rectified linear unit, while Identity means the activation function is the identity function (see Innes et al., 2018). Both NNs take as input the initial condition which has dimensions Channels × Width × Height = 1 × 100 × 100.

between the frequency of occurrence of low and high values, we introduced an exponential weight on the pixels in the loss function. We modulate this exponential weight through a scalar hyperparameter w, for the field in the ith lattice position in the loss function as

$$\mathcal{L}_{i\beta}^{(\alpha)} = \exp(-(\langle i|\mathbf{1}\rangle - \langle i|y_\beta\rangle)/w) \cdot \left(\langle i|\hat{y}_\beta\rangle - \langle i|y_\beta\rangle\right)^\alpha \ , \qquad (1)$$

where α is 1 or 2 for MAE or MSE, respectively and β tags the tuple in the data set (input and target). Here $\langle|\rangle$ denotes the inner product and $|i\rangle$ is a unitary vector with the same size as $|y_\beta\rangle$ with all components equal to zero except the element in position i which is equal to one. $|\mathbf{1}\rangle$ is a vector with all components equal to 1 and with size equal to that of $|y_\beta\rangle$. Then $\langle i|y_\beta\rangle$ is a scalar corresponding to the pixel value at the ith position in $|y_\beta\rangle$, whereas $\langle i|\mathbf{1}\rangle = 1$ for all i. Notice that high pixel values will then have an exponential weight ≈ 1 while low pixel values will have an exponential weight $\approx \exp(-1/w)$. This implies that the error associated to high pixels will have a larger value than that for low pixels. The loss function $\mathcal{L}^{(\alpha)}$ is the mean value over all pixels (i) and a given data set (β):

$$\mathcal{L}^{(\alpha)} = \langle \mathcal{L}_{i\beta}^{(\alpha)} \rangle \ , \qquad (2)$$

where $\langle\rangle$ denotes average. In our initial trial training runs, we noticed that the loss function always reached a plateau by 800 epochs, so we trained the NNs over 800 epochs for all runs reported in this paper. Because the training is stochastic, the loss function can increase as well as decrease between epochs as seen in **Figure 4**. At the end of 800 epochs we adopted the network configuration with the lowest loss function regardless of the epoch at which it was achieved.

While the trendline (averaged over 5 or 10 epochs) of the loss function value tends to decrease during training, the stochasticity of the training means that the value of the loss function often increases significantly between successive epochs, even by one or two orders of magnitude (see **Figure 4**). In some cases, the loss function decreases back to its trend after one or two epochs, in other cases (which we call jumps), it stays at the higher value, resetting the trend line to the higher value and only gradually begins to decrease afterwards. In this case all of the epochs after the jump have larger loss functions than the epoch immediately before the jump, as shown for the evolution of the loss function for a typical training run in **Figure 4A**. This behavior indicates that the stochastic optimization algorithm has pursued an unfavorable branch. To avoid this problem, we added a *roll-back* algorithm to the training, as proposed in Geoffrey (2020). We set a loss threshold value, $\mathcal{L}_{thrs}$, such that if the ratio of loss value from epoch n to $n+1$ is larger than $\mathcal{L}_{thrs}$, then the training algorithm reverts (rolls back) to the NN state corresponding to epoch $n-s$ and tries again. The stochasticity of training means that roll-back has an effect similar to training an ensemble of models with the same hyperparameters and selecting the model with the lowest loss function value, however, the roll-back optimization takes much less computer time than a large ensemble. We set $s = 5$ and set the threshold value $\mathcal{L}_{thrs}$ to

$$\mathcal{L}_{thrs} = C\frac{1}{m}\sum_{ep=n-m+1}^{n} \mathcal{L}^{(\alpha)}(ep) \ . \qquad (3)$$

Here we chose $C = 5$ and $m = 20$ where ep stands for epoch, i.e., we set the threshold value to 5 times the average loss function value over the previous $m = 20$ epochs. We chose these values empirically. In **Figure 4B**, we have plotted a typical example of the evolution of the loss function during training when we train using roll-back. A typical

TABLE 2 | Trained models with their corresponding hyperparameters.

Model	Weight (w)	p_1	p_2	D_1 D_2 D_3 D_4	Loss	⟨res⟩ (10^{-3})	99-P res (10^{-2})	Max res
1	1000	1	1	0.3 0.3 0.3 0.3	MSE	2.77	2.26	0.35
2	1	1	1	0.3 0.3 0.3 0.3	MSE	2.91	2.25	0.37
3	1	1	1	0.4 0.4 0.1 0.1	MSE	3.49	2.03	0.34
4	1	0	1	– – 0.3 0.3	MSE	2.49	1.97	0.38
5	1	0	1	– – 0.1 0.1	MSE	2.04	1.89	0.35
6	1	1	0	0.3 0.3 – –	MSE	75.8	16.5	0.47
7	1	1	0	0.4 0.4 – –	MSE	79.9	21.6	0.65
8	100	1	1	0.3 0.3 0.3 0.3	MAE	2.62	2.59	0.33
9	100	1	1	0.4 0.4 0.1 0.1	MAE	2.08	2.02	0.30
10	1	1	1	0.3 0.3 0.3 0.3	MAE	3.19	3.53	0.40
11	1	1	1	0.4 0.4 0.1 0.1	MAE	2.36	2.66	0.25
12	1	0	1	– – 0.1 0.1	MAE	2.12	2.17	0.34
13	10	0	1	– – 0.3 0.3	MAE	3.15	3.39	0.36
14	10	0	1	– – 0.1 0.1	MAE	2.30	2.46	0.33

*Each model is numbered for reference. The weight w is defined in Equation (1). The D_i for $i = 1, \ldots, 4$ are the dropout values (see **Table 1**). D_1 and D_2 apply to NN 1 whereas D_3 and D_4 apply to NN 2. p_1 and p_2 are Boolean variables. $p_i = 0$ ($p_i = 1$) implies NN i is turned off (on). If $p_1 = 0$ then the values of D_1 and D_2 are irrelevant, while $p_2 = 0$ makes the values of D_3 and D_4 irrelevant. The loss column specifies the loss function, either MSE for mean squared error ($\alpha = 2$) or mean absolute error MAE ($\alpha = 1$), respectively (see Equation 1). The mean res, 99-P res and max res columns show the mean, 99-percentile and maximum residual for each model computed over the test set.*

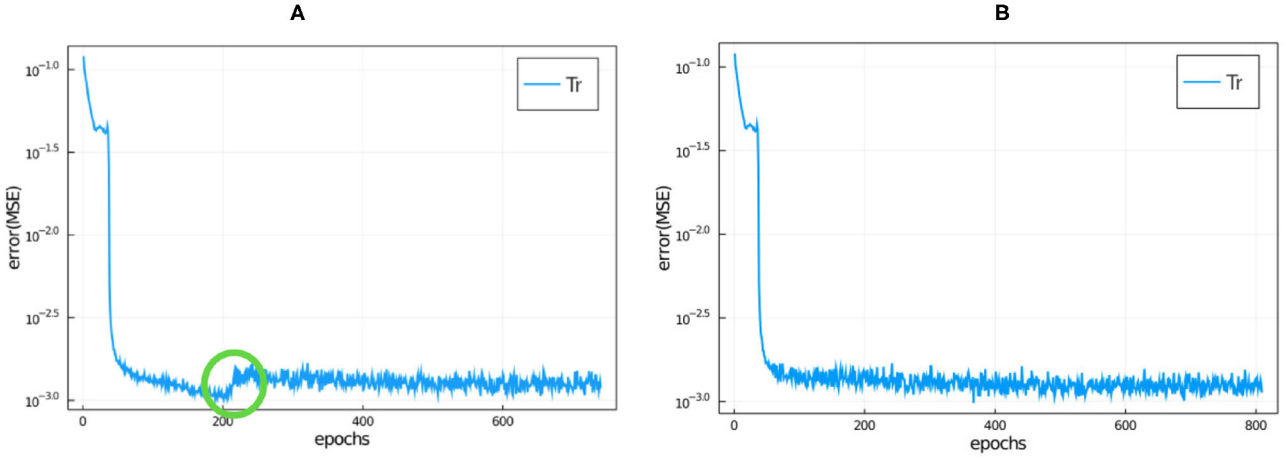


FIGURE 4 | Training loss function vs. epochs for model 9 (the hyperparameters are specified in **Table 2** and the NN details are described in the main text) without roll-back **(A)** and with roll-back **(B)** using the same seed. We have circled in green where a jump occurred during this training run (see main text for discussion).

number of roll-backs is 40, i.e., this number is the number of epochs where the jump was higher than the threshold during the training.

3. RESULTS

Quite commonly, the mean residual is the estimator used to judge the goodness of a given model. However, there are cases where the worst predictions are highly informative and can be used to make basic decisions about which features of the NN do not add value. In **Figures 5A–C** we show 20 different inputs, targets and predictions, respectively. The predictions in **Figure 5C** were obtained using model 12 (see **Table 2**) and qualitatively show very good results. For each model we computed the residual, i.e., the absolute value of the difference between the ground truth and the NN prediction pixel-by-pixel, as shown in **Figure 6B**. We also analyzed the relative residual, *i.e.*, the residual divided by the ground truth pixel-by-pixel, as shown in **Figure 6C**. Models 6 and 7, which only use NN 1 ($p_1 = 1$ and $p_2 = 0$), yield mean residuals an order of magnitude larger than models that use both or only NN 2. Therefore, we reject the NN 1-only models and do not analyze them further.

Table 2 summarizes the hyperparameter values for each model we trained. The choice of these parameters was empirically driven. Since we had the field values bounded between 0 and 1

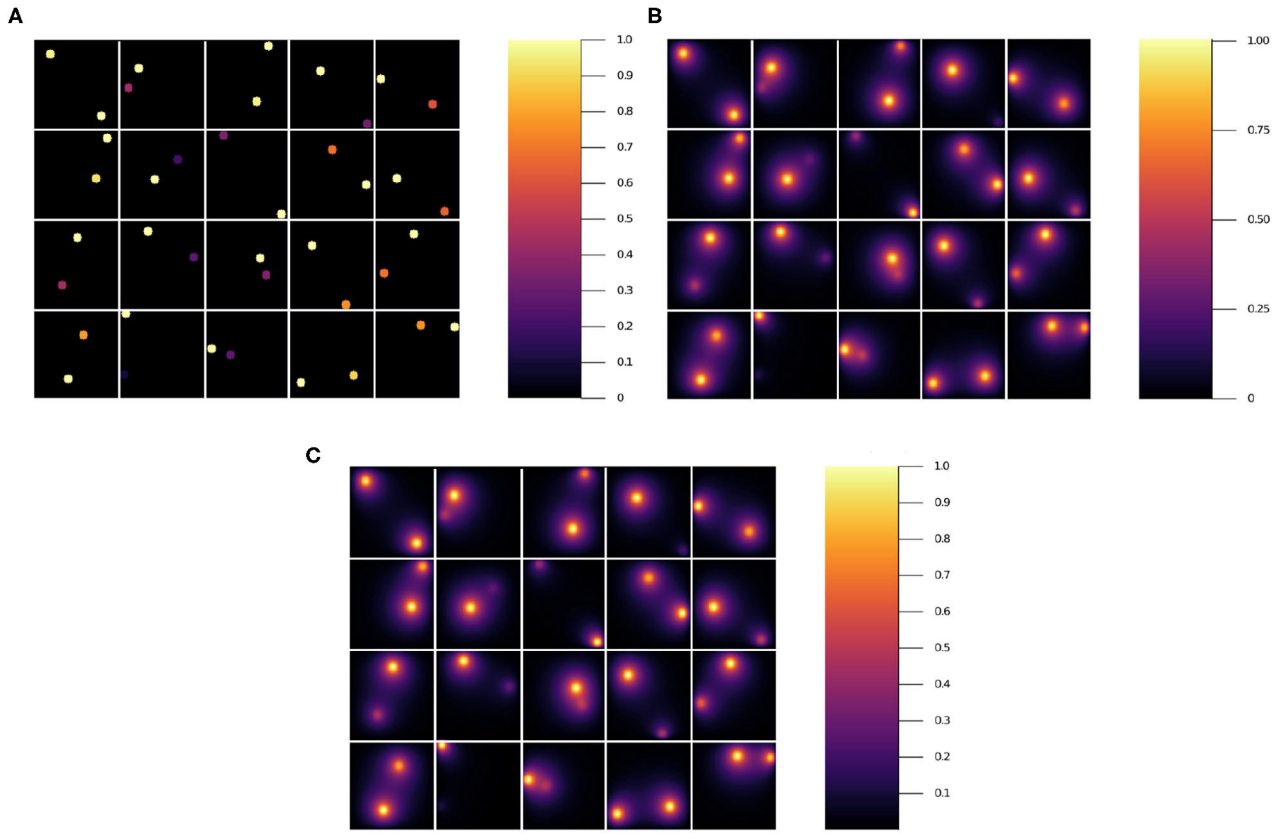


FIGURE 5 | Results for 20 randomly selected test data sets'. **(A)** input, **(B)** ground truth (target output), and **(C)** NN surrogate prediction of steady-state diffusion field output for the input.

similar to black and white images, we tested different L-norms, namely, mean absolute value (MAE), mean squared value (MSE), and mean to the fourth power, often used in neural networks applied to images. In this paper we show the results for MAE and MSE. We also tested different hyperparameters values for the dropout. We found that low dropout values for NN 2 yield the best results.

In **Figure 6D**, we have plotted the mean residual value, the 99-Percentile residual value and the maximum residual value computed over the test set. Notice that the 99-Percentile residual value is ten times the mean residual value and the maximum residual value is 10 times the 99-Percentile residual value. This suggests that the residual distribution contains outliers, i.e., there is a 1% residual that deviate from mean residual 10 to 100 times. Furthermore, these outliers correspond to regions between the source and the border, near the source, where the source is close to the border as suggested by **Figure 6B**. While the largest values in absolute residual come from pixels near the source as shown in **Figure 6B**, the relative error near the source is small whereas the relative error near boundaries is large, as shown in **Figure 6C**. In **Figure 6A** we show the stationary solution for the same batch shown in **Figures 6B,D**. Since we are considering absorbing boundary conditions, the field at the boundary is always equal to zero, thus strictly speaking the relative residual value has a singularity at the boundary. Thus, at the boundaries there is a larger relative error due to the boundary conditions. Since our method has a small absolute error independent of the mean value, the relative error is a poor measure of accuracy for small mean values, since it diverges as the mean approaches zero. Since we have zero-value boundary conditions, at the boundaries there is a larger relative error due to the boundary conditions and therefore the relative error is not a functionally meaningful measure of error unless the system being modeled is highly sensitive to small values of the field. Oxygen levels in normal tissues fluctuate significantly in space and time. For instance, in the retina, oxygen concentration fluctuates dramatically in space, time and depending on illumination. Short-term temporal fluctuations range from 5 to 50% depending on depth in cat retina (Linsenmeier and Zhang, 2017). This intrinsic oxygenation fluctuation in tissues suggests that biologically, 5% relative error at low concentrations is an acceptable accuracy for oxygen concentration estimation.

Models 5, 11, and 12 have low mean residuals with model 5 being the smallest. Focusing instead on the mean residual and the 99-Percentile, we notice that models 3, 4, 5, 11, and 12 yield the best results. Finally, considering the maximum residual together with the previous estimators, we notice that model 9 has low mean residual, low 99-percentile residual and the lowest

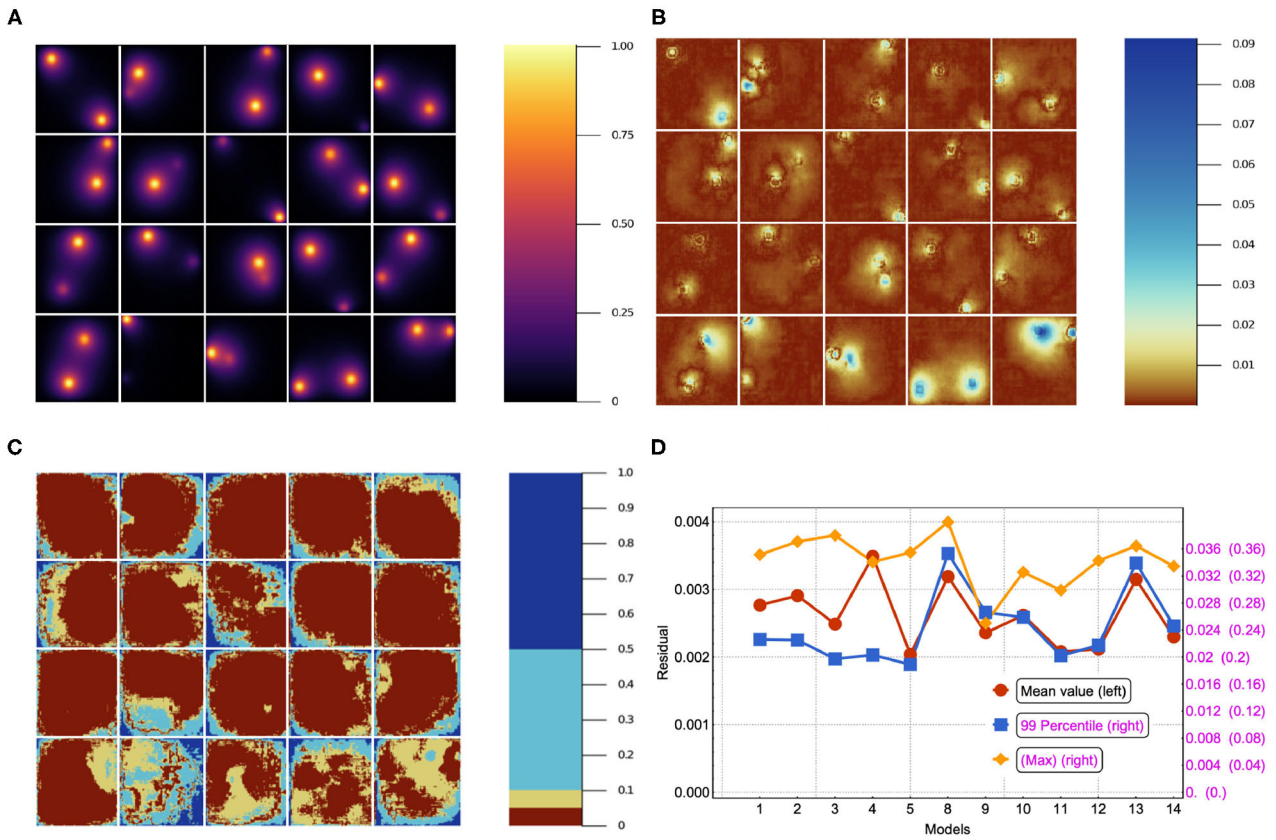


FIGURE 6 | (A) The stationary solution for the same batch in the test set. **(B)** Residual (absolute error, i.e., $||y_\beta\rangle - |\hat{y}_\beta\rangle|$) for 20 sample source images in the test set trained using model 12 in **Table 2**. **(C)** Residual/true value (relative error) for the corresponding images. **(D)** Mean, 99-Percentile, and maximum residual for all of the models in **Table 2**. Left scale for mean value, right scale for 99-Percentile residual value and right scale in parentheses for max residual value.

max residual. Depending on the user's needs, one estimator will be more relevant than others. In this sense, defining a *best* model is relative. Nevertheless, having more metrics (e.g., relative error for large values and absolute error for small values) helps to characterize each model's performance. In future work we'll consider more adaptable metrics, as well as mixed error functions that incorporate multiple estimators.

Figure 8 plots the prediction vs. the target for each pixel in each image in the training and test sets for models 9 and 11. Notice that for the test sets the results are qualitatively similar between models, for the training set the dispersion is larger in model 11 than in model 9. This suggests model 11 is overfitting the training data. Models 9 and 11 have the same hyperparameters except for the weight w. In the former $w = 100$ while in the latter $w = 1$. This suggests that the exponential weight helps reduce overfitting.

In **Figure 7**, we show the prediction from NN 1 (**Figure 7A**) and NN 2 (**Figure 7B**). Notice that NN 1 is able to detect the sources whereas NN 2 is able to predict the field. Using both neural networks improves the results as can be seen in **Figure 6D**. As previously mentioned, pixels with low (near 0) field values are much more common than pixels with high (near 1) field values. While the exponential factor in the loss function compensates for this bias, the residual in **Figure 6D** does not. To address this issue we compute the mean residual over small field intervals. This will tell us how well the model predicts for each range of absolute values. Furthermore, this method can be used to emphasize accuracy or relative accuracy in different value ranges. The way we do this is as follows. In **Figure 8**, we take 10 slices of size 0.1 in the direction $y = x$. We then compute the mean residual and standard deviation per slice. In **Supplementary Material** (section 1), we have plotted the PDF (probability density function) per slice (blue bins) and a Gaussian distribution (red curve) with mean and standard deviation set to the mean residual and standard deviation per slice, respectively. We did this for all models in **Table 2**. In **Figure 9**, we plotted the mean residual vs. for each model for each slice for the test and training sets. The error envelop shows the residual standard deviation per slice. Notice that models trained with MSE have a smaller residual standard deviation than models trained with MAE in the case of the training set, which suggest that MSE contributes to overfitting more that MAE. Recall that the difference between the MSE gradient and the MAE gradient is that the former is linear with the residual value whereas the latter is a constant. Therefore, training with MAE generalizes better

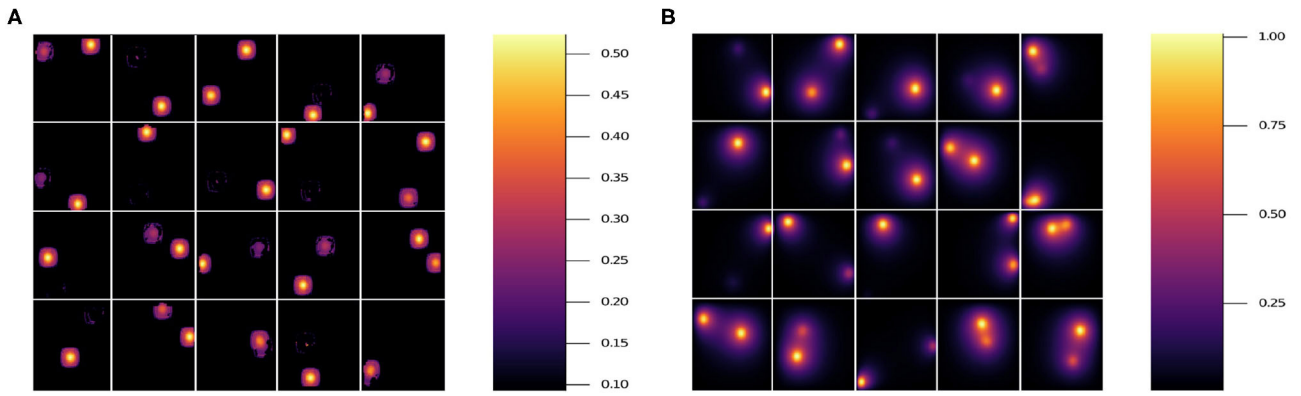


FIGURE 7 | Results for 20 randomly selected test data sets'. **(A)** Prediction using model 7, which only uses NN 1. **(B)** Prediction using model 5, which only uses NN 2 (see **Table 2**). Note the different scale on the color bars.

than MSE. Additionally, notice the dispersion increases with the slice number.

In **Figure 10**, we plotted the average and maximum over the residual mean value per slice (see **Figure 10A**) and the residual standard deviation per slice (see **Figure 10B**) for each model's test and training sets. Notice that in this approach, by slicing the residual values and computing the average residual over the set of slices, we are giving equal weight to each mean residual per slice and, therefore, compensating for the imbalance in frequency of low and high value pixels. An interesting feature from using MSE or MAE comes from the PDF of the field values. Training using MAE makes the PDF prediction quite accurate as the prediction completely overlaps with the ground truth (see **Figure 11**). In comparison, when training with MSE, the PDF is not as good and the overlap between ground truth and prediction is not complete. There is a mismatch for low field values in the sense that the NN does not predict low non-zero field values correctly. Thus, we recommend using MAE to avoid this issue.

4. DISCUSSION

In large-scale mechanistic simulations of biological tissues, calculations of the diffusion of molecular species can be a significant fraction of the total computational cost. Because biological responses to concentrations often have a stochastic overlay, high precision may not be essential in these calculations Because NN surrogate estimates are significantly faster than the explicit calculation of the steady-state diffusion field for a given configuration of sources and sinks, an effective NN surrogate could greatly increase the practical size of simulated tissues, e.g., in cardiac simulations (Kerckhoffs et al., 2007; Sundnes et al., 2014), cancer simulations (Bruno et al., 2020), and orthopedic simulations (Erdemir et al., 2007). In our case, using a NVIDIA Quadro RTX 6000, each diffusion solution is about 1,000 times faster using the trained NN solver compared to the Julia code.

In order to decide if this acceleration is useful, we have to consider how long it takes to run the direct simulation, how long the NN takes to train and how long it takes to execute the NN once it has been trained (Fox and Jha, 2019). If each diffusion calculation takes δ seconds to run, conducting N calculations directly takes $t_{direct} = N\delta$. If each neural network surrogate takes ϵ seconds to run, and the number of replicas in the training set is M and the training time is E, the total time for the neural network simulation is the time to generate the training set, the training time plus the simulation time, $t_{neuro} = M\delta + E + N\epsilon$. To estimate these times, we ran 20,000 explicit simulations in Julia, which took ~6 h and 30 min, yielding roughly 1.16s each. The NN training time was 12 h on average. While the speedup for an individual simulation is $\delta/\epsilon \approx 1,000$, the ratio $\tau_{neuro}/\tau_{direct}$ must be smaller than 1 in order to have a useful acceleration. Equating this ratio to 1 and solving for N yields

$$N_{min} = \frac{M + E/\delta}{1 - \epsilon/\delta} \approx M + \frac{E}{\delta}\,. \tag{4}$$

N_{min} gives the number of replicas necessary for the total time using the NN to be the same as the direct calculation. Of course, the exact times will depend on the specific hardware used for the direct and NN calculations. In our case, from Equation (4) we obtain that $N_{min} \approx 57,300$, we would need to use the neural network more than 57,300 times for the total time using the NN to be faster than the direct calculation. Thus the NN acceleration is primarily useful in simulations that will be run many, many times for the specific situation for which the NN is appropriate. Consider for example if one wishes to include a variable number of sources, different lattice sizes, different dimensionalities (e.g., 3D) and boundary conditions. The more general the NN the more training data it will require, the longer training will take, and the slower the individual NN calculations will be. Currently virtual-tissue simulation studies often run thousands to tens of thousands of replicas and each replica often takes tens of minutes to tens of hours to run. This computational cost makes detailed parameter identification and uncertainty quantification impractical, since simulations often have dozens of parameters to explore. If using a NN-based diffusion solver accelerated these simulations by 100× it would permit practical studies with hundreds of thousands to

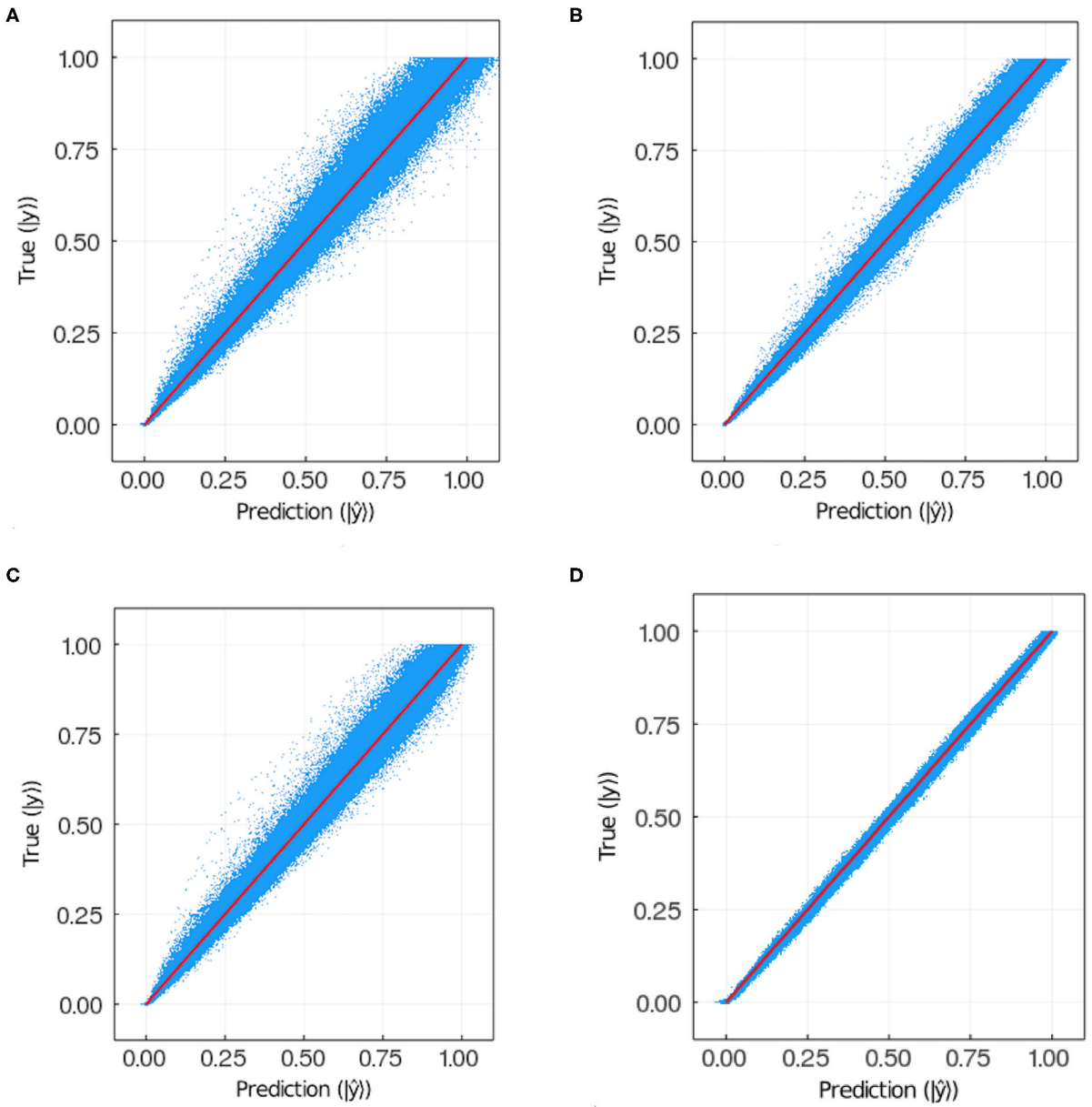


FIGURE 8 | Ground truth vs. prediction for **(A)** test set and **(B)** training set in the case of model 9; **(C)** test set and **(D)** training set in the case of model 11 (see **Table 2**). The number of points plotted in each panel is $3.75 \cdot 10^7$.

millions of replicas, greatly expanding the feasible exploration of parameter space for parameter identification and uncertainty quantification. It is worthwhile mentioning that there are other numerical methods for diffusion in 2D and 3D models that can also exploit the GPU parallelization such as the one in Secomb (2016) based on a discretization of Green's function. Our focus is on the ability of neural-network surrogates to solve the time-independent diffusion equation, however it would be interesting to extensively optimize the mechanistic methods we used to generate our training data sets. Generating training data is so time consuming, applications of deep neural networks will benefit greatly from using faster mechanistic methods to generate training data.

While there isn't a protocol for setting up a diffusion-solver surrogate, there are several things that must be considered. First one needs to frame the problem similar to how one would do when performing mechanistic modeling. One needs to settle on the dimensionality, e.g., 1D, 2D, 3D,… or n-dimensions; the system size which in our case we settled on 100×100; the type of sources to consider, e.g., sinks, sources, or both; the boundary conditions e.g., absorbing, reflective, periodic, or mixed; the distribution of sources in space; and it is also

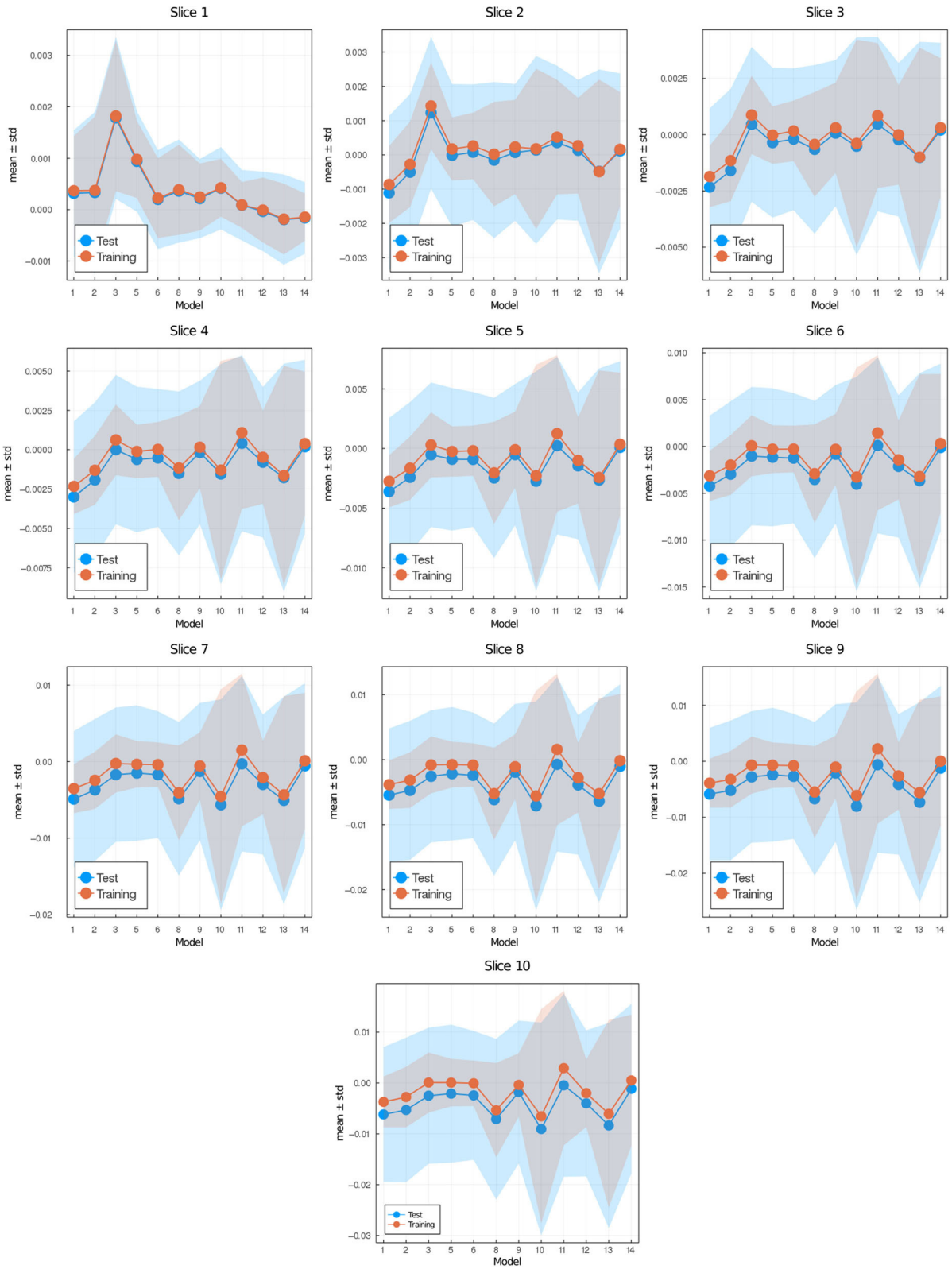


FIGURE 9 | Mean (data points) ± standard deviation (envelop) per slice vs. models (see **Table 1**) for test set (blue) and training set (red). Slice i corresponds to field values in the interval $[0.1 \cdot (i-1), 0.1 \cdot i]$ where $i = 1, \ldots, 10$.

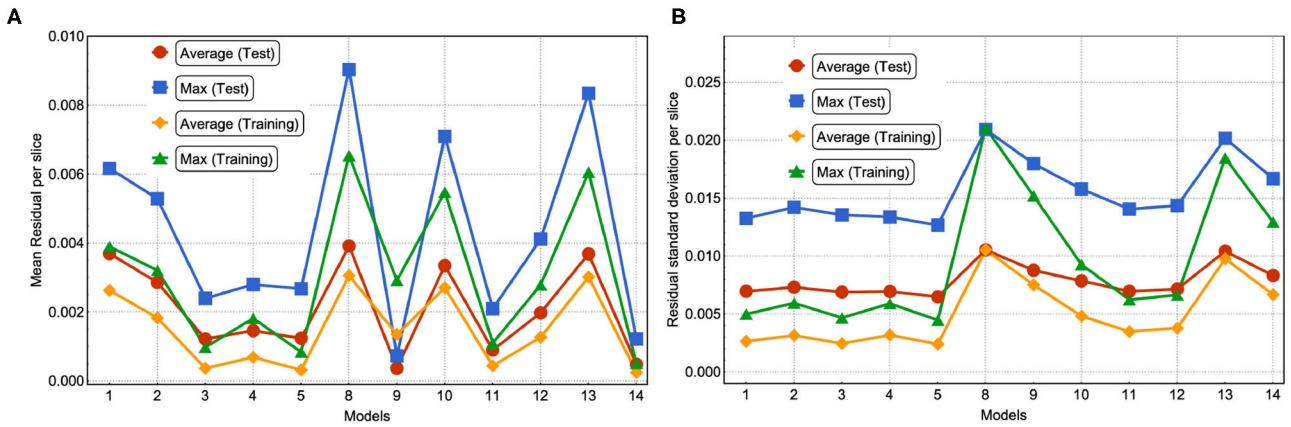


FIGURE 10 | (A) For each model, we show the average and maximum over the residual mean value per slice. **(B)** For each model, we show the average and maximum over the residual standard deviation per slice (see **Figure 9**). This was done for the test and training set.

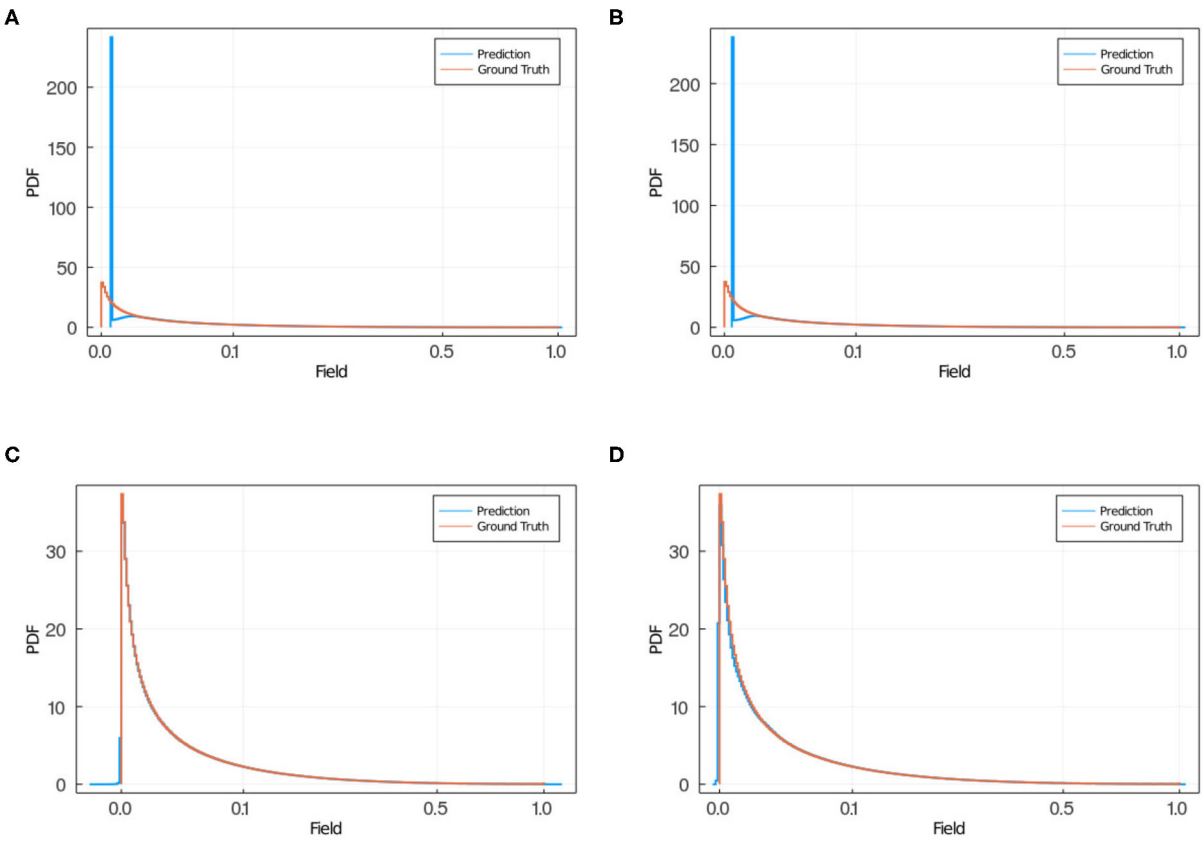


FIGURE 11 | PDF of field obtained via NN (blue) and ground truth (red) in the case of training using MSE, for **(A)** model 2 and **(B)** model 3 and for training using MAE, for **(C)** model 11 and **(D)** model 12. When using MSE **(A,B)** the NN predicts zero field values instead of low non-zero field values as the predicted PDF has a larger peak in zero than the ground truth PDF, and a smaller PDF for small non-zero field values compared with the ground truth PDF. When training using MAE **(C,D)** the prediction and ground truth PDFs overlap completely.

important to think about the accuracy required from the neural network. There isn't a rigorous way to determine the size of the required training dataset, although the size will depend on the problem one is addressing and the decisions made in the previous step. We recommend to start with a training dataset size of the order $\sim 10^4$. Then one needs to decide on the network architecture. For the network architecture the number of options is large. For instance, the depth of the neural network, e.g., deep or shallow; the type of layers, e.g., convolutional layers, fully-connected layers, recurrent layers, or mixed layers; the

activation functions, e.g., ReLU, sigmoid, tanh, etc. Evidence suggests that using deep neural networks as opposed to shallow neural networks will increase the non-linearity of the neural network, which ultimately broadens the learning capabilities of the neural network (Bianchini and Scarselli, 2014). But as depth increases the gradient of the loss function can grow or diminish significantly leading to instabilities or a regime of *zero learning* where the gradient becomes zero but the loss function value is large. Very deep neural networks can also lead to an *overflow* or *underflow* situation. Therefore, the neural network depth is a feature that should be set in way that meets a middle ground. The right choice of activation functions, regularizer layers (i.e., DropOut, BatchNorm, etc.) and weight initializers can hinder the unwanted features of instabilities or *zero learning*. Choosing the right optimizer for training is also something to consider. However, unless there are some very specific needs, the standard rule is to use the ADAM optimizer (Kingma and Ba, 2014). Choosing the loss function is crucial as this is the metric the neural network will use to measure how *good* the outcome is. Typically for these type of problems MSE, MAE and other similar norms are used. Additionally, there are a number of (hyper)parameters to be chosen. For instance, some activation functions, regularizer layer and optimizers have hyperparameters. Also the number of epochs and size of minibatch are hyperparameters. To set the hyperparameters' values, one can start by using the values reported in the literature but the scope should be to explore the space of hyperparameters by training an ensemble of neural networks with different hyperparameters and then choosing the model that performed best on the validation set.

In a real tissue, the oxygen tension on the surface of the blood vessel and in the tissue as a whole involves complex feedback among many factors, including spatial and temporal variation in the supply and consumption of oxygen; supply at a given location could depend on the degree of local blood-vessel dilation, the rate of blood flow, and levels of oxygen in the blood to name a few examples. A realistic model of oxygenation in tissue would need to include spatial and temporal models of all of these processes individually and of their coupling. Clearly such a model is much more complex than our simple example of calculating the steady-state oxygen field given a fixed set of circular sources with fixed oxygen tensions and a fixed uniform consumption rate in the tissue implemented as linear decay.

While developing NN surrogates to solve the entire complex problem of oxygenation would be worthwhile, we believe that deep neural network surrogates will (at least initially) not replace the entire simulation, but to replace the most computationally costly components of the simulation. In this case, looking for surrogates for specific commonly-used calculations, which can be used in many different applications and which can provide a substantial speed-up is appropriate. Many biophysical and engineering problems require solving the diffusion equation for fixed sources. Despite the improvements to direct solution mentioned in Secomb (2016), solving the diffusion equation still often contributes much of the computational cost of the full problem solution. In these cases, the faster the "diffusion step" is computed, the faster the solution of the multiscale model as a whole. To train an optimal diffusion surrogate for a particular problem one has to choose a set of appropriate loss functions and combine them to minimize the errors of the metrics one defines as most relevant to the specific problem being addressed. How to choose loss functions and their weighting to achieve macroscopic desired outcomes is not well understood as a general problem. Even in our very simple example, we had to explore a wide variety of loss functions to achieve reasonable convergence of our NN during training and reasonable final absolute and relative accuracy of our surrogate.

5. CONCLUSIONS

Neural networks provide many possible approaches to generating surrogate diffusion solvers. Given the type of problem setting, we were interested in a neural network that could predict the stationary field. We considered a deep convolutional neural network, an autoencoder and their combination. We considered two loss functions, *viz.* mean squared error and mean absolute error. We considered different hyperparameters for dropout and an exponential weight to compensate the under-sampling of high field values. The exponential weight also helped reduce overfitting as shown in **Figure 8**.

The range of scientific and engineering applications for diffusion solvers is very broad. Depending on the specific application, the predictions by the neural network will have to meet a specific set of criteria quantified in the form of statistical estimators (e.g., mean error, max error, percentiles, mean relative error, *etc*.). In this paper we studied several reasonable error metrics, namely, mean residual, maximum residual, 99-Percentile residual, mean relative residual, mean weighted residual and the weighted standard deviation residual. The last two metrics compensate for the low frequency of high field values, ones that usually occur in small regions around sources. The autoencoders are commonly used in generative models which is applicable, as we have shown here, to the case of a diffusion surrogate. The field predictions are accurate on all the metrics we considered. This is appears to be due to collapsing the input into a one-dimensional vector and then decoding back to the initial size, which forces the network to learn the relevant features (Kingma and Welling, 2019). While some models had high errors across all metrics, no single model had the smallest error for all error metrics. Different networks and hyperparameters were optimal for different metrics, e.g., model 5 had the lowest mean residual, whereas model 9 yielded relatively good results on all metrics. Model 9 uses both neural networks with the dropout values for the deep convolutional network were set to $D_{1,2} = 0.4$, and for the autoencoder to $D_{3,4} = 0.1$. The weight hyperparameter was set to 100. Recall that large weight hyperparameter values make the loss function weight high field values over low field values. This is important since the largest absolute error happens close to sources and close to boundaries because of the under-representation of these kinds of configurations. We also noticed that this choice reduced the overfitting as was shown in **Figure 8**.

Additionally, we tested several loss function. Here we reported the results using mean squared error and mean absolute error. We noticed two key differences. With MSE the weighted standard deviation (see **Figure 9**) is smaller than for MAE for the training set. However, for the test set, the results for both loss functions are comparable. This difference between training and test sets suggests that MSE is more prone to overfitting the data than MAE. The other key difference is that for the MAE, the predicted field probability function consistently overlapped the ground truth completely, whereas for MSE there is a mismatch in that the NN does not predict low non-zero field values correctly (see **Figure 11**). Therefore, we recommend using MAE as the loss function for surrogate calculations where the field values are well bounded, as we have shown it produces better predictions than MSE. The autoencoder (NN 2) is capable of approximating the diffusion field on its own, the convolutional network (NN 1) is not. However, if we use the two networks together we find that the prediction is more accurate than NN 2 alone.

These encouraging results suggest that we should pursue NN surrogates for acceleration of simulations in which the solution to the diffusion equation contributes a considerable fraction of the total computational cost. An effective NN diffusion solver surrogate would need to be able to solve diffusion fields for arbitrary sources and sinks in two or three dimensions with variable diffusivity, a much higher dimensional set of conditions than the two circular sources in a uniform two-dimensional square domain that we investigated in this paper. A key question will be the degree to which NNs are able to generalize, e.g., from n sources to $n+1$ sources or from circular sources to more complex shapes. In addition, here we only considered absorbing boundary conditions, ultimately mixed boundary conditions are desirable. It is unclear if the best approach would be a single NN capable of doing multiple boundary conditions, or better to develop unique NNs for each boundary condition scenario. While in this paper we have only considered zero-field boundary conditions mainly due to feasibility purposes for the neural network, we will consider different boundary conditions in future work.

Increasing the number and size of vessels is a combinatorial problem in the dimensionality of the training set, but it ultimately doesn't change the nature of the diffusion equation. Thus, we expect that a straightforward approach consisting using a bigger training set including a greater variety of source and sink sizes, shapes, and number, should still work, though it will take more computing time to generate the training data and train the network. The ability of greens-function methods to solve the diffusion equation for arbitrary numbers of sources and sinks suggests (though it does not prove) that such generalization should work also for neural network solvers.

To solve 3D diffusion problems, the most straightforward extension of our method would be to use 3D convolutional neural networks. However, there may be some difficulties with a naive extension of our convolutional methods to 3D. If we have a linear dimension of L then the output layer of the NN has L^2 elements in 2D and L^3 in 3D. Thus, for a given value of L, the network size is much larger in 3D. Besides the size of the network, the training set will also be larger. For N sources, the number of possible configurations grows roughly as L^{2N} in 2D, while the number of configurations in 3D is L^{3N}. In addition, if we wish to represent realistic sources in 3D, like blood vessels, we need to sample over appropriately spatially-correlated patterns of sources rather than the randomly located spherical sources we used in our 2D example. Naively these very high dimensions of possible source configurations suggest that the 3D problem would require impossibly large training datasets. However, one of the outstanding features of deep neural networks is their capacity to extrapolate from apparently severely undersampled training sets, so increasing the number of possible configurations exponentially does not necessarily imply the need to increase the training set exponentially. Another approach to develop diffusion solver surrogates in 3D is to build physically informed neural networks (PINNs) (Raissi et al., 2019) where the ODE describing the process, the initial conditions and the boundary conditions are embedded in the loss function. Other efforts attempt to tackle the *curse of dimensionality* by physical intuition embedded in the neural network architecture (Roberts, 2021). We will explore these issues in future work.

DATA AVAILABILITY STATEMENT

The datasets presented in this study can be found in online repositories. The names of the repository/repositories and accession number(s) can be found at: https://github.com/jquetzalcoatl/DiffusionSurrogateDeepLearning.

AUTHOR CONTRIBUTIONS

JG and JS proposed the project. JT-M and GF built the models. JT-M trained the models. All authors analyzed the results and wrote the manuscript.

FUNDING

This research was supported in part by Lilly Endowment, Inc., through its support for the Indiana University Pervasive Technology Institute. This work was partially supported by the National Science Foundation (NSF) through awards nanoBIO 1720625, CINES 1835598 and Global Pervasive Computational Epidemiology 1918626, and Cisco University Research Program Fund grant 2020-220491. This work was partially supported by the Biocomplexity Institute at Indiana University, National Institutes of Health, grant NIGMS R01 GM122424. This work was partially supported by the DOE ASCR Award DE-SC0021418 "FAIR Surrogate Benchmarks Supporting AI and Simulation Research".

ACKNOWLEDGMENTS

We thank the referees for suggesting different ways to extend our work. This is a short text to acknowledge the contributions of specific colleagues, institutions, or agencies that aided the efforts of the authors.

SUPPLEMENTARY MATERIAL

The Supplementary Material for this article can be found online at: https://www.frontiersin.org/articles/10.3389/fphys.2021.667828/full#supplementary-material

REFERENCES

Baur, C., Denner, S., Wiestler, B., Navab, N., and Albarqouni, S. (2020). Autoencoders for unsupervised anomaly segmentation in brain MR images: a comparative study. *Med. Image Anal.* 2020:101952. doi: 10.1016/j.media.2020.101952

Bianchini, M., and Scarselli, F. (2014). On the complexity of neural network classifiers: a comparison between shallow and deep architectures. *IEEE Trans. Neural Netw. Learn. Syst.* 25, 1553–1565. doi: 10.1109/TNNLS.2013.2293637

Bruno, R., Bottino, D., de Alwis, D. P., Fojo, A. T., Guedj, J., Liu, C., et al. (2020). Progress and opportunities to advance clinical cancer therapeutics using tumor dynamic models. *Clin. Cancer Res.* 26, 1787–1795. doi: 10.1158/1078-0432.CCR-19-0287

Cai, S., Wang, Z., Wang, S., Perdikaris, P., and Karniadakis, G. (2021). Physics-informed neural networks (PINNS) for heat transfer problems. *J. Heat Transf.* 143:060801. doi: 10.1115/1.4050542

Champion, K., Lusch, B., Kutz, J. N., and Brunton, S. L. (2019). Data-driven discovery of coordinates and governing equations. *Proc. Natl. Acad. Sci. U.S.A.* 116, 22445–22451. doi: 10.1073/pnas.1906995116

Chen, M., Shi, X., Zhang, Y., Wu, D., and Guizani, M. (2017). Deep features learning for medical image analysis with convolutional autoencoder neural network. *IEEE Trans. Big Data*. 1-1. doi: 10.1109/TBDATA.2017.2717439

Chen, R. T. Q., Rubanova, Y., Bettencourt, J., and Duvenaud, D. K. (2018). "Advances in neural information processing systems," in *Neural Ordinary Differential Equations*, Vol. 31, eds S. Bengio, H. Wallach, H. Larochelle, K. Grauman, N. Cesa-Bianchi, R. Garnett (Curran Associates, Inc.)

Chen, W., and Ferguson, A. L. (2018). Molecular enhanced sampling with autoencoders: on-the-fly collective variable discovery and accelerated free energy landscape exploration. *J. Comput. Chem.* 39, 2079–2102. doi: 10.1002/jcc.25520

Dubois, P., Gomez, T., Planckaert, L., and Perret, L. (2020). Data-driven predictions of the lorenz system. *Phys. D* 2020:132495. doi: 10.1016/j.physd.2020.132495

Edalatifar, M., Tavakoli, M. B., Ghalambaz, M., and Setoudeh, F. (2020). Using deep learning to learn physics of conduction heat transfer. *J. Therm. Anal. Calorim.* 1–18. doi: 10.1007/s10973-020-09875-6

Erdemir, A., McLean, S., Herzog, W., and van den Bogert, A. J. (2007). Model-based estimation of muscle forces exerted during movements. *Clin. Biomech.* 22, 131–154. doi: 10.1016/j.clinbiomech.2006.09.005

Farimani, A. B., Gomes, J., and Pande, V. S. (2017). Deep learning the physics of transport phenomena. *arXiv preprint. arXiv:1709.02432*.

Fox, G., and Jha, S. (2019). "Learning everywhere: a taxonomy for the integration of machine learning and simulations," in *2019 15th International Conference on eScience (eScience)*, (San Diego, CA), 439–448. doi: 10.1109/eScience.2019.00057

Geoffrey, F. (2020). *Draft Deep Learning for Spatial Time Series*. Technical Report. Available online at: https://www.dsc.soic.indiana.edu

Gkeka, P., Stoltz, G., Farimani, A. B., Belkacemi, Z., Ceriotti, M., Chodera, J., et al. (2020). Machine learning force fields and coarse-grained variables in molecular dynamics: application to materials and biological systems. *arXiv preprint arXiv:2004.06950*. doi: 10.1021/acs.jctc.0c00355

He, H., and Pathak, J. (2020). An unsupervised learning approach to solving heat equations on chip based on auto encoder and image gradient. *arXiv preprint arXiv:2007.09684*.

Innes, M. (2018). Flux: elegant machine learning with Julia. *J. Open Source Softw.* 3:602. doi: 10.21105/joss.00602

Innes, M., Saba, E., Fischer, K., Gandhi, D., Rudilosso, M. C., Joy, N. M., et al. (2018). Fashionable modelling with flux. *CoRR, abs/1811.01457*.

Kasim, M., Watson-Parris, D., Deaconu, L., Oliver, S., Hatfield, P., Froula, D. H., et al. (2020). Up to two billion times acceleration of scientific simulations with deep neural architecture search. *arXiv preprint arXiv:2001.08055*.

Kerckhoffs, R. C., Neal, M. L., Gu, Q., Bassingthwaighte, J. B., Omens, J. H., and McCulloch, A. D. (2007). Coupling of a 3d finite element model of cardiac ventricular mechanics to lumped systems models of the systemic and pulmonic circulation. *Ann. Biomed. Eng.* 35, 1–18. doi: 10.1007/s10439-006-9212-7

Kingma, D. P., and Ba, J. (2014). Adam: Amethod for stochastic optimization. *arXiv preprint arXiv:1412.6980*.

Kingma, D. P., and Welling, M. (2019). An introduction to variational autoencoders. *arXiv preprint arXiv:1906.02691*. doi: 10.1561/9781680836233

Lee, Y., Veerubhotla, K., Jeong, M. H., and Lee, C. H. (2020). Deep learning in personalization of cardiovascular stents. *J. Cardiovasc. Pharmacol. Therap.* 25, 110–120. doi: 10.1177/1074248419878405

Li, A., Chen, R., Farimani, A. B., and Zhang, Y. J. (2020). Reaction diffusion system prediction based on convolutional neural network. *Sci. Rep.* 10, 1–9. doi: 10.1038/s41598-020-60853-2

Li, Z., Kovachki, N., Azizzadenesheli, K., Liu, B., Bhattacharya, K., Stuart, A., et al. (2020). Fourier neural operator for parametric partial differential equations. *arXiv preprint arXiv:2010.08895*.

Linsenmeier, R. A., and Zhang, H. F. (2017). Retinal oxygen: from animals to humans. *Prog. Retinal Eye Res.* 58, 115–151. doi: 10.1016/j.preteyeres.2017.01.003

Noé, F., Olsson, S., Köhler, J., and Wu, H. (2019). Boltzmann generators: sampling equilibrium states of many-body systems with deep learning. *Science* 365:eaaw1147. doi: 10.1126/science.aaw1147

Noé, F., Tkatchenko, A., Müller, K.-R., and Clementi, C. (2020). Machine learning for molecular simulation. *Annu. Rev. Phys. chem.* 71, 361–390. doi: 10.1146/annurev-physchem-042018-052331

Phillips, R. (2018). "Membranes by the numbers," in *Physics of Biological Membranes*, eds P. Bassereau and S. Pierre (Springer), 73–105. doi: 10.1007/978-3-030-00630-3_3

Rackauckas, C., Innes, M., Ma, Y., Bettencourt, J., White, L., and Dixit, V. (2019). Diffeqflux.jl - A julia library for neural differential equations. *CoRR, abs/1902.02376*.

Rackauckas, C., and Nie, Q. (2017). Differentialequations.jl-a performant and feature-rich ecosystem for solving differential equations in Julia. *J. Open Res. Softw.* 5:15. doi: 10.5334/jors.151

Raissi, M., Perdikaris, P., and Karniadakis, G. E. (2019). Physics-informed neural networks: a deep learning framework for solving forward and inverse problems involving nonlinear partial differential equations. *J. Comput. Phys.* 378, 686–707. doi: 10.1016/j.jcp.2018.10.045

Roberts, D. A. (2021). *Why is AI Hard and Physics Simple*? Technical report. *arxiv preprint. arXiv:2104.00008*

Schiesser, W. E. (2012). *The Numerical Method of Lines: Integration of Partial Differential Equations*. San Diego, CA: Elsevier.

Secomb, T. W. (2016). A green's function method for simulation of time-dependent solute transport and reaction in realistic microvascular geometries. *Math. Med. Biol.* 33, 475–494. doi: 10.1093/imammb/dqv031

Sharma, R., Farimani, A. B., Gomes, J., Eastman, P., and Pande, V. (2018). Weakly-supervised deep learning of heat transport via physics informed loss. *arXiv preprint arXiv:1807.11374*.

Sundnes, J., Wall, S., Osnes, H., Thorvaldsen, T., and McCulloch, A. D. (2014). Improved discretisation and linearisation of active tension in strongly coupled cardiac electro-mechanics simulations. *Comput. Methods Biomech. Biomed. Eng.* 17, 604–615. doi: 10.1080/10255842.2012.704368

Tikhonov, A. N., and Samarskii, A. A. (2013). *Equations of Mathematical Physics*. New York, NY: Courier Corporation. Available online at: https://books.google.ca/books?hl=en&lr=&id=PTmoAAAAQBAJ&oi=fnd&pg=PP1&dq=Tikhonov+and+Samarskii,+2013&ots=kYk8H8xNep&sig=pNN3S9TRC0RhHhlZYdobzBWT66Y&redir_esc=y#v=onepage&q=Tikhonov%20and%20Samarskii%2C%202013&f=false

Toledo-Marin, J. Q. (2020). *Stationary Diffusion State Ml Surrogate Using Flux and Cuarrays*. Available online at: https://github.com/jquetzalcoatl/DiffusionSurrogate (accessed June 09, 2021).

Xie, D., Li, Y., Yang, H., Song, D., Shang, Y., Ge, Q., et al. (2019). "Bold fMRI-based brain perfusion prediction using deep dilated wide activation networks," in *International Workshop on Machine Learning in Medical Imaging* (Shenzhen: Springer), 373–381. doi: 10.1007/978-3-030-32692-0_43

Zhang, H., Hippalgaonkar, K., Buonassisi, T., Løvvik, O. M., Sagvolden, E., and Ding, D. (2019). Machine learning for novel thermal-materials discovery: early successes, opportunities, and challenges. *arXiv preprint arXiv:1901.05801*. doi: 10.30919/esee8c209

Zhang, J.-S., and Xiao, X.-C. (2000). Predicting chaotic time series using recurrent neural network. *Chinese Phys. Lett.* 17:88. doi: 10.1088/0256-307X/17/2/004

Conflict of Interest: The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

frontiers in Physiology

ORIGINAL RESEARCH
published: 16 July 2021
doi: 10.3389/fphys.2021.702975

Bayesian Physics-Based Modeling of Tau Propagation in Alzheimer's Disease

Amelie Schäfer[1], Mathias Peirlinck[1], Kevin Linka[2], Ellen Kuhl[1] and the Alzheimer's Disease Neuroimaging Initiative (ADNI)[†]*

[1] Department of Mechanical Engineering, Stanford University, Stanford, CA, United States, [2] Institute of Continuum and Materials Mechanics, Hamburg University of Technology, Hamburg, Germany

OPEN ACCESS

Edited by:
Nicole Y. K. Li-Jessen,
McGill University, Canada

Reviewed by:
Sara Garbarino,
University of Genoa, Italy
Bratislav Misic,
McGill University, Canada

***Correspondence:**
Amelie Schäfer
amesch@stanford.edu

[†] Data used in preparation of this article were obtained from the Alzheimer's Disease Neuroimaging Initiative (ADNI) database (adni.loni.usc.edu). As such, the investigators within the ADNI contributed to the design and implementation of ADNI and/or provided data but did not participate in analysis or writing of this report. A complete listing of ADNI investigators can be found at: http://adni.loni.usc.edu/wp-content/uploads/how_to_apply/ADNI_Acknowledgement_List.pdf

Specialty section:
This article was submitted to Computational Physiology and Medicine, a section of the journal Frontiers in Physiology

Received: 30 April 2021
Accepted: 22 June 2021
Published: 16 July 2021

Citation:
Schäfer A, Peirlinck M, Linka K, Kuhl E and the Alzheimer's Disease Neuroimaging Initiative (ADNI) (2021) Bayesian Physics-Based Modeling of Tau Propagation in Alzheimer's Disease. Front. Physiol. 12:702975. doi: 10.3389/fphys.2021.702975

Amyloid-β and hyperphosphorylated tau protein are known drivers of neuropathology in Alzheimer's disease. Tau in particular spreads in the brains of patients following a spatiotemporal pattern that is highly sterotypical and correlated with subsequent neurodegeneration. Novel medical imaging techniques can now visualize the distribution of tau in the brain *in vivo*, allowing for new insights to the dynamics of this biomarker. Here we personalize a network diffusion model with global spreading and local production terms to longitudinal tau positron emission tomography data of 76 subjects from the Alzheimer's Disease Neuroimaging Initiative. We use Bayesian inference with a hierarchical prior structure to infer means and credible intervals for our model parameters on group and subject levels. Our results show that the group average protein production rate for amyloid positive subjects is significantly higher with 0.019 ± 0.27/yr, than that for amyloid negative subjects with -0.143 ± 0.21/yr ($p = 0.0075$). These results support the hypothesis that amyloid pathology drives tau pathology. The calibrated model could serve as a valuable clinical tool to identify optimal time points for follow-up scans and predict the timeline of disease progression.

Keywords: Alzheimer's disease, network diffusion model, tau PET, Bayesian inference, hierarchical modeling, uncertainty quantification

1. INTRODUCTION

Alzheimer's disease currently affects one out of 10 adults over the age of 65 in the United States (Association, 2019). Due to demographic changes worldwide, the prevalence and public health impact of this neurodegenerative disease is projected to more than double in the next 30 years. Effective therapeutic interventions require early diagnosis and a detailed understanding of the early mechanisms driving pathology. For Alzheimer's disease, this poses a particular challenge since clinical diagnosis is currently possible only with the appearance of cognitive impairment at late disease stages. We now know that the first pathological changes which initiate the disease may happen up to decades before the presence of cognitive symptoms (Bateman et al., 2012; Jack et al., 2013). Investigating these early disease mechanisms is crucial, if we want to understand the timeline of disease progression and identify early access points for intervention.

It is well accepted that two proteins, amyloid-β and tau, play a major role in disease initiation and represent important biomarkers for disease progress (Duyckaerts et al., 2009). Amyloid and tau are both present in the healthy brain, but have been found to accumulate and aggregate in abnormal amounts and pathological forms in the brains of Alzheimer's patients. The amyloid hypothesis

states that at the early stages of disease, amyloid-β starts to accumulate widely across the neocortex. Subsequently, hyperphosphorylated tau starts to accumulate and aggregate in neurofibrillary tangles in more and more areas of the brain, ultimately causing neurodegeneration and cognitive impairment (Jack and Holtzman, 2013). The sequence of when and where neurofibrillary tangles of tau emerge has been shown to follow a highly reproducible pattern. Cross-sectional autopsy studies have confirmed that tangles first appear in the transentorhinal and entorhinal cortex in early disease stages, then emerge in the neighboring hippocampus and regions of the temporal lobe, before ultimately spreading into more distantly connected areas of the neocortex (Braak and Braak, 1991; Braak et al., 2006). There is strong evidence from animal and imaging studies that hyperphosphorylated tau spreads intracellularly along axons in the brain (De Calignon et al., 2012; Liu et al., 2012; Jones et al., 2017; Pereira et al., 2019), explaining how the pathology propagates from the entorhinal cortex to connected regions. Several studies have found links between amyloid and tau, suggesting that amyloid pathology is a precursor for tau pathology and influences the distribution of neurofibrillary tangles in the brain (Price and Morris, 1999; Musiek and Holtzman, 2012; Jack et al., 2013). Tau itself has been found to be strongly correlated with tissue atrophy and neurodegeneration, making it a predictor for cognitive impairment at later disease stages (Harrison et al., 2019; La Joie et al., 2020).

The consistency of tau's spatiotemporal progression and its confirmed direct correlation with neurodegeneration make it an optimal target for computational modeling. Personalized models of tau pathology could serve as a tool to predict individual disease progression timelines and as simulated controls in clinical trials. In the latter context, the model may be leveraged to predict how tau would develop in a test subject over time without intervention which can then be compared to the actual developments in the test subject with interventions targeting tau aggregation (Congdon and Sigurdsson, 2018). Multiple groups have proposed network diffusion and epidemic spreading models to simulate the spatiotemporal propagation in the brain for pathological proteins in general (Iturria-Medina et al., 2014; Weickenmeier et al., 2019; Garbarino et al., 2021), and for tau in particular (Raj et al., 2012; Torok et al., 2018; Fornari et al., 2019; Vogel et al., 2020) with good qualitative results. Until recently, the only way to measure the distribution of tau in the brain was through postmortem histology or by making assumptions about the relationship between tau and tissue atrophy observed in structural MRI scans (Raj et al., 2012; Torok et al., 2018). The resulting lack of data has posed significant challenges for calibration of computational tau models. However, an emerging molecular imaging technique, positron emission tomography (PET), now enables us to track the distribution of hyperphosphorylated tau in the brain *in vivo* (Johnson et al., 2016; Villemagne et al., 2018). As the technique is maturing, the amount of available data is growing steadily, allowing us to computationally comprehend the tau pathology in individual subjects over time and use this data for model calibration.

In a recent study, we have shown that we can successfully fit a network diffusion model based on a weighted Laplacian graph of the axonal connectome to longitudinal tau PET data of 46 subjects using a deterministic optimization approach (Schäfer et al., 2020). With tau PET becoming a more established component of longitudinal imaging studies, the amount of available data is steadily increasing, setting the ground for data-driven modeling techniques. Here we use Bayesian hierarchical modeling (Peirlinck et al., 2019) to calibrate the same network diffusion model to longitudinal imaging data from 76 subjects from the Alzheimer's Disease Neuroimaging Initiative (ADNI, 2020). Introducing this probabilistic approach to replace our previous deterministic optimization allows us to account for potential uncertainties in image acquisition and processing, and at the same time, quantify the uncertainty in our model calibration. Identifying the uncertainty in our model parameters is essential to determine the accuracy of our personalized model predictions. If clinical scientists and study designers are to use our model, it is crucial to quantify the accuracy of the simulation. Only then can they determine for which subjects the disease course can be confidently inferred from the available data and for which subjects additional data may be needed to make accurate enough projections. It may also inform them at which time points to acquire additional data to most efficiently improve model accuracy. The hierarchical structure we chose here to represent our model parameters on group and subject levels, will help us gain a better understanding of variability and commonalities of tau pathology between subjects.

2. MATERIALS AND METHODS

Figure 1 gives an overview of our methods. In summary, we obtain regional tau uptake values from longitudinal tau PET images through a process of image registration, segmentation, and region of interest analysis. We assume that the propagation of misfolded tau in the brain can be described by a network diffusion model characterized by two model parameters, diffusion coefficient and production rate. After defining weakly informative prior distributions for those model parameters we use a Markov Chain Monte Carlo algorithm to smartly sample from the priors. Inserting the sampled parameters into our model and comparing the resulting simulated tau uptake with the observed data then allows us to rate each sample based on its likelihood and apply Bayes' theorem to determine the posterior distributions of most likely parameter values for each subject.

2.1. Network Diffusion Model

We model the accumulation and propagation of hyperphosphorylated tau in the brain's connectome as a diffusion problem on a weighted, undirected graph $\mathcal{G}$ with N nodes, representing different brain regions, and E edges, representing axonal connections between those brain regions. We use the Budapest Reference Connectome v. 3.0 (Szalkai et al., 2017) to obtain the graph $\mathcal{G}$ from processed diffusion tensor imaging data of 418 healthy subjects collected through the Human Connectome Project (McNab et al., 2013). From the original graph with $N = 1015$ nodes, we create a reduced graph with $N = 83$ nodes representing 83 cortical and subcortical brain regions. The edge weights of the network are defined

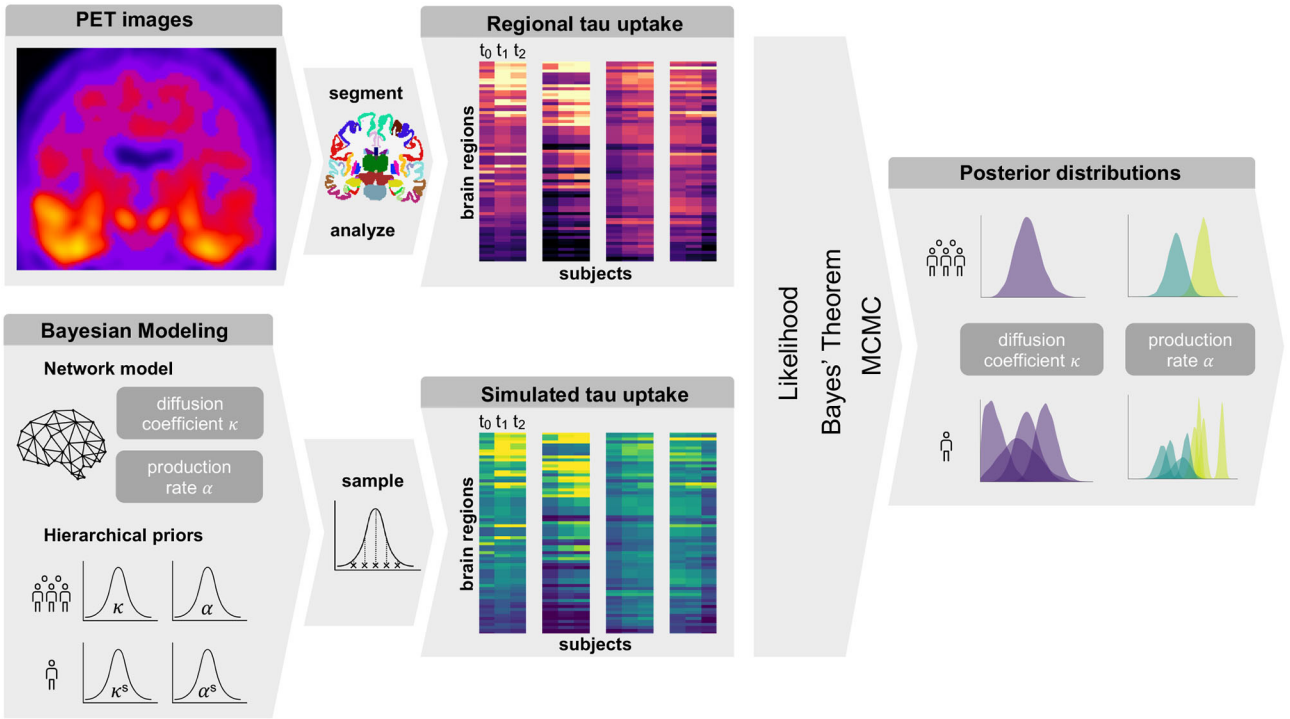


FIGURE 1 | Summary of Methods. Summary of our workflow including PET image analysis and Bayesian modeling to obtain personalized posterior distributions of two model parameters describing the tau pathology in the examined group of subjects.

by the number of fibers n_{ij} detected along the respective edge between the pair of nodes i and j, divided by the fiber length l_{ij} along this edge averaged across all 418 brains. The adjacency matrix A_{ij} of the graph, containing the edge weights for all connections, is thus computed as $A_{ij} = n_{ij}/l_{ij}$. The resulting network and its adjacency matrix are illustrated in **Figure 2**, showing a small number of strong and medium connections within and between the lobes of each hemisphere and only few connections between hemispheres.

We characterize the aggregation and spread of pathological tau within the brain connectome as a nonlinear reaction-diffusion problem governed by the Fisher-Kolmogorov equation (Fisher, 1937; Kolmogorov et al., 1937). This equation describes how the concentration of misfolded protein c evolves over time based on the assumption that tau pathology develops in a prion-like fashion (Jucker and Walker, 2011; Fornari et al., 2019, 2020).

$$\frac{\mathrm{d}c}{\mathrm{d}t} = \nabla \cdot (\mathbf{D} \cdot \nabla c(t)) + \alpha\, c(t)\,[\,1 - c(t)\,], \qquad (1)$$

Here, $\mathbf{D}$ denotes the diffusion tensor, which determines the speed and directionality of corruptive tau seed propagation, and α the local production rate, which captures the processes of protein production, clearance and conversion from healthy to unhealthy seeds (Fornari et al., 2019). In order to apply the diffusion model to our brain network, we discretize Equation (1) on the weighted graph $\mathcal{G}$. This leads to a discretized diffusion equation expressing for each node of the network $i = 1, ..., N$ the change in nodal concentration of misfolded protein c_i as

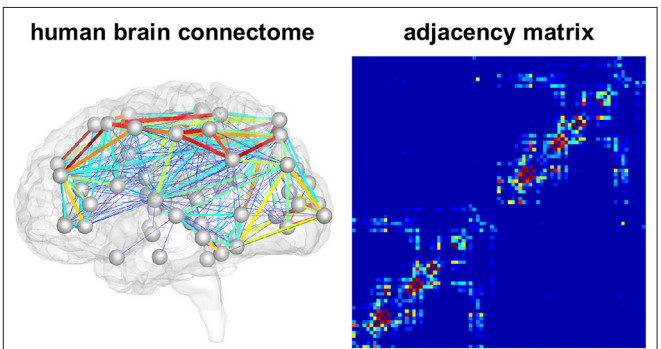


FIGURE 2 | Brain network model. Connectivity-weighted brain network and corresponding adjacency matrix. Colors represent the connection strength between two regions. Connectivity is moderate to strong within the two brain hemispheres while there are only few and weak connections between hemispheres.

$$\frac{\mathrm{d}c_i}{\mathrm{d}t} = -\kappa \sum_{j=1}^{N} L_{ij}\, c_j(t) + \alpha\, c_i(t)[\,1 - c_i(t)]. \qquad (2)$$

Equation (2) contains two model parameters, κ and α, which we can calibrate to individual patient data to reflect differences in disease dynamics across individuals. The diffusion coefficient κ

determines the transport rate of misfolded protein between two regions and α the production or clearance of pathological protein at each node. We assume these model parameters to be identical at all nodes $i = 1, ..., N$, but different between individuals. The weighted graph Laplacian L_{ij} summarizes the connectivity of the graph. Its diagonal terms contain information about how much protein diffuses out of node i into other nodes j and its non-diagonal terms describe how much protein enters node i from all other nodes j. The Laplacian is a square matrix constructed by subtracting the adjacency matrix A_{ij} from the degree matrix D_{ii},

$$L_{ij} = D_{ij} - A_{ij}. \tag{3}$$

The degree matrix D_{ii} is a diagonal matrix with each entry representing the sum of elements along a row of the adjacency matrix A_{ij},

$$D_{ii} = \text{diag} \sum_{j=1, j\neq i}^{N} A_{ij}. \tag{4}$$

2.2. Image Data

We use longitudinal imaging data from the Alzheimer's Disease Neuroimaging Initiative (ADNI) ADNI (2020) to initialize and calibrate our model. From the database, we select 76 subjects with at least three consecutive tau PET scans, which were acquired on average 1 year (1.07 ± 0.31) apart. This group contains a variety of clinical diagnoses, 31 subjects are diagnosed as cognitively normal, 15 with significant memory concern, 28 with mild cognitive impairment, and two with clinically confirmed Alzheimer's disease. Previously evaluated β-amyloid PET images identify 46 subjects as amyloid positive (Landau et al., 2013), meaning the average measured amyloid concentration in their brain exceeds a certain threshold value. We conduct our analysis blind to clinical diagnosis, but take amyloid status into account in our model structure.

All acquired AV1451-PET scans were processed according to standard ADNI protocols (ADNI, 2020). For each subject, we co-register the PET images to a corresponding high resolution T1 weighted magnetic resonance image (MRI) which we segmented into 68 cortical and 45 subcortical regions according to the Desikan-Killiany atlas (Desikan et al., 2006) using FreeSurfer (FreeSurfer, 2020). We use this segmentation to compute regional tracer uptake values from the PET images for the same 83 regions represented in our network model. We normalize these regional uptake values with respect to the uptake in the inferior cerebellar gray matter, which serves as our reference region, in order to gain regional standardized uptake value ratios (SUVR). Since PET recordings in subcortical regions and the hippocampus are known to be contaminated by off-target binding in the choroid plexus and nearby vascular structures (Lowe et al., 2016; Marquié et al., 2017; Lemoine et al., 2018), we focus our model calibration on the remaining 66 cortical regions.

Our network diffusion model delivers regional normalized tau concentrations c^{sim}, between zero, indicating that no misfolded protein is present, and one, indicating that a maximum amount of misfolded protein is present, $0 \leq c^{\text{sim}} \leq 1$. To compare simulated with observed protein concentrations, we need to map the tau PET standardized uptake value ratios into the same zero-to-one interval. To this end, we identify a lower threshold for tau positivity by fitting a Gaussian mixture model with two components to the cumulative raw tau PET data c^{raw} from all subjects, time points, and regions. Assuming that many of the included regions must be free from pathological tau, this allows us to determine the minimum raw PET value that should be considered positive. We declare all values below to this threshold of $c^{\text{raw}} = 1.1$ to be zero and normalize the remaining raw values such that $0 \leq c^{\text{pet}} \leq 1$.

2.3. Hierarchical Bayesian Inference

For each subject, we infer a personalized diffusion coefficient κ^{s} and protein production rate α^{s} most accurately reproducing the image data and quantify the uncertainty in our calibration using Bayesian inference. For each subject, we set the initial conditions of our model to the tau uptake values measured in the baseline PET scan $c^{\text{sim}}(t = 0) = c^{\text{pet}}(t_0)$. Starting from this initial distribution of tau, Bayesian inference allows us to find the parameters that, when inserted into the model, minimize the difference between the model predictions $c^{\text{sim}}(t_{\text{i}})$ and the longitudinal tau PET data $c^{\text{pet}}(t_{\text{i}})$ for each subject. The timepoints t_{i} ($i = 1, ..., M$) for model evaluation are dictated by the timepoints of PET scan acquisition, with the number of follow-up scans M ranging from two to four depending on data availability for each respective subject. To define the prior distributions for our Bayesian inference, we employ the hierarchical structure illustrated in **Figure 3**. The hierarchical approach allows us to gain personalized posterior distributions while taking into account commonalities between subjects (Gelman and Hill, 2006). Specifically, we assume that the personalized diffusion coefficient κ^{s} is represented by a normal distribution bounded to positive values. Additionally, we propose that the hyperparameters μ^{κ} and σ^{κ}, representing mean and standard deviation of this bounded normal distribution, are drawn from common hyperdistributions for all subjects. To account for potential deviations in pathology based on the subjects' amyloid status, we assume that the personalized production rate, $\alpha^{\text{s}}_{\text{A}\beta+}$ or $\alpha^{\text{s}}_{\text{A}\beta-}$, is drawn from a different normal distribution depending on whether the subject has been identified to be amyloid positive or negative. To account for similarities across subjects within one amyloid status group, we postulate that the hyperparameters $\mu^{\alpha}_{\text{A}\beta+}$ and $\sigma^{\alpha}_{\text{A}\beta+}$ are drawn from common hyperdistributions for all amyloid positive subjects, while the hyperparameters $\mu^{\alpha}_{\text{A}\beta-}$ and $\sigma^{\alpha}_{\text{A}\beta-}$ are drawn from common hyperdistributions for all amyloid negative subjects.

We postulate that the likelihood between the time-dependent PET imaging data $\hat{D}(t)$ and our model predictions $D(t, \boldsymbol{\vartheta}, \boldsymbol{\varphi})$ is normally distributed around the modeled values with a width of σ^{err}.

$$p(\hat{D}(t)|\boldsymbol{\vartheta}, \boldsymbol{\varphi}) \sim \text{Normal}(\text{mean} = D(t, \boldsymbol{\vartheta}, \boldsymbol{\varphi}), \text{width} = \sigma^{\text{err}}). \tag{5}$$

To complete our statistical model in a Bayesian setting, we select weakly informative priors for our set of model parameters $\boldsymbol{\vartheta} = \{\kappa^{\text{s}}, \alpha^{\text{s}}_{\text{A}\beta+}, \alpha^{\text{s}}_{\text{A}\beta-}\}$ and our set of hyperparameters $\boldsymbol{\varphi} = \{\mu^{\kappa}, \sigma^{\kappa}, \mu^{\alpha}_{\text{A}\beta+}, \sigma^{\alpha}_{\text{A}\beta+}, \mu^{\alpha}_{\text{A}\beta-}, \sigma^{\alpha}_{\text{A}\beta-}\}$ as summarized in **Table 1**.

Finally, we compute the posterior distributions $p(\boldsymbol{\vartheta}, \boldsymbol{\varphi}|\hat{D}(t))$ for the model parameters $\boldsymbol{\vartheta}$ and hyperparameters $\boldsymbol{\varphi}$ using Bayes' theorem,

$$p(\boldsymbol{\vartheta}, \boldsymbol{\varphi}|\hat{D}(t)) = \frac{p(\hat{D}(t)|\boldsymbol{\vartheta}, \boldsymbol{\varphi})\, p(\boldsymbol{\vartheta}, \boldsymbol{\varphi})}{p(\hat{D}(t))}, \tag{6}$$

with $p(\boldsymbol{\vartheta})$ denoting the prior distributions from **Table 1**. Since we cannot solve for the posterior distributions analytically, we adopt approximate-inference techniques to calibrate our model to the imaging data. Specifically, we use the No-U-Turn sampler (NUTS) (Hoffman and Gelman, 2014), a type of Hamiltonian Monte Carlo algorithm implemented in the python package PyMC3 (Salvatier et al., 2016) to numerically approximate the posterior distributions. We run two chains with 1,600 tuning samples and 2,000 post-tuning samples each. After convergence of the posterior distributions, we draw 4,000 posterior predictive samples of different parameter combinations which allow us to quantify the uncertainty on the inferred parameters. Additionally, we sample from the posterior distributions to predict the evolution of tau in three brain regions of interest in 35 subjects with a positive production rate. Specifically, we predict how the tau concentration is projected to change over the next 30 years in the entorhinal cortex (EC), the middle temporal gyrus (MTG) and the superior temporal gyrus (STG). Post mortem histological studies have shown that these regions are affected by hyperphosphorylated tau and neurofibrillary tangles at different disease stages, the entorhinal cortex falling into Braak stage II, the middle temporal gyrus into Braak stage IV and the superior temporal gyrus into Braak stage V (Braak et al., 2006). By propagating the uncertainty from the parameter inference through the posterior predictions, we gain an ensemble of forecasts enabling us to determine the credible intervals around our predictions.

TABLE 1 | Hierarchical Bayesian inference.

Parameter	Distribution
μ^{κ}	BoundNormal(> 0,1,20)
σ^{κ}	HalfCauchy($\beta = 1$)
κ^{s}	BoundNormal(> 0,μ^{κ},σ^{κ})
$\mu^{\alpha}_{A\beta+}$	Normal(0,2)
$\sigma^{\alpha}_{A\beta+}$	HalfCauchy($\beta = 1$)
$\alpha^{s}_{A\beta+}$	Normal($\mu^{\alpha}_{A\beta+}$,$\sigma^{\alpha}_{A\beta+}$)
$\mu^{\alpha}_{A\beta-}$	Normal(0,2)
$\sigma^{\alpha}_{A\beta-}$	HalfCauchy($\beta = 1$)
$\alpha^{s}_{A\beta-}$	Normal($\mu^{\alpha}_{A\beta-}$,$\sigma^{\alpha}_{A\beta-}$)
σ^{err}	HalfCauchy($\beta = 1$)

Prior distributions for the personalized diffusion coefficient and its hyperparameters, the personalized production rate and its hyperparameters, and the width of the likelihood.

3. RESULTS

3.1. Posterior Distributions

Figure 4 shows the posterior distribution density plots for the personalized model parameters κ^{s}, $\alpha^{s}_{A\beta+}$ and $\alpha^{s}_{A\beta-}$, as well as for the hyperparameters μ^{κ}, $\mu^{\alpha}_{A\beta+}$ and $\mu^{\alpha}_{A\beta-}$. The personalized diffusion coefficient, characterizing how fast tau spreads along a single connection between two regions, is physically constrained to be positive. We found that this parameter takes on values of up to 4.38 μm/yr. Across all subjects we identified an average diffusion coefficient of 1.304 ± 0.69 μm/yr. The protein production rate can take on positive or negative values, depending on whether clearance or production of pathological protein dominate in a particular subject. Both amyloid groups contain subjects with positive and

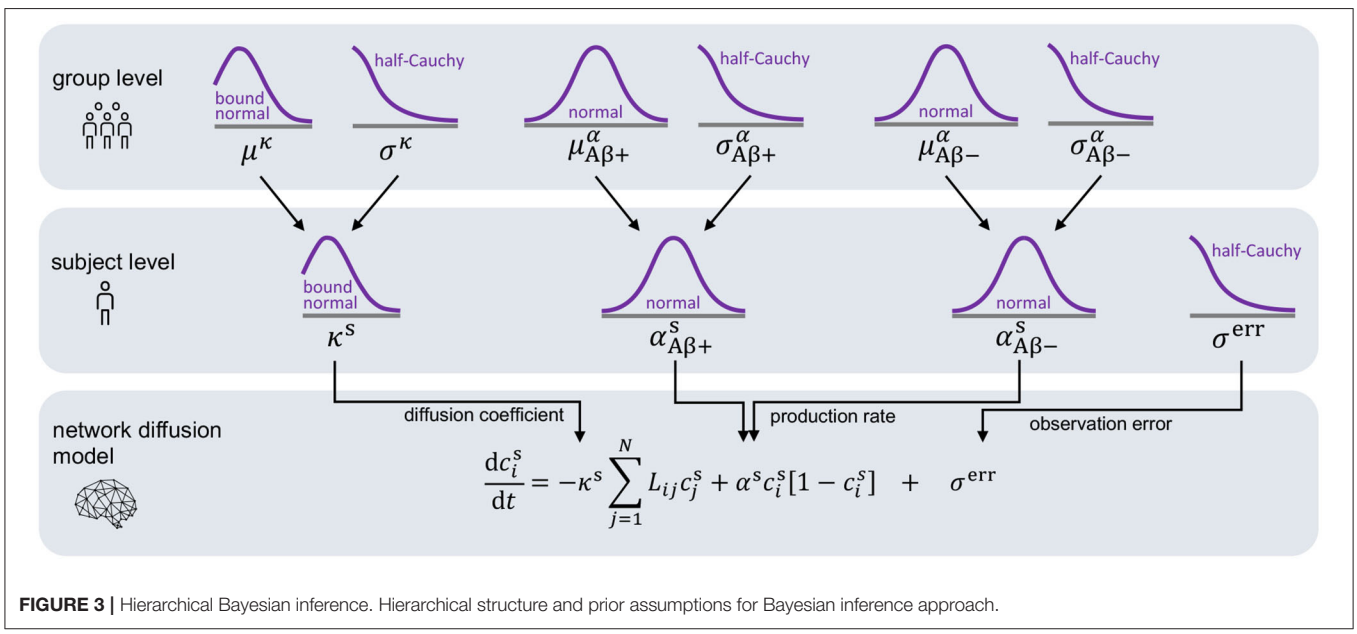


FIGURE 3 | Hierarchical Bayesian inference. Hierarchical structure and prior assumptions for Bayesian inference approach.

subjects with negative production rates. However, the density plots of the hyperparameters show that there is a noticeable difference in the group-level mean production rate depending on amyloid status. Subjects with negative amyloid status tend to exhibit a lower protein production rate than subjects with positive amyloid status. We identified an average production rate of -0.143 ± 0.21/yr across all amyloid negative subjects and 0.019 ± 0.27/yr across all amyloid positive subjects. **Table 2** summarizes the mean, maximum, and minimum inferred values for all personalized model parameters.

The boxplot in **Figure 5** further illustrates the effect of amyloid status on the inferred personalized production rate. When comparing the average production rates associated with amyloid positive and negative groups in an independent *t*-test, we found that the difference is significant with $p = 0.0075$. While there are some outliers toward negative values in the amyloid positive group, overall the production rate associated with amyloid positive subjects is significantly higher than the production rate associated with amyloid negative subjects. Our results did not show any significant and consistent trends in diffusion coefficients or production rates associated with different clinical diagnoses, e.g., cognitively normal, mild cognitive impairment, or Alzheimer's disease (see **Supplementary Figure 1**).

TABLE 2 | Posterior distributions.

Parameter	Diffusion coefficient κ^s [μm/yr]	Production rate $\alpha^s_{A\beta-}$ [1/yr]	Production rate $\alpha^s_{A\beta+}$ [1/yr]
Mean	1.304	−0.143	0.019
Std	±0.69	±0.21	±0.27
Min	0.15	−0.49	−1.01
Max	4.38	0.27	0.44

Mean values, standard deviations, maximum and minimum values for personalized model parameters.

3.2. Posterior Predictive Modeling

Posterior predictive modeling allows us to propagate the uncertainty from the Bayesian inference process through the model and illustrate its impact on model predictions. **Figures 6–8** show our projections for tau evolution over 30 years after

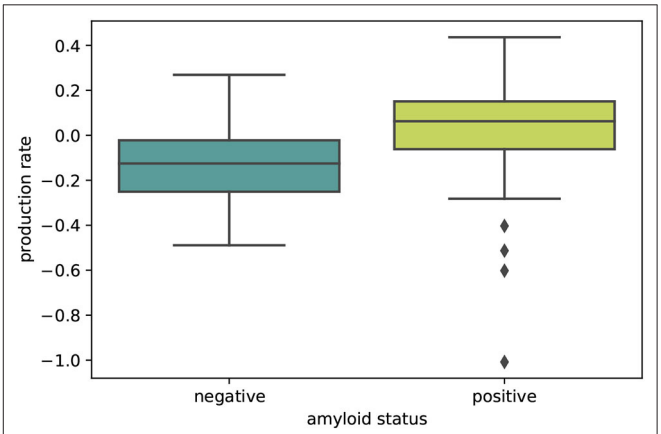


FIGURE 5 | Amyloid status. Boxplot illustrating the distributions of personalized production rates in amyloid negative and amyloid positive subject groups. The difference between the two groups is significant with $p = 0.0075$.

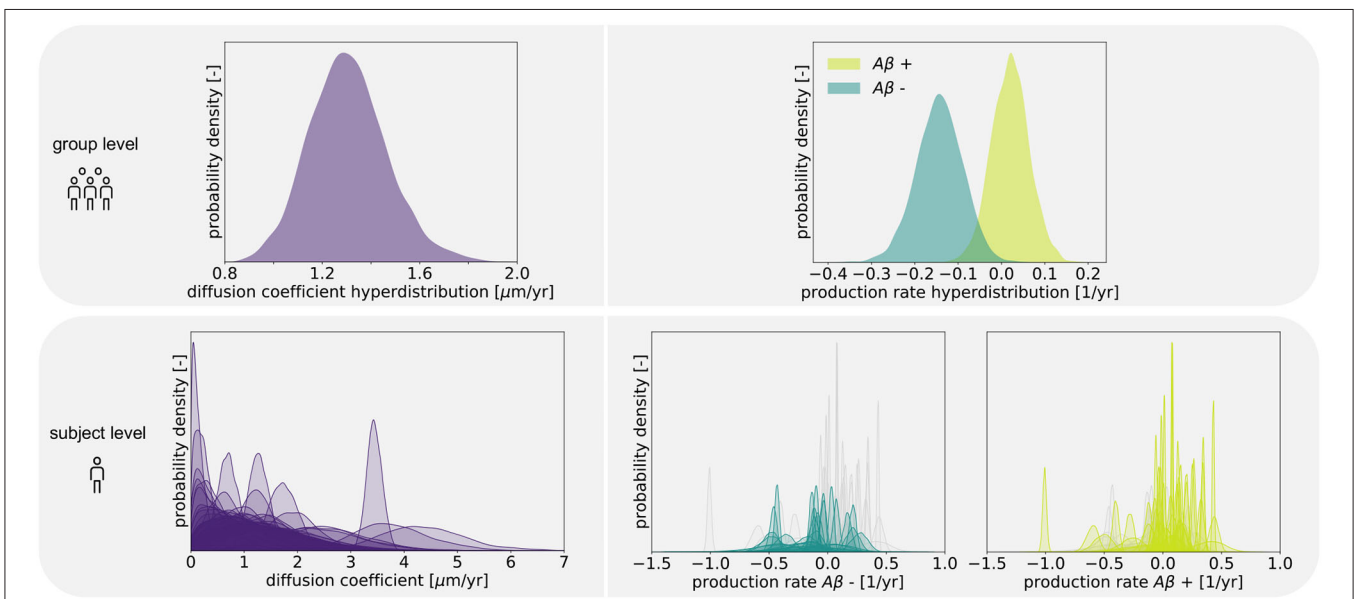


FIGURE 4 | Posterior distributions. Posterior distributions for diffusion coefficient and protein production rate on group and subject levels. Subject-wise distributions for the production rate are depicted in light and dark green in separate plots based on amyloid status. In each of those plots, the individual distributions associated with the other amyloid status group are depicted in light gray for comparison.

the first PET scan in 35 subjects and three different brain regions. The sigmoid like shape of the curves is characteristic for the combined diffusion production equation we use to model the spread of pathological tau and local conversion from healthy to unhealthy proteins. The shaded area around the curves represents the 95% credible interval, quantifying the uncertainty in our predictions as established by the probabilistic approach. Narrow credible intervals indicate high confidence in our predictions. The curves are fairly symmetrical across left and right hemisphere. When comparing the predictions for different subjects within entorhinal cortex (EC), middle temporal gyrus (MTG) and superior temporal gyrus (STG), we can identify a number of subjects for which the credible interval is narrow, confirming high certainty for our predictions. Specifically, there are seven subjects for whom the credible interval does not exceed a width of 0.2 over 30 years in any of the three examined brain regions. For multiple other subjects however, the credible interval is rapidly widening after only a few years. For these subjects, the available imaging data did not yet contain enough information to confidently infer personalized model parameters with our probabilistic approach. In those instances, additional data from PET scans at future time points may improve the prediction certainty. The vertical gray lines in **Figures 6–8** indicate the year at which the width of the credible interval exceeds a critical threshold of 0.2. If the goal is to collect additional data to increase confidence in our projections, these time points would be reasonable choices for additional scans. The value of 0.2 was chosen arbitrarily to illustrate how our uncertainty predictions can inform future study design if there is a known confidence requirement for predictions.

4. DISCUSSION

In this study we used a probabilistic approach based on hierarchical modeling and Bayesian inference to identify personalized model parameters of a physics-based network diffusion model for misfolded tau propagation. We calibrated our model to longitudinal tau PET data of 76 subjects and created personalized predictions for disease progression over a course of 30 years. Propagating the uncertainty from our parameter search through the posterior predictions allowed us to determine the credibility associated with our predictions for different subjects and brain regions.

We based the structure of our hierarchical model on the assumption that the protein production rate, summarizing the process of healthy protein production, protein clearance and conversion from healthy to misfolded protein, may vary between amyloid positive and amyloid negative subjects. This assumption makes sense in light of the amyloid hypothesis, which identifies amyloid pathology as the primary hallmark of Alzheimer's disease. In fact, tau pathology has been found in medial temporal limbic areas before the appearance of any amyloid plaques (Braak and Del Tredici, 2011). However, these

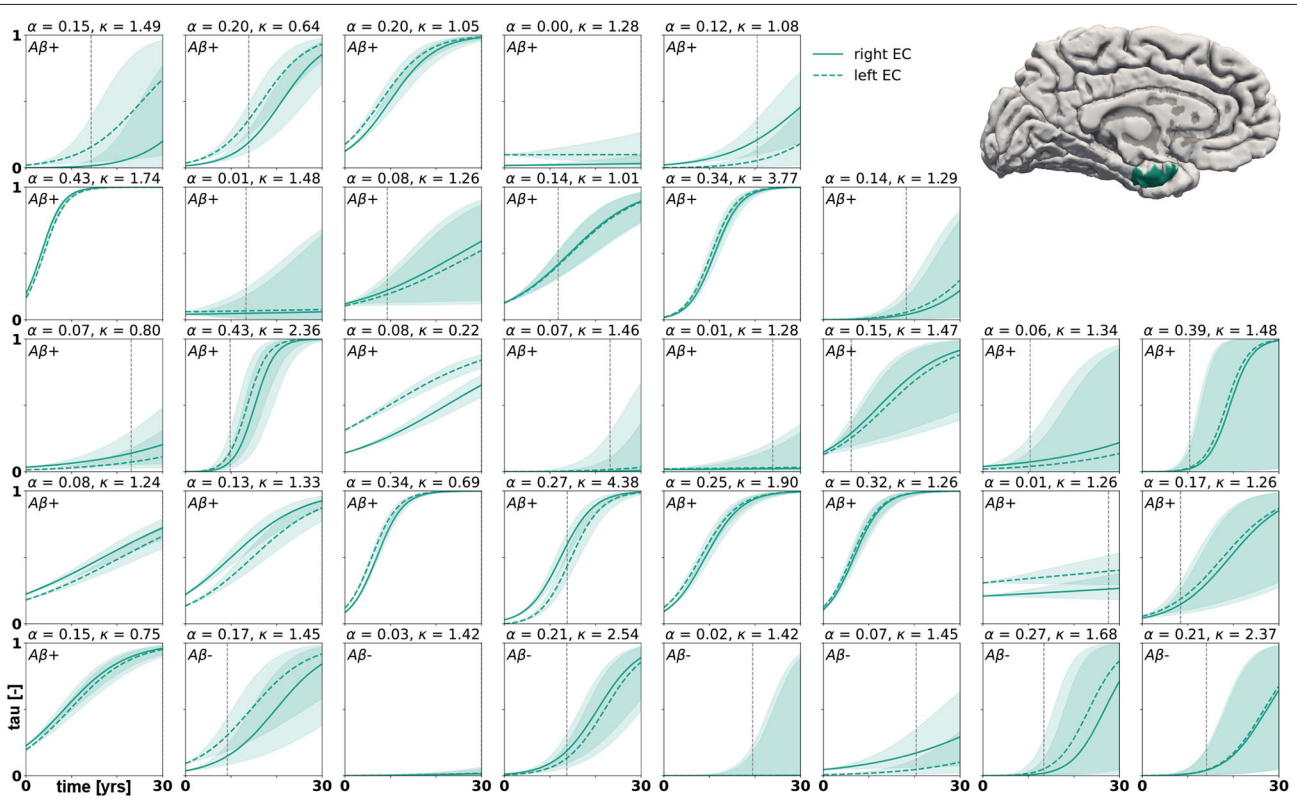


FIGURE 6 | Posterior predictive modeling. Predictions for the change in tau in the entorhinal cortex in 30 years after first tau PET scan for 35 subjects. Shaded areas around the curves represent the 95% credible intervals of the predictions.

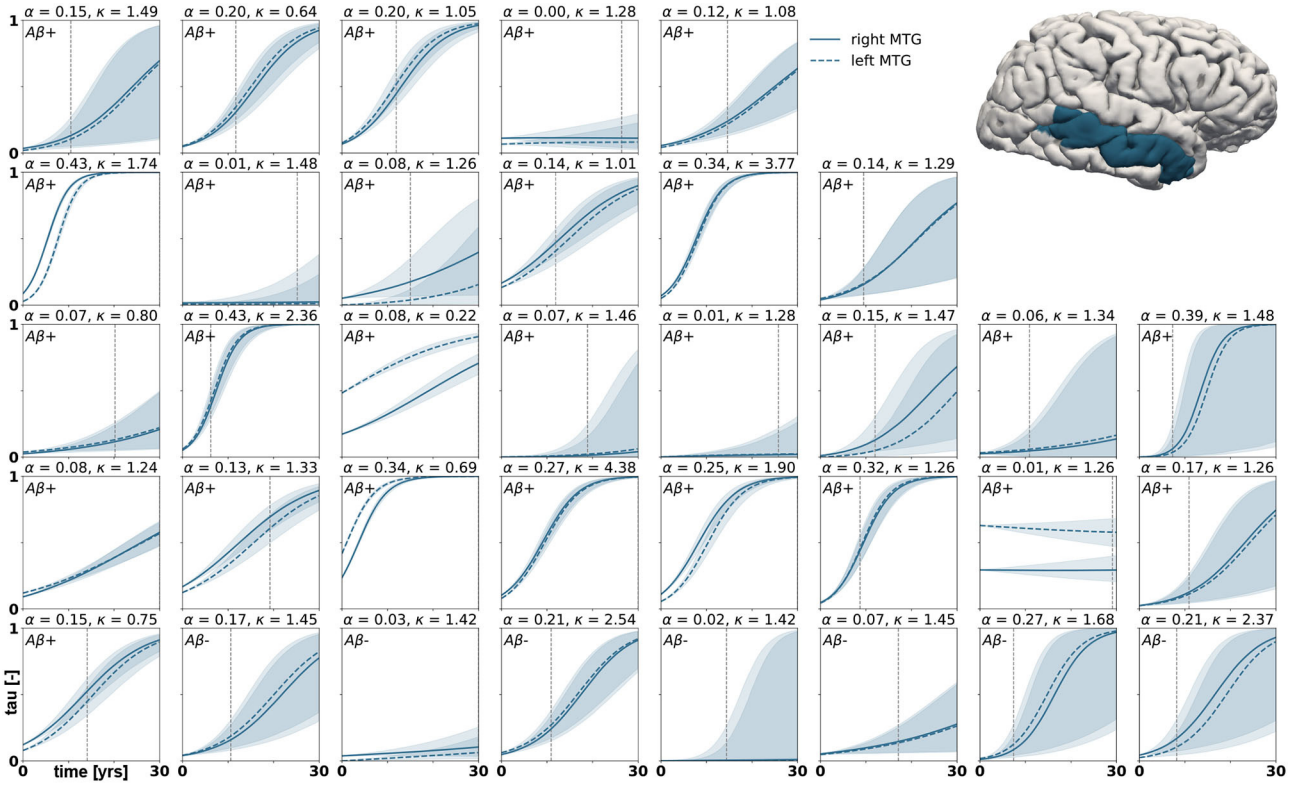


FIGURE 7 | Posterior predictive modeling. Predictions for the change in tau in the middle temporal gyrus in 30 years after first tau PET scan for 35 subjects. Shaded areas around the curves represent the 95% credible intervals of the predictions.

early tau accumulations are usually so small that they can only be detected by immunostaining methods and are rather related to normal aging than to Alzheimer's disease. It has been suggested that, independent from previously existing minor tauopathy, amyloid pathology intensifies and accelerates any existing tauopathy through currently unknown mechanisms (Price and Morris, 1999; Jack et al., 2013). This may allow hyperphosphorylated tau to spread widely across the neocortex (Musiek and Holtzman, 2012; Kevrekidis et al., 2020; Thompson et al., 2020). Our results support the hypothesis that amyloid pathology is a driver for tau pathology. Even though the structure of our probabilistic model does not enforce a difference between the production rate hyperdistributions for amyloid positive and negative groups, we found that misfolded tau production rates were significantly higher in amyloid positive subjects than in amyloid negative subjects.

Across all examined subjects, we identified an average tau diffusion rate of 1.304 μm/yr. *In vivo* experiments in mice determined that healthy tau proteins move as part of the slow component of axonal transport at 0.2–0.4 mm/day in retinal ganglion cell axons (Mercken et al., 1995). There seems to be a strong disconnect between the slow time scale of tau pathology evolution in Alzheimer's disease, which is known to typically stretch over more than a decade (Bateman et al., 2012), and the fast time scale of axonal transport. If misfolded tau spread in the brain at the speed measured for healthy tau, it would easily contaminate the whole brain in just a few months. This scenario appears inconsistent with the slow propagation and discrete stageing of neurofibrillary tangles and neuropathology that has been observed in histopathological (Braak and Braak, 1991) and imaging studies (Jack et al., 2018).

A possible explanation for this discrepancy in time scales could be related to how protein diffusion and production contribute to tau pathology to varying extents depending on the stage of disease. The model we use here to describe the propagation of tau pathology in the brain is based on the hypothesis that misfolded tau contaminates the brain in a prion-like fashion (Jucker and Walker, 2011; Fornari et al., 2020). We assume that hyperphosphorylated tau proteins act as proteopathic seeds that can travel along the axon, leave the cell and be taken up into previously unaffected neurons (Clavaguera et al., 2009; Liu et al., 2012). Additionally, we adopt the hypothesis that misfolded tau seeds replicate and aggregate locally (Iba et al., 2013). Once the chain reaction consisting of the spread of proteopathic seeds and the local multiplication of seeds is initiated, it results in an overall increase of misfolded tau across the brain. However, it is difficult, if not impossible, to determine experimentally which of the two components, diffusion or local production, dominates during different stages of disease.

A recent study aimed at quantifying the chemical kinetics of tau replication and spreading from several modalities of data, including seed amplification assays, histopathology, and tau PET, found that protein replication, not spreading, is the dominant and limiting component of tau accumulation after a

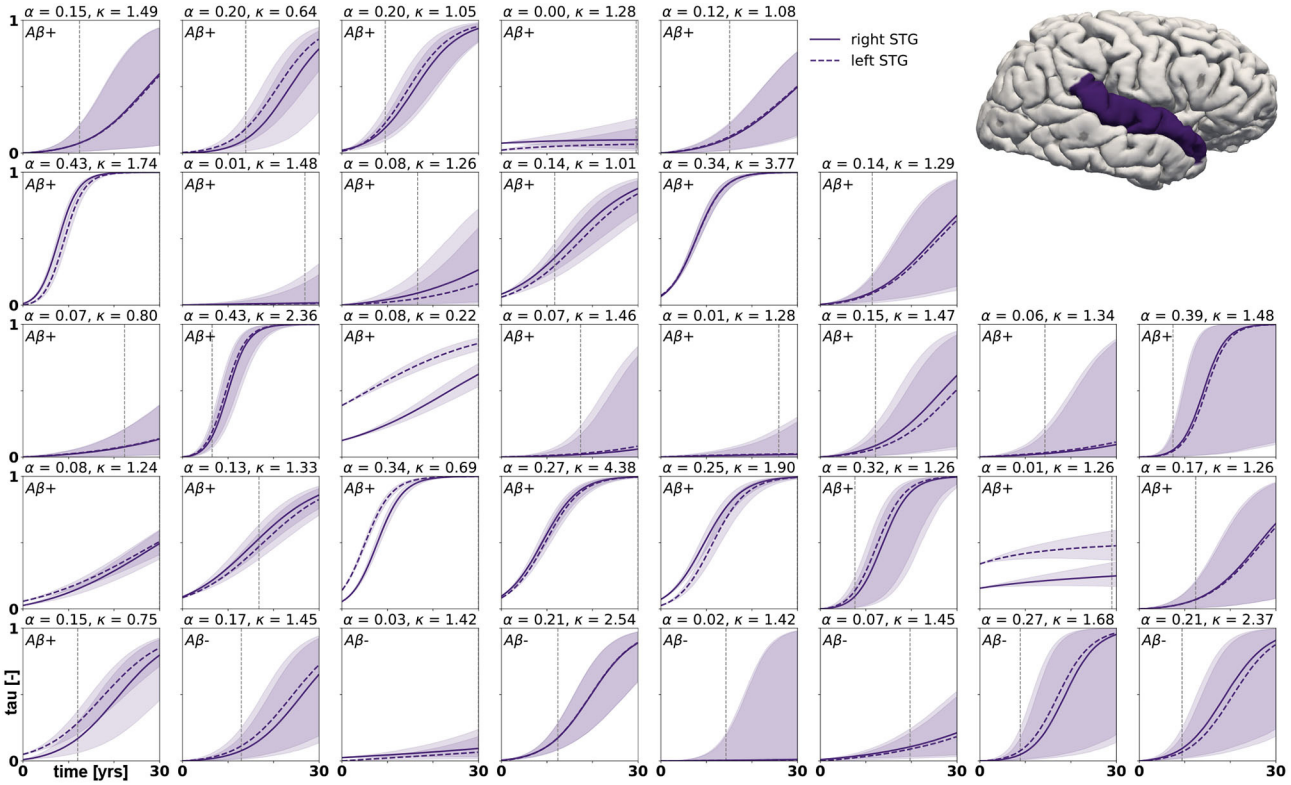


FIGURE 8 | Posterior predictive modeling. Predictions for the change in tau in the superior temporal gyrus in 30 years after first tau PET scan for 35 subjects. Shaded areas around the curves represent the 95% credible intervals of the predictions.

certain stage of disease (Meisl et al., 2020). The authors argue that misfolded tau seeds spread very fast early in the disease process, consistent with the fast axonal transport rates known for other proteins. After this initial fast spread, small but significant amounts of proteopathic seeds are already present in numerous brain regions, and further kinetics are largely determined by the local replication and aggregation of those seeds. These findings are consistent with our results, which indicate very low diffusion coefficients. Unless our data were capturing the very beginning of tau pathology, we will not be able to infer the fast transport rates that might determine disease progression initially. Since we use the regional tau distribution from each subject's baseline PET scan instead of an artificial seeding approach, it is common that small amounts of tau are already measured in a majority of the brain regions. The protein production rates we identified for our subject sample are comparable to the average replication rate of 0.14/yr reported in the study above (Meisl et al., 2020).

We computed personalized 30-year predictions of tau evolution in three different brain areas, the entorhinal cortex, the middle temporal gyrus and the superior temporal gyrus. Since there is a well established correlation between tau distribution and neurodegeneration, these predictions not only contain information about the amount of protein in these areas, but also provide important clinical insight into when certain brain functions might be affected. A recently conducted study compared tau PET distribution at baseline visit to the amount and distribution of atrophy detected between baseline and follow-up visit (La Joie et al., 2020). It was found that tau is a strong predictor for regional atrophy presenting around 15 months after the PET scan. In the healthy adult brain, the neurons in the entorhinal cortex provide a number of functionalities, but are mainly thought to be responsible for spatial memory and spatial association tasks (Kerr et al., 2007; Van Strien et al., 2009; Kuruvilla and Ainge, 2017). If this area atrophies after serious invasion of misfolded tau protein, these functions may be impaired or lost. The middle temporal gyrus, which is part of the inferior temporal lobe, has been suggested to play a central role in visual learning and memory (Buckley et al., 1997) and lesions in this region may cause object and face recognition deficits (Purves et al., 2001). The superior temporal gyrus contains the auditory cortex and is involved in speech and auditory processing (Gernsbacher and Kaschak, 2003) as well as social cognition (Adolphs, 2003; Bigler et al., 2007). The quantified uncertainty on disease progression showcased in **Figures 6**–**8** indicates the credibility associated with each subject's prediction specifically, taking into account the behavior of the whole cohort. This framework may provide an interesting tool for clinical prognosis, informing clinical practitioners and caregivers when cognitive symptoms related to loss of the functions above maybe be expected in a certain patient. It may also provide a new means to assess the optimal time for a follow-up scan, smartly maximizing prognosis credibility while minimizing the number of scans performed.

This study comes with a number of limitations. First, the amount of longitudinal tau PET data available today is limited. The small sample size and limited follow-up data included in our study result in large credible intervals and reduced confidence in the model parameters. As tau PET becomes a more established technique in longitudinal imaging studies over the next years, more data will naturally become available, allowing us to constantly improve our hierarchical model, as Bayesian methods are inherently tailored to analyzing data that are continuously updated in time. Adding more subjects to our data set will further increase the learning effect we achieve through the hierarchical structure, which will in turn increase the credibility of all personalized predictions. Larger sample sizes of data in the future will also allow us to explore more complex models, e.g., models introducing regionally varying protein production rates based on local gene expression (Grothe et al., 2018), without the risk of overfitting. Second, we use the same anatomical brain network to compute tau spreading in all subjects. This network was extracted from averaged diffusion tensor images of over 400 brains. In reality, the connectivity is different in every brain, potentially affecting the diffusion dynamics observed here. We attempt to surmount this issue by introducing the diffusion coefficient as a personalized parameter. It would be reasonable to assume that the transport rate of misfolded tau along the axon is a biological parameter that is similar in all brains. However, by allowing this parameter to vary between subjects, we provide the option to scale the adjacency matrix and thereby introduce variations in connection strength for the otherwise non-personalized network. In the future, we plan to surpass the potential over-generalization that using an average network introduces by extracting personalized connectomes for each subject from diffusion tensor images when available. Another consideration for future studies is to correct the weighting of connections in our network for the varying surface areas between brain regions. Additionally, it could be of interest to compare the performance of our model on the structural connectome with its performance on other reference networks, and thus test the hypothesis that misfolded tau spreads along the axonal network. Third, since the ADNI data base only provides one tau PET scan for each time point, this study does not explicitly take into account uncertainties arising from imaging protocols. However, since we expect this error to be Gaussian, it is partially accounted for by the stochastic nature of the observation error in our Bayesian inference framework. In contrast to deterministic optimization algorithms, our probabilistic approach inherently accounts for observation errors through the likelihood width.

The approach we used here is optimal for understanding the applicability of the physics-based network diffusion model to longitudinal brain imaging data and for quantifying the range of model parameters presented in this data. Additionally, the Bayesian inference framework inherently provides information on the uncertainty in our model, intrinsically informing us on model applicability. In the future, we will expand our model to a more predictive approach using a combination of deep learning, Bayesian inference, and physics-based modeling, with the goal to create personalized predictions of tau spreading dynamics from a single baseline PET scan. Furthermore, we will explore a coupled model of tau pathology and resulting tissue atrophy (Schäfer et al., 2019) calibrated to longitudinal tau PET and structural MRI.

5. CONCLUSION

We presented a probabilistic approach to calibrate the parameters of a physics-based network diffusion model to longitudinal tau PET data. We obtained posterior probability distributions for two personalized model parameters, the diffusion coefficient and the protein production rate, using Bayesian inference combined with a hierarchical prior structure. This approach allowed us to identify the characteristics of tau propagation for each individual subject while taking into account expected commonalities between subjects. We inferred an average diffusion coefficient of 1.304 ± 0.69 μm/yr, a protein production rate of 0.019 ± 0.27/yr for the amyloid positive group, and a production rate of -0.143 ± 0.21/yr for the amyloid negative group. The significantly higher tau production rate associated with the presence of amyloid-β supports the hypothesis that amyloid pathology drives tau pathology. The small magnitude of our inferred diffusion coefficients is inconsistent with experimentally identified axonal transport rates for healthy tau, but consistent with the slow disease progression known for Alzheimer's disease. Extrapolating our model based on the posterior distributions of model parameters allowed us to create personalized predictions of tau evolution in three brain regions associated with distinct cognitive functions. These predicitons and associated credibility intervals may serve as a tool to estimate the timeline of regional tau pathology and function-specific cognitive impairment in individual patients. Our findings could serve as simulated controls in therapeutic trials or as a means to smartly schedule follow-up PET scans that most benefit model prediction certainty.

DATA AVAILABILITY STATEMENT

Publicly available datasets were analyzed in this study. This data can be found here: http://adni.loni.usc.edu/.

AUTHOR CONTRIBUTIONS

AS was responsible for conception and design of the study, implementation of the model, analysis and interpretation of results, and draft of the manuscript. MP, KL, and EK contributed to and guided study conception and design and provided critical revision of the manuscript for intellectual content. All authors approved the final version of the article to be published.

FUNDING

This work was supported by a Brit and Alex d'Arbeloff Stanford Graduate Fellowship to AS, a Belgian American Educational Foundation (B.A.E.F.) Postdoctoral Research Fellowship to MP, a DAAD Fellowship to KL, and a National Science Foundation Grant CMMI 1727268 and BioX-IIP Seed Grant to EK.

Data collection and sharing for this project was funded by the Alzheimer's Disease Neuroimaging Initiative (ADNI) (National Institutes of Health Grant U01 AG024904) and DOD ADNI (Department of Defense award number W81XWH-12-2-0012). ADNI is funded by the National Institute on Aging, the National Institute of Biomedical Imaging and Bioengineering, and through generous contributions from the following: AbbVie, Alzheimer's Association; Alzheimer's Drug Discovery Foundation; Araclon Biotech; BioClinica, Inc.; Biogen; Bristol-Myers Squibb Company; CereSpir, Inc.; Cogstate; Eisai Inc.; Elan Pharmaceuticals, Inc.; Eli Lilly and Company; EuroImmun; F. Hoffmann-La Roche Ltd and its affiliated company Genentech, Inc.; Fujirebio; GE Healthcare; IXICO Ltd.; Janssen Alzheimer Immunotherapy Research & Development, LLC.; Johnson & Johnson Pharmaceutical Research & Development LLC.; Lumosity; Lundbeck; Merck & Co., Inc.; Meso Scale Diagnostics, LLC.; NeuroRx Research; Neurotrack Technologies; Novartis Pharmaceuticals Corporation; Pfizer Inc.; Piramal Imaging; Servier; Takeda Pharmaceutical Company; and Transition Therapeutics. The Canadian Institutes of Health Research is providing funds to support ADNI clinical sites in Canada. Private sector contributions are facilitated by the Foundation for the National Institutes of Health (www.fnih.org). The grantee organization is the Northern California Institute for Research and Education, and the study is coordinated by the Alzheimer's Therapeutic Research Institute at the University of Southern California. ADNI data are disseminated by the Laboratory for Neuro Imaging at the University of Southern California.

SUPPLEMENTARY MATERIAL

The Supplementary Material for this article can be found online at: https://www.frontiersin.org/articles/10.3389/fphys.2021.702975/full#supplementary-material

REFERENCES

ADNI (2020). *Alzheimer's Disease Neuroimaging Initiative*. Available online at: http://adni.loni.usc.edu.

Adolphs, R. (2003). Cognitive neuroscience of human social behaviour. *Nat. Rev. Neurosci*. 4, 165–178. doi: 10.1038/nrn1056

Association, A. (2019). 2019 Alzheimer's disease facts and figures. *Alzheimers Dement*. 15, 321–387. doi: 10.1016/j.jalz.2019.01.010

Bateman, R. J., Xiong, C., Benzinger, T. L., Fagan, A. M., Goate, A., Fox, N. C., et al. (2012). Clinical and biomarker changes in dominantly inherited Alzheimer's disease. *New Engl. J. Med*. 367, 795–804. doi: 10.1056/NEJMoa1202753

Bigler, E. D., Mortensen, S., Neeley, E. S., Ozonoff, S., Krasny, L., Johnson, M., et al. (2007). Superior temporal gyrus, language function, and autism. *Dev. Neuropsychol*. 31, 217–238. doi: 10.1080/87565640701190841

Braak, H., Alafuzoff, I., Arzberger, T., Kretzschmar, H., and Del Tredici, K. (2006). Staging of Alzheimer disease-associated neurofibrillary pathology using paraffin sections and immunocytochemistry. *Acta Neuropathol*. 112, 389–404. doi: 10.1007/s00401-006-0127-z

Braak, H., and Braak, E. (1991). Neuropathological stageing of Alzheimer-related changes. *Acta Neuropathol*. 82, 239–259. doi: 10.1007/BF00308809

Braak, H., and Del Tredici, K. (2011). The pathological process underlying alzheimer's disease in individuals under thirty. *Acta Neuropathol*. 121, 171–181. doi: 10.1007/s00401-010-0789-4

Buckley, M., Gaffan, D., and Murray, E. (1997). Functional double dissociation between two inferior temporal cortical areas: perirhinal cortex versus middle temporal gyrus. *J. Neurophysiol*. 77, 587–598. doi: 10.1152/jn.1997.77.2.587

Clavaguera, F., Bolmont, T., Crowther, R. A., Abramowski, D., Frank, S., Probst, A., et al. (2009). Transmission and spreading of tauopathy in transgenic mouse brain. *Nat. Cell Biol*. 11, 909–913. doi: 10.1038/ncb1901

Congdon, E. E., and Sigurdsson, E. M. (2018). Tau-targeting therapies for alzheimer disease. *Nat. Rev. Neurol*. 14, 399–415. doi: 10.1038/s41582-018-0013-z

De Calignon, A., Polydoro, M., Suárez-Calvet, M., William, C., Adamowicz, D. H., Kopeikina, K. J., et al. (2012). Propagation of tau pathology in a model of early Alzheimer's disease. *Neuron* 73, 685–697. doi: 10.1016/j.neuron.2011.11.033

Desikan, R. S., Ségonne, F., Fischl, B., Quinn, B. T., Dickerson, B. C., Blacker, D., et al. (2006). An automated labeling system for subdividing the human cerebral cortex on MRI scans into gyral based regions of interest. *Neuroimage* 31, 968–980. doi: 10.1016/j.neuroimage.2006.01.021

Duyckaerts, C., Delatour, B., and Potier, M.-C. (2009). Classification and basic pathology of Alzheimer disease. *Acta Neuropathol*. 118, 5–36. doi: 10.1007/s00401-009-0532-1

Fisher, R. A. (1937). The wave of advance of advantageous genes. *Ann. Eugen*. 7, 355–369. doi: 10.1111/j.1469-1809.1937.tb02153.x

Fornari, S., Schäfer, A., Jucker, M., Goriely, A., and Kuhl, E. (2019). Prion-like spreading of Alzheimer's disease within the brain's connectome. *J. R. Soc. Interface* 16, 20190356. doi: 10.1098/rsif.2019.0356

Fornari, S., Schäfer, A., Kuhl, E., and Goriely, A. (2020). Spatially-extended nucleation-aggregation-fragmentation models for the dynamics of prion-like neurodegenerative protein-spreading in the brain and its connectome. *J. Theor. Biol*. 486:110102. doi: 10.1016/j.jtbi.2019.110102

FreeSurfer (2020). *FreeSurfer Software Suite*. Available online at: http://surfer.nmr.mgh.harvard.edu.

Garbarino, S., Lorenzi, M., Initiative, A. D. N., et al. (2021). Investigating hypotheses of neurodegeneration by learning dynamical systems of protein propagation in the brain. *Neuroimage* 235:117980. doi: 10.1016/j.neuroimage.2021.117980

Gelman, A., and Hill, J. (2006). *Data Analysis Using Regression and Multilevel/Hierarchical Models*. Cambridge; New York, NY: Cambridge University Press.

Gernsbacher, M. A., and Kaschak, M. P. (2003). Neuroimaging studies of language production and comprehension. *Ann. Rev. Psychol*. 54, 91–114. doi: 10.1146/annurev.psych.54.101601.145128

Grothe, M. J., Sepulcre, J., Gonzalez-Escamilla, G., Jelistratova, I., Schöll, M., Hansson, O., et al. (2018). Molecular properties underlying regional vulnerability to Alzheimer's disease pathology. *Brain* 141, 2755–2771. doi: 10.1093/brain/awy189

Harrison, T. M., La Joie, R., Maass, A., Baker, S. L., Swinnerton, K., Fenton, L., et al. (2019). Longitudinal tau accumulation and atrophy in aging and Alzheimer disease. *Ann. Neurol*. 85, 229–240. doi: 10.1002/ana.25406

Hoffman, M. D., and Gelman, A. (2014). The no-u-turn sampler: adaptively setting path lengths in hamiltonian monte carlo. *J. Mach. Learn. Res*. 15, 1593–1623.

Iba, M., Guo, J. L., McBride, J. D., Zhang, B., Trojanowski, J. Q., and Lee, V. M.-Y. (2013). Synthetic tau fibrils mediate transmission of neurofibrillary tangles in a transgenic mouse model of alzheimer's-like tauopathy. *J. Neurosci*. 33, 1024–1037. doi: 10.1523/JNEUROSCI.2642-12.2013

Iturria-Medina, Y., Sotero, R. C., Toussaint, P. J., Evans, A. C., Initiative, A. D. N., et al. (2014). Epidemic spreading model to characterize misfolded proteins propagation in aging and associated neurodegenerative disorders. *PLoS Comput. Biol*. 10:e1003956. doi: 10.1371/journal.pcbi.1003956

Jack, C. R., and Holtzman, D. M. (2013). Biomarker modeling of Alzheimer's disease. *Neuron* 80, 1347–1358. doi: 10.1016/j.neuron.2013.12.003

Jack, C. R., Knopman, D. S., Jagust, W. J., Petersen, R. C., Weiner, M. W., Aisen, P. S., et al. (2013). Tracking pathophysiological processes in Alzheimer's disease: an updated hypothetical model of dynamic biomarkers. *Lancet Neurol*. 12, 207–216. doi: 10.1016/S1474-4422(12)70291-0

Jack, C. R., Wiste, H. J., Schwarz, C. G., Lowe, V. J., Senjem, M. L., Vemuri, P., et al. (2018). Longitudinal tau PET in ageing and Alzheimer's disease. *Brain* 141, 1517–1528. doi: 10.1093/brain/awy059

Johnson, K. A., Schultz, A., Betensky, R. A., Becker, J. A., Sepulcre, J., Rentz, D., et al. (2016). Tau positron emission tomographic imaging in aging and early Alzheimer disease. *Ann. Neurol.* 79, 110–119. doi: 10.1002/ana.24546

Jones, D. T., Graff-Radford, J., Lowe, V. J., Wiste, H. J., Gunter, J. L., Senjem, M. L., et al. (2017). Tau, amyloid, and cascading network failure across the Alzheimer's disease spectrum. *Cortex* 97, 143–159. doi: 10.1016/j.cortex.2017.09.018

Jucker, M., and Walker, L. C. (2011). Pathogenic protein seeding in Alzheimer disease and other neurodegenerative disorders. *Ann. Neurol.* 70, 532–540. doi: 10.1002/ana.22615

Kerr, K. M., Agster, K. L., Furtak, S. C., and Burwell, R. D. (2007). Functional neuroanatomy of the parahippocampal region: the lateral and medial entorhinal areas. *Hippocampus* 17, 697–708. doi: 10.1002/hipo.20315

Kevrekidis, P., Thompson, T. B., and Goriely, A. (2020). Anisotropic diffusion and traveling waves of toxic proteins in neurodegenerative diseases. *Phys. Lett. A* 384, 126935. doi: 10.1016/j.physleta.2020.126935

Kolmogorov, A., Petrovskii, I., and Piskunov, N. (1937). A study of the equation of diffusion with increase in the quantity of matter, and its application to a biological problem. *Byul. Moskovskogo Gos. Univ.* 1, 1–25.

Kuruvilla, M. V., and Ainge, J. A. (2017). Lateral entorhinal cortex lesions impair local spatial frameworks. *Front. Syst. Neurosci.* 11:30. doi: 10.3389/fnsys.2017.00030

La Joie, R., Visani, A. V., Baker, S. L., Brown, J. A., Bourakova, V., Cha, J., et al. (2020). Prospective longitudinal atrophy in Alzheimer's disease correlates with the intensity and topography of baseline tau-PET. *Sci. Transl. Med.* 12:eaau5732. doi: 10.1126/scitranslmed.aau5732

Landau, S. M., Lu, M., Joshi, A. D., Pontecorvo, M., Mintun, M. A., Trojanowski, J. Q., et al. (2013). Comparing positron emission tomography imaging and cerebrospinal fluid measurements of β-amyloid. *Ann. Neurol.* 74, 826–836. doi: 10.1002/ana.23908

Lemoine, L., Leuzy, A., Chiotis, K., Rodriguez-Vieitez, E., and Nordberg, A. (2018). Tau positron emission tomography imaging in tauopathies: the added hurdle of off-target binding. *Alzheimers Dementia* 232–236. doi: 10.1016/j.dadm.2018.01.007

Liu, L., Drouet, V., Wu, J. W., Witter, M. P., Small, S. A., Clelland, C., et al. (2012). Trans-synaptic spread of tau pathology *in vivo*. *PLoS ONE* 7:e31302. doi: 10.1371/journal.pone.0031302

Lowe, V. J., Curran, G., Fang, P., Liesinger, A. M., Josephs, K. A., Parisi, J. E., et al. (2016). An autoradiographic evaluation of AV-1451 Tau PET in dementia. *Acta Neuropathol. Commun.* 4, 58. doi: 10.1186/s40478-016-0315-6

Marquié, M., Normandin, M. D., Meltzer, A. C., Siao Tick Chong, M., Andrea, N. V., Antón-Fernández, A., et al. (2017). Pathological correlations of [F-18]-AV-1451 imaging in non-alzheimer tauopathies. *Ann. Neurol.* 81, 117–128. doi: 10.1002/ana.24844

McNab, J. A., Edlow, B. L., Witzel, T., Huang, S. Y., Bhat, H., Heberlein, K., et al. (2013). The Human Connectome Project and beyond: initial applications of 300 mT/m gradients. *Neuroimage* 80, 234–245. doi: 10.1016/j.neuroimage.2013.05.074

Meisl, G., Zuo, Y., Allinson, K., Rittman, T., DeVos, S., Sanchez, J. S., et al. (2020). Amplification, not spreading limits rate of tau aggregate accumulation in Alzheimer's disease. *bioRxiv*. doi: 10.1101/2020.11.16.384727

Mercken, M., Fischer, I., Kosik, K., and Nixon, R. (1995). Three distinct axonal transport rates for tau, tubulin, and other microtubule-associated proteins: evidence for dynamic interactions of tau with microtubules *in vivo*. *J. Neurosci.* 15, 8259–8267. doi: 10.1523/JNEUROSCI.15-12-08259.1995

Musiek, E. S., and Holtzman, D. M. (2012). Origins of alzheimer's disease: Reconciling csf biomarker and neuropathology data regarding the temporal sequence of aβ and tau involvement. *Curr. Opin. Neurol.* 25, 715. doi: 10.1097/WCO.0b013e32835a30f4

Peirlinck, M., Costabal, F. S., Sack, K., Choy, J., Kassab, G., Guccione, J., et al. (2019). Using machine learning to characterize heart failure across the scales. *Biomech. Model. Mechanobiol.* 18, 1987–2001. doi: 10.1007/s10237-019-01190-w

Pereira, J. B., Ossenkoppele, R., Palmqvist, S., Strandberg, T. O., Smith, R., Westman, E., et al. (2019). Amyloid and tau accumulate across distinct spatial networks and are differentially associated with brain connectivity. *eLife* 8:e50830. doi: 10.7554/eLife.50830

Price, J. L., and Morris, J. C. (1999). Tangles and plaques in nondemented aging and "preclinical" Alzheimer's disease. *Ann. Neurol.* 45, 358–368. doi: 10.1002/1531-8249(199903)45:3andlt;358::AID-ANA12andgt;3.0.CO;2-X

Purves, D., Augustine, G., and Fitzpatrick, D. E. A. (2001). *Neuroscience*. Sunderland: Sinauer Associates, Inc.

Raj, A., Kuceyeski, A., and Weiner, M. (2012). A network diffusion model of disease progression in dementia. *Neuron* 73, 1204–1215. doi: 10.1016/j.neuron.2011.12.040

Salvatier, J., Wiecki, T. V., and Fonnesbeck, C. (2016). Probabilistic programming in python using pymc3. *PeerJ Comput. Sci.* 2:e55. doi: 10.7717/peerj-cs.55

Schäfer, A., Mormino, E. C., and Kuhl, E. (2020). Network diffusion modeling explains longitudinal tau pet data. *Front. Neurosci.* 14:1370. doi: 10.3389/fnins.2020.566876

Schäfer, A., Weickenmeier, J., and Kuhl, E. (2019). The interplay of biochemical and biomechanical degeneration in Alzheimer's disease. *Comput. Methods Appl. Mech. Eng.* 352, 369–388. doi: 10.1016/j.cma.2019.04.028

Szalkai, B., Kerepesi, C., Varga, B., and Grolmusz, V. (2017). Parameterizable consensus connectomes from the human connectome project: the budapest reference connectome server v3. 0. *Cogn. Neurodyn.* 11, 113–116. doi: 10.1007/s11571-016-9407-z

Thompson, T. B., Chaggar, P., Kuhl, E., Goriely, A., and Initiative, A. D. N. (2020). Protein-protein interactions in neurodegenerative diseases: a conspiracy theory. *PLoS Comput. Biol.* 16:e1008267. doi: 10.1371/journal.pcbi.1008267

Torok, J., Maia, P. D., Powell, F., Pandya, S., and Raj, A. (2018). A method for inferring regional origins of neurodegeneration. *Brain* 141, 863–876. doi: 10.1093/brain/awx371

Van Strien, N., Cappaert, N., and Witter, M. (2009). The anatomy of memory: an interactive overview of the parahippocampal-hippocampal network. *Nat. Rev. Neurosci.* 10, 272–282. doi: 10.1038/nrn2614

Villemagne, V. L., Doré, V., Burnham, S. C., Masters, C. L., and Rowe, C. C. (2018). Imaging tau and amyloid-β proteinopathies in Alzheimer disease and other conditions. *Nat. Rev. Neurol.* 14, 225–236. doi: 10.1038/nrneurol.2018.9

Vogel, J. W., Iturria-Medina, Y., Strandberg, O. T., Smith, R., Levitis, E., Evans, A. C., et al. (2020). Spread of pathological tau proteins through communicating neurons in human Alzheimer's disease. *Nat. Commun.* 11:2612. doi: 10.1038/s41467-020-15701-2

Weickenmeier, J., Jucker, M., Goriely, A., and Kuhl, E. (2019). A physics-based model explains the prion-like features of neurodegeneration in Alzheimer's disease, Parkinson's disease, and amyotrophic lateral sclerosis. *J. Mech. Phys. Solids* 124, 264–281. doi: 10.1016/j.jmps.2018.10.013

Conflict of Interest: The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

frontiers in Physiology

ORIGINAL RESEARCH
published: 02 August 2021
doi: 10.3389/fphys.2021.700655

Extracting the Auditory Attention in a Dual-Speaker Scenario From EEG Using a Joint CNN-LSTM Model

Ivine Kuruvila[1], *Jan Muncke*[1], *Eghart Fischer*[2] *and Ulrich Hoppe*[1]*

[1] Department of Audiology, ENT-Clinic, Friedrich-Alexander-Universität Erlangen-Nürnberg (FAU), Erlangen, Germany, [2] WS Audiology, Erlangen, Germany

Human brain performs remarkably well in segregating a particular speaker from interfering ones in a multispeaker scenario. We can quantitatively evaluate the segregation capability by modeling a relationship between the speech signals present in an auditory scene, and the listener's cortical signals measured using electroencephalography (EEG). This has opened up avenues to integrate neuro-feedback into hearing aids where the device can infer user's attention and enhance the attended speaker. Commonly used algorithms to infer the auditory attention are based on linear systems theory where cues such as speech envelopes are mapped on to the EEG signals. Here, we present a joint convolutional neural network (CNN)—long short-term memory (LSTM) model to infer the auditory attention. Our joint CNN-LSTM model takes the EEG signals and the spectrogram of the multiple speakers as inputs and classifies the attention to one of the speakers. We evaluated the reliability of our network using three different datasets comprising of 61 subjects, where each subject undertook a dual-speaker experiment. The three datasets analyzed corresponded to speech stimuli presented in three different languages namely German, Danish, and Dutch. Using the proposed joint CNN-LSTM model, we obtained a median decoding accuracy of 77.2% at a trial duration of 3 s. Furthermore, we evaluated the amount of sparsity that the model can tolerate by means of magnitude pruning and found a tolerance of up to 50% sparsity without substantial loss of decoding accuracy.

Keywords: EEG, cocktail party effect, auditory attention, long short term memory networks, hearing aids, speech enhancement, speech separation, convolutional neural network

OPEN ACCESS

Edited by:
Nicole Y. K. Li-Jessen,
McGill University, Canada

Reviewed by:
Behtash Babadi,
University of Maryland, United States
Michael Georg Metzen,
McGill University, Canada

***Correspondence:**
Ulrich Hoppe
ulrich.hoppe@uk-erlangen.de

Specialty section:
This article was submitted to Computational Physiology and Medicine, a section of the journal Frontiers in Physiology

Received: *26 April 2021*
Accepted: *05 July 2021*
Published: *02 August 2021*

Citation:
Kuruvila I, Muncke J, Fischer E and Hoppe U (2021) Extracting the Auditory Attention in a Dual-Speaker Scenario From EEG Using a Joint CNN-LSTM Model. Front. Physiol. 12:700655. doi: 10.3389/fphys.2021.700655

1. INTRODUCTION

Holding a conversation in presence of multiple noise sources and interfering speakers is a task that people with normal hearing carry out exceptionally well. The inherent ability to focus the auditory attention on a particular speech signal in a complex mixture is known as the cocktail party effect (Cherry, 1953). However, an automatic machine based solution to the cocktail party problem is yet to be discovered despite the intense research for more than half a century. Such a solution is highly desirable for a plethora of applications such as human-machine interface (e.g., Amazon Alexa), automatic captioning of audio/video recordings (e.g., YouTube, Netflix), advanced hearing aids etc.

In the domain of hearing aids, people with hearing loss suffer from deteriorated speech intelligibility when listening to a particular speaker in a multispeaker scenario. Hearing aids

currently available in the market often do not provide sufficient amenity in such scenarios due to their inability to distinguish between the attended speaker and the ignored ones. Hence, additional information about the locus of attention is highly desirable. In visual domain, selective attention is explained in terms of visual object formation where an observer focuses on a certain object in a complex visual scene (Feldman, 2003). This was extended to auditory domain where it was suggested that phenomena such as cocktail party effect could be better understood using auditory object formation (Shinn-Cunningham, 2008). In other words, brain forms objects based on the multiple speakers present in an auditory scene and selects those objects belonging to a particular speaker during attentive listening (top-down or late selection). However, flexible locus of attention theory was concurrently proposed where the late selection is hypothesized to occur at low cognitive load and early selection is hypothesized to occur at high cognitive load (Vogel et al., 2005). This has inspired investigation into whether cortical signals could provide additional information that helps to discriminate between the attended speaker and interfering speakers. In a dual-speaker experiment, it was observed that the cortical signals measured using implanted electrodes track the salient features of the attended speaker stronger than the ignored speaker (Mesgarani and Chang, 2012). Similar results were obtained using magnetoencephalography and electroencephalography (EEG) (Ding and Simon, 2012; O'Sullivan et al., 2014). In recent years, EEG analyses have become the commonly used methodology in attention research which is lately known as auditory attention decoding (AAD).

Both low level acoustic features (speech envelope or speech spectrogram) and high level features (phonemes or phonetics) have been used to investigate the speech tracking in cortical signals (Aiken and Picton, 2008; Lalor and Foxe, 2010; Di Liberto et al., 2015; Broderick et al., 2019). State-of-the-art AAD algorithms are based on linear systems theory where acoustic features are linearly mapped on to the EEG signals. This mapping can be either in the forward direction (Lalor and Foxe, 2010; Fiedler et al., 2017; Kuruvila et al., 2020) or in the backward direction (O'Sullivan et al., 2014; Mirkovic et al., 2015; Biesmans et al., 2017). These algorithms have been successful in providing insights into the underlying neuroscientific processes through which brain suppresses the ignored speaker in a dual-speaker scenario. Using speech envelope as the input acoustic feature, linear algorithms could generate system response functions that characterize the auditory pathway in the forward direction. These system response functions are referred to as temporal response function (TRF) (Lalor and Foxe, 2010). Analysis of the shape of TRFs has revealed that the human brain encodes the attended speaker different to that of the ignored speaker. Specifically, TRFs corresponding to the attended speaker have salient peaks around 100 and 200 ms which are weaker in TRFs corresponding to the ignored speaker (Fiedler et al., 2019; Kuruvila et al., 2021). Similar attention modulation effects were observed when the acoustic input was modified to using speech spectrogram or higher level features such as phonetics (Di Liberto et al., 2015). Likewise using backward models, the input stimulus can be reconstructed from EEG signals (stimulus reconstruction method) and a listener's attention could be inferred by comparing the reconstructed stimulus to the input stimuli (O'Sullivan et al., 2014). These findings give the possibility of integrating AAD algorithms into hearing aids which in combination with robust speech separation algorithms could greatly enhance the amenity provided to the users.

It has been well-established that the human auditory system is inherently non-linear (Zwicker and Fastl, 2013) and AAD analysis based on linear systems theory addresses the issue of non-linearity to a certain extend in the preprocessing stage. For example, during speech envelope extraction. Another limitation of linear methods is the longer time delay required to classify attention (Fuglsang et al., 2017; Geirnaert et al., 2019), although there were attempts to overcome this limitation (Miran et al., 2018; Kuruvila et al., 2021). In the last few years, deep neural networks have become popular especially in the field of computer vision and natural language processing. Since neural networks have the ability to model non-linearity, they have been used to estimate the dynamic state of brain from EEG signals (Craik et al., 2019). Similarly in AAD paradigm, convolutional neural network (CNN) based models were proposed where the stimulus reconstruction algorithm was implemented using the CNN model to infer attention (Ciccarelli et al., 2019; de Taillez et al., 2020). A direct classification of attention which bypasses the regression task of stimulus reconstruction, instead classifies whether the attention is to speaker 1 or speaker 2 directly was proposed in Ciccarelli et al. (2019) and Vandecappelle et al. (2021). In a non-competing speaker experiment, classifying attention as successful vs unsuccessful or match vs mismatch was further addressed in Monesi et al. (2020) and Tian and Ma (2020).

All aforementioned neural network models either did not use speech features or made use of only speech envelope as the input feature. As neural networks are data driven models, additional data/information about the speech stimuli may improve the performance of the network. In speech separation algorithms based on neural networks, spectrogram is used as the input feature to separate multiple speakers from a speech mixture (Wang and Chen, 2018). Inspired by the joint audio-visual speech separation model (Ephrat et al., 2018), we present a novel neural network framework that make use the speech spectrogram of multiple speakers and the EEG signals as inputs to classify the auditory attention.

The rest of the paper is organized as follows. In section 2, details of the datasets that were used to train and validate the neural network are provided. In section 3, the neural network architecture is explained in detail. The results are presented in sections 4, 5 provides a discussion on the results.

2. MATERIALS AND METHODS

2.1. Examined EEG Datasets

We evaluated the performance of our neural network model using three different EEG datasets. The first dataset was collected at our lab and it will be referred to as FAU_Dataset (Kuruvila et al., 2021). The second and third datasets are publicly available and they will be referred to as DTU_Dataset (Fuglsang et al., 2018) and KUL_Dataset (Das et al., 2019), respectively.

2.1.1. FAU_Dataset

This dataset comprised of EEG collected from 27 subjects who were all native German speakers. A cocktail party effect was simulated by presenting two speech stimuli simultaneously using loudspeakers and the subject was asked to attend selectively to one of the two stimuli. Speech stimuli were taken from the slowly spoken news section of the German news website www.dw.de and were read by two male speakers. The experiment consisted of six different presentations with each presentation being approximately five minutes long making it a total of 30 min. EEG was collected using 21 AgCl electrodes placed over the scalp according to the 10–20 EEG format. The reference electrode was placed at the right mastoid, the ground electrode was placed at the left earlobe and the EEG signals were sampled at 2,500 Hz. More details of the experiment could be found in Kuruvila et al. (2021).

2.1.2. DTU_Dataset

This is a publicly available dataset that was part of the work presented in Fuglsang et al. (2017). The dataset consisted of 18 subjects who selectively attended to one of the two simultaneous speakers. Speech stimuli were excerpts taken from Danish audiobooks that were narrated by a male and a female speaker. The experiment consisted of 60 segments with each segment being 50 s long making it a total of 50 min. EEG were recorded using 64 electrodes and were sampled at 512 Hz. The reference electrode was chosen either as the left mastoid or as the right mastoid after visual inspection. Further details can be found in Fuglsang et al. (2017, 2018).

2.1.3. KUL_Dataset

The final dataset that was analyzed is another publicly available dataset where 16 subjects undertook selective attention experiment. Speech stimuli consisted of four Dutch stories narrated by male speakers. Each story was 12 min long which was further divided into two 6 min presentations. EEG was recorded using 64 electrodes and were sampled at 8,196 Hz. The reference electrode was chosen either as TP7 or as TP8 electrode after visually inspecting the quality of the EEG signal measured at these locations. The experiment consisted of three different conditions namely HRTF, dichotic and repeated stimuli. In this work, we analyzed only the dichotic condition which was 24 min long. Additional details about the experiment and the dataset can be found in Das et al. (2016, 2019).

Details of the datasets are summarized again in **Table 1**. A total of 34.9 h of EEG data were examined in this work. However, the speech stimuli used were identical across subjects per dataset and they totalled 104 min of dual-speaker data. In all the three datasets that were analyzed, the two speakers read out different stimuli. Moreover, the stimuli were presented only once to the subject in order to avoid any learning effect. For each subject, the training and the test data were split as 75–25% and we ensured that no part of the EEG or the speech used in the test data was part of the training data. The test data were further divided equally into two halves and one half was used as a validation set during the training procedure.

2.2. Data Analysis

As EEG signals analyzed were collected at different sampling frequencies, they were all low pass filtered at a cut off frequency of 32 Hz and downsampled to 64 Hz sampling rate. Additionally, signals measured at only 10 electrode locations were considered for analysis and they were F7, F3, F4, F8, T7, C3, Cz, C4, T8, Pz. We analyzed four different trial durations in this study namely 2, 3, 4, and 5 s. For 2 s trials, an overlap of 1 s was applied. Thus, there were 118,922 trials in total for analysis. In order to maintain the total number of trials constant, 2 s of overlap was used in case of 3 s trial, 3 s of overlap was used in case of 4 s trial and 4 s overlap was used in case of 5 s trial. EEG signals in each trial were further high pass filtered with a cut off frequency of 1 Hz and the filtered signals were normalized to have zero mean and unit variance at each electrode location.

Speech stimuli were initially low pass filtered with a cut off frequency of 8 kHz and were downsampled to a sampling rate of 16 kHz. Subsequently, they were segmented into trials with a duration of 2, 3, 4, and 5 s at an overlap of 1, 2, 3, and 4 s, respectively. The speech spectrogram for each trial was obtained by taking the absolute value of the short-time Fourier transform (STFT) coefficients. The STFT was computed using a Hann window of 32 ms duration with a 12 ms overlap. Most of the analysis in our work was performed using 3 s trial and other trial durations were used only for comparison purposes. A summary of the dimensions of EEG signals and speech spectrogram after preprocessing for different trial durations is provided in **Table 2**.

3. NETWORK ARCHITECTURE

A top level view of the proposed neural network architecture is shown in **Figure 1**. It consists of three subnetworks namely EEG_CNN, Audio_CNN, and AE_Concat.

TABLE 1 | Details of the EEG datasets analyzed.

Name	Number of subjects	Duration per subject (minutes)	Total duration (hours)	Experiment type	Language
FAU_Dataset	27	30	13.5	Male + Male	German
DTU_Dataset	18	50	15	Male + Female	Danish
KUL_Dataset	16	24	6.4	Male + Male	Dutch

TABLE 2 | Trial duration vs. dimension of the input.

Trial duration (sec)	EEG data (time × num_electrodes)	Speech data (time × freq)
2	128 × 10	101 × 257
3	192 × 10	151 × 257
4	256 × 10	201 × 257
5	320 × 10	251 × 257

3.1. EEG_CNN

The EEG subnetwork comprised of four different convolutional layers as shown in **Table 3**. The kernel size of the first layer was chosen as 24 and it corresponded to a latency of 375 ms in the time domain. A longer kernel was chosen because previous studies have shown that the TRFs corresponding to attended and unattended speakers differ around 100 and 200 ms (Fiedler et al., 2019; Kuruvila et al., 2021). Therefore, a latency of 375 ms could help us to extract features that modulate the attention to different speakers in a dual-speaker environment. All other layers were initialized with kernels of shorter duration as shown in **Table 3**. All convolutions were performed using a stride of 1 × 1 and after the convolutions, max pooling was used to reduce the dimensionality. To prevent overfitting on the training data and improve generalization, dropout (Srivastava et al., 2014), and batch normalization (BN) (Ioffe and Szegedy, 2015) were applied. Subsequently, the output was passed through a non-linear activation function which was chosen as rectified linear unit (ReLU). The dimension of the input to EEG_CNN varied according to the length of the trial (**Table 2**) but the dimension of the output was fixed at 48 × 32. The max pooling parameter was slightly modified for different trial durations to obtain the fixed output dimension. The first dimension (48) corresponded to the temporal axis and the second dimension (32) corresponded to the number of convolution kernels. The dimension of the output that mapped the EEG signals measured at different electrodes was reduced to one by the successive application of max pooling along the electrode axis.

3.2. Audio_CNN

The audio subnetwork that processed the speech spectrogram consisted of five convolution layers whose parameters are shown in **Table 4**. All standard procedures such as max pooling, batch normalization, dropout, and ReLU activation were applied to the convolution output. Similar to the EEG_CNN, dimension of the input to the Audio_CNN varied according to the trial duration (**Table 2**) but the dimension of the output feature map was always

TABLE 3 | CNN parameters of the EEG subnetwork.

	Number of kernels	Kernel size	Dilation	Padding	Maxpool
Layer 1	32	24 × 1	1.1	12.0	2.1
Layer 2	32	7 × 1	2.1	6.0	1.2
Layer 3	32	7 × 5	1.1	3.2	2.5
Layer 4	32	7 × 1	1.1	3.0	1.1

TABLE 4 | CNN parameters of the Audio subnetwork.

	Number of kernels	Kernel size	Dilation	Padding	Maxpool
Layer 1	32	1 × 7	1.1	0.3	1.1
Layer 2	32	7 × 1	1.1	0.0	1.4
Layer 3	32	3 × 5	8.8	0.16	1.2
Layer 4	32	3 × 3	16.16	0.16	1.1
Layer 5	1	1 × 1	1.1	0.0	2.2

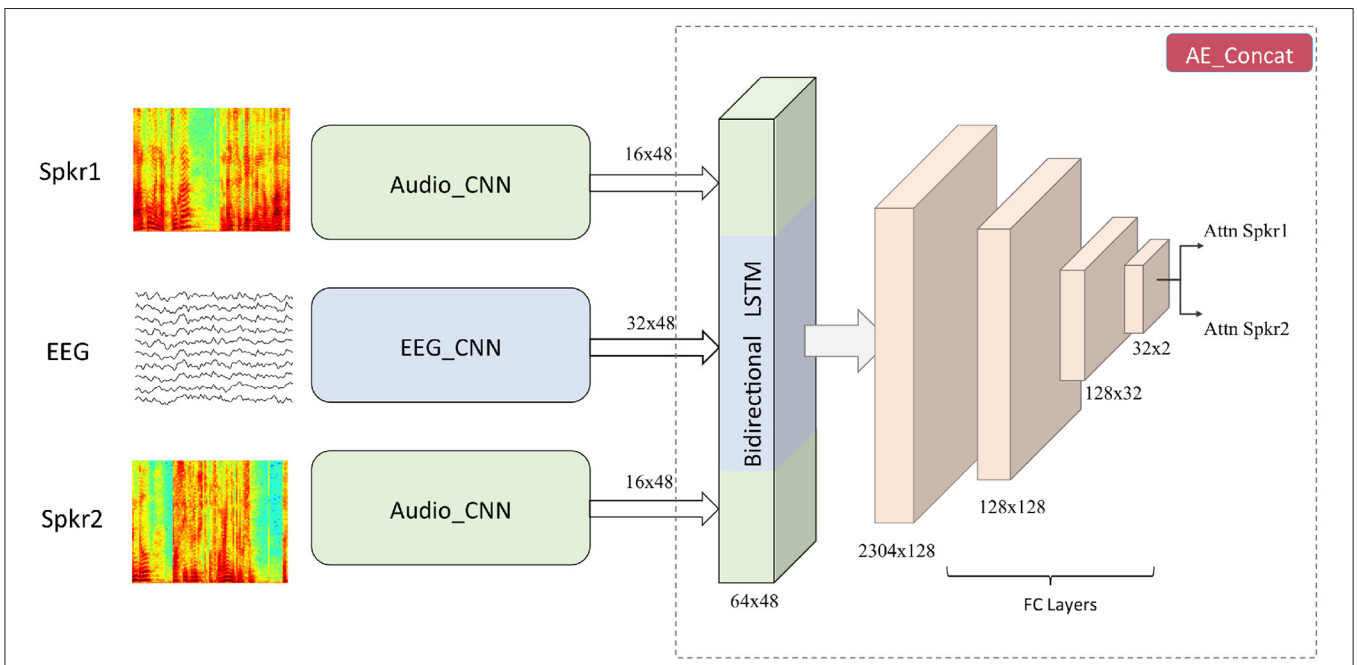


FIGURE 1 | The architecture of the proposed joint CNN-LSTM model. Input to the audio stream is the spectrogram of speech signals and input to the EEG stream is the downsampled version of EEG signals. Number of Audio_CNNs depends on the number of speakers present in the auditory scene (here two). From the outputs of Audio_CNN and EEG_CNN, speech and EEG embeddings are created which are concatenated together and passed to a BLSTM layer followed by FC layers.

fixed at 48 × 16. As the datasets considered in this study were taken from dual-speaker experiments, the Audio_CNN was run twice resulting in two sets of output.

3.3. AE_Concat

The feature maps obtained from EEG_CNN and Audio_CNN were concatenated along the temporal axis and the dimension of the feature map after concatenation was 48 × 64. In this way, we ensured that half of the feature map was contributed from the EEG data and half of the feature map was contributed from the speech data. This also provides the flexibility to extend to more than two speakers such as the experiment performed in Schäfer et al. (2018). The concatenated feature map was passed through a bidirectional long short-term memory (BLSTM) layer (Hochreiter and Schmidhuber, 1997; Schuster and Paliwal, 1997) which was followed by four fully connected (FC) layers. For the first three FC layers, ReLU activation was used and for the last FC layer, softmax activation was applied which helps us to classify the attention to speaker 1 or speaker 2.

The total number of EEG samples and audio samples (trials) available was 118,922 and 75% of the total available samples (89,192) were used to train the network and the rest of the available samples (29,730) were equally split as validation and test data. The network was trained for 80 epochs using a mini batch size of 32 samples and with a learning rate of $5 * 10^{-4}$. The drop out probability was set to 0.25 for the EEG_CNN and the AE_Concat subnetworks but it was increased to 0.4 for the Audio_CNN subnetwork. A larger drop out probability was used for the Audio_CNN because speech stimuli were identical across subjects for a particular dataset. Hence, when trained on data from multiple subjects, the speech data remain identical and the network may remember the speech spectrogram of the training data. The network was optimized using Adam optimizer (Kingma and Ba, 2014) and the loss function used was binary cross entropy. As neural network training can result in random variations from epoch to epoch, the test accuracy was calculated as the median accuracy of the last five epochs (Goyal et al., 2017). The network was trained using an Nvidia Geforce RTX-2060 (6 GB) graphics card and took ∼36 h to complete the training. The neural network model was developed in PyTorch and the python code is available at: https://github.com/ivine-GIT/joint_CNN_LSTM_AAD.

3.4. Sparse Neural Network: Magnitude Pruning

Despite neural network learning being a sophisticated algorithm, it is still not widely used in embedded devices due to the high memory and computational power requirements. Sparse neural networks have been recently proposed to overcome these challenges and enable running these models on embedded devices (Han et al., 2015). In sparse networks, majority of the model parameters are zeros and zero-valued multiplications can be ignored thereby reducing the computational requirement. Similarly, only non-zero weights need to be stored on the device and for all the zero-valued weights, only their position needs to be known reducing the memory footprint. Empirical evidences have shown that neural networks tolerate high level of sparsity (Han et al., 2015; Narang et al., 2017; Zhu and Gupta, 2017).

Sparse neural networks are found out by using a procedure known as network pruning. It consists of three steps. First, a large over-parameterized network is trained in order to obtain a high test accuracy as over-parameterization has stronger representation power (Luo et al., 2017). Second, from the trained over-parameterized network, only important weights based on certain criterion are retained and all other weights are assumed to be redundant and reinitialized to zero. Finally, the pruned network is fine-tuned by training it further using only the retained weights so as to improve the performance. Searching for the redundant weights can be based on simple criteria such as magnitude pruning (Han et al., 2015) or based on complex algorithms such as variational dropout (Molchanov et al., 2017) or L0 regularization (Louizos et al., 2017). However, it was shown that introducing sparsity using magnitude pruning could achieve comparable or better performance than complex techniques such as variational dropout or L0 regularization (Gale et al., 2019). Hence, we will present results based on only magnitude pruning in this work.

4. RESULTS

4.1. Attention Decoding Accuracy

To evaluate the performance of our neural network, we trained the model under different scenarios using a trial duration of 3 s. In the first scenario (*Ind set train*), attention decoding accuracies were calculated per individual dataset. In other words, to obtain the test accuracy of subjects belonging to FAU_Dataset, the model was trained using training samples only from FAU_Dataset leaving out DTU_dataset and KUL_Dataset. Similarly, to obtain the test accuracy for DTU_Dataset, the model was trained using training samples only from DTU_Dataset. The same procedure was repeated for KUL_Dataset. The median decoding accuracy was 72.6% for FAU_Dataset, 48.1% for DTU_Dataset, and 69.1% for KUL_Dataset (**Figure 2**). In the second scenario (*Full set train*), accuracies were calculated by combining training samples from all the three datasets together. The median decoding accuracies obtained in this scenario were 84.5, 52.9, and 77.9% for FAU_Dataset, DTU_Dataset, and KUL_Dataset, respectively. The results from the second scenario showed a clear improvement over the first scenario ($p_{\_FAU} < 0.001$; $p_{\_DTU} < 0.05$; $p_{\_KUL} < 0.01$) suggesting that the model generalizes better in the *Full set train*. Furthermore, to evaluate the cross-set training performance, we trained the model using one dataset and tested it on the other two datasets. For example, the training would be performed using FAU_Dataset and testing would be performed on both DTU and KUL datasets. The same procedure was repeated by training using the DTU dataset and the KUL dataset. The decoding accuracies obtained were all at chance level across the three cross-set training scenarios (**Figure 3**). Consequently, all results presented further in this paper are based on *Full set train*. The statistical analyses are based on paired Wilcoxon signed-rank test with sample sizes given in **Table 1**.

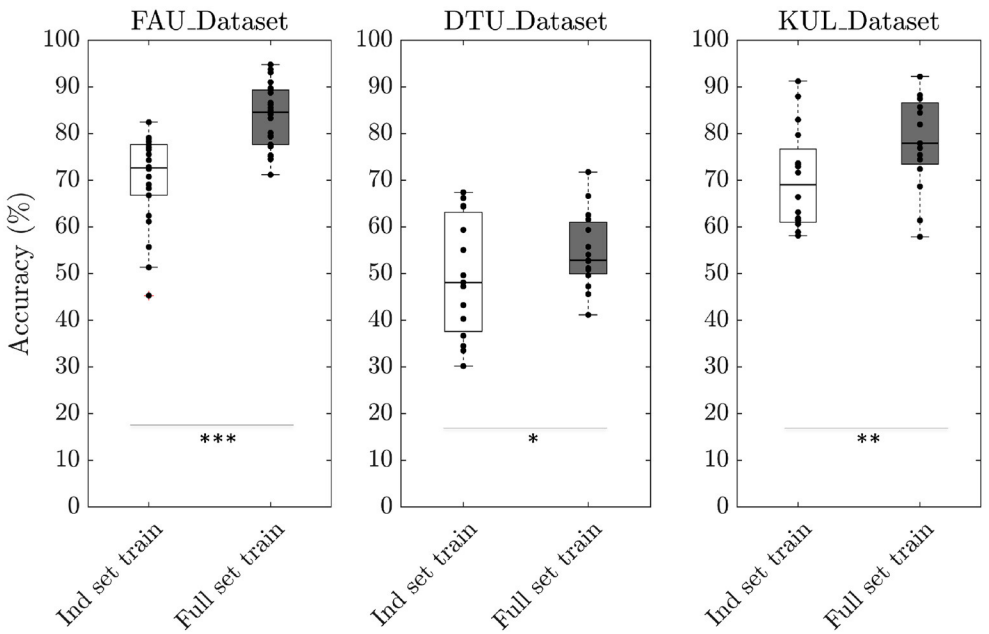


FIGURE 2 | Boxplot depicting the decoding accuracies obtained using two different training scenarios. In the first scenario (*Ind set train*), individual dataset accuracies were obtained by using training samples only from that particular dataset. For example, to calculate the test accuracy of FAU_Dataset, training samples were taken only from FAU_Dataset. In the second scenario (*Full set train*), individual dataset accuracies were obtained using training samples from all the three datasets combined. As a result, there are more training samples in the second scenario compared to the first (*$p < 0.05$; **$p < 0.01$; ***$p < 0.001$ based on paired Wilcoxon signed-rank test).

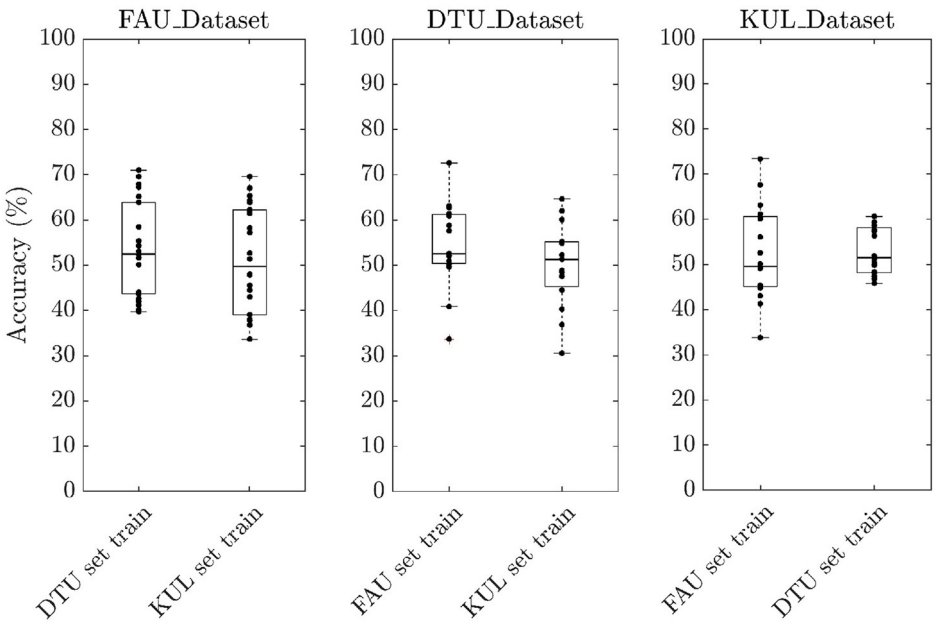


FIGURE 3 | Boxplot showing the decoding accuracies obtained for cross-set training scenario. The accuracies obtained were all at chance level.

4.2. Decoding Accuracy vs. Trial Duration

To analyse the effect of trial duration on the attention decoding accuracy, the model was trained using trials of length 2, 3, 4, and 5 s. For every trial, only 1 s of new data were added and the remaining data were populated by overlapping to the previous trial using a sliding window. Specifically, for 2 s trial, 1 s of

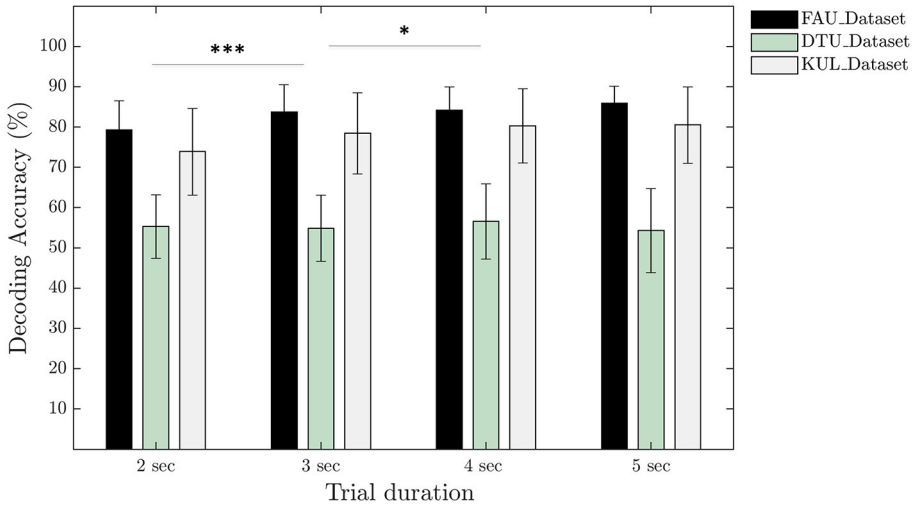


FIGURE 4 | Comparison of the decoding accuracies calculated for different trial durations per dataset. Statistical analysis based on paired Wilcoxon signed-rank test and pooled over all subjects together from the three datasets (*$p < 0.05$; ***$p < 0.001$).

overlap was used and for 3 s trial, 2 s of overlap was used, and so on. In this way, total number of training samples remained constant for different trial durations considered in our analysis. The mean decoding accuracy across all subjects and all datasets in case of 2 s trial duration was 70.9 ± 13.2%. The mean accuracy improved to 73.9 ± 14.8% when the trial duration was increased to 3 s ($p < 0.001$, $r = 0.60$). Using a trial duration of 4 s, the mean accuracy obtained was 75.2 ± 14.3% which is a slight improvement over 3 s trials ($p < 0.05$, $r = 0.31$). For 5 s trials, our neural network model resulted in a mean accuracy of 75.5 ± 15.7% that was statistically identical to the accuracy obtained using 4 s trials ($p > 0.05$, $r = 0.10$). **Figure 4** depicts the accuracy calculated for individual datasets.

4.3. Ablation Analysis

In order to gain further insights into the architecture and understand the contribution of different parts of our neural network, we performed ablation analysis using a trial duration of 3 s. To this end, we modified the neural network architecture by removing specific block such as the BLSTM layer or the FC layers one at a time and retrained the modified network. Similarly, to understand the importance of the audio input feature, decoding accuracies were calculated by zeroing out the EEG input and to understand the importance of the EEG input feature, decoding accuracies were calculated by zeroing out the audio input. As shown in **Figure 5**, the median decoding accuracy by zeroing out the EEG input was 48.6% whereas zeroing out the audio input resulted in an accuracy of 53.6% resulting in no significant difference ($p > 0.05$). When the network was retrained by removing the BLSTM layer only, the median decoding accuracy obtained was 68.3% and on removing the FC layers only, median decoding accuracy was 74.7%. Hence, the BLSTM layer contributes more toward the network learning than the FC layer ($p < 0.001$). To compare, the median decoding accuracy calculated using the full the network was 77.2%.

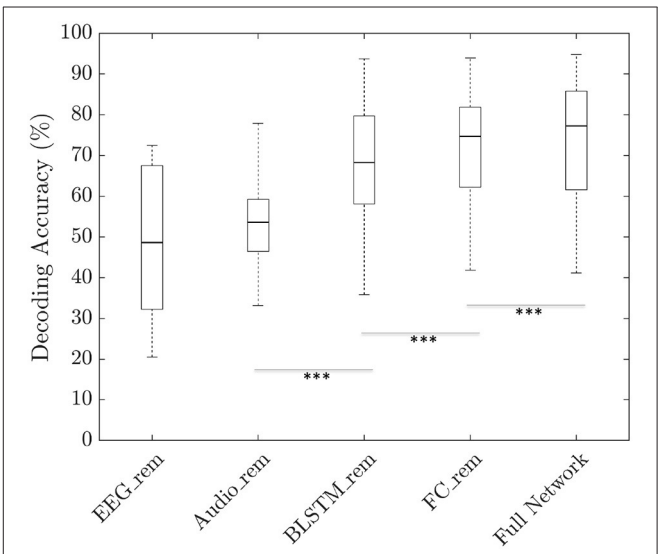


FIGURE 5 | Boxplots showing the decoding accuracies obtained by ablating the different blocks such as FC layer or BLSTM layer. To obtain the test accuracies after ablating, the ablated network was trained from scratch in case of FC_rem and BLSTM_rem. However, in case of Audio_rem and EEG_rem, accuracies were calculated by zeroing out the corresponding input features before passing them to a fully trained network. The obtained accuracy did not demonstrate a statistically significant difference between Audio_rem and EEG_rem ($p > 0.05$). For all other cases, there was a significant difference (***$p < 0.001$ based on paired Wilcoxon signed-rank test).

4.4. Sparse Neural Network Using Magnitude Pruning

To investigate the degree of sparsity that our neural network can tolerate, we pruned the model at 40, 50, 60, 70, and 80% sparsity

using the 3 s trial duration. In order to fine-tune the pruned neural network, there are two options: (1) sequential or (2) one-shot. In sequential fine-tuning, weights of the trained original model are reinitialized to zero in smaller steps per epoch until the required sparsity is attained. In one-shot fine-tuning, weights of the trained original model are reinitialized to zero at one shot in the first epoch and the sparse model is further trained to improve performance. We observed that the sequential fine-tuning is less efficient than one-shot fine-tuning in terms of training time budget. Therefore, all results presented here are based on one-shot fine-tuning. We achieved a median decoding accuracy of 76.9% at a sparsity of 40% which is statistically identical to the original model at 77.2% ($p > 0.05$). When the sparsity was increased to 50%, the median decoding accuracy decreased to 75.7% which was lower than the original model ($p < 0.001$). Increasing the sparsity level further resulted in deterioration of decoding accuracy reaching 63.2% at a sparsity of 80% (**Figure 6**). Total number of learnable parameters in our model was 416,741 and to find the sparse network, we pruned only the weights leaving the bias and BN parameters unchanged.

5. DISCUSSION

People with hearing loss suffer from deteriorated speech intelligibility in noisy acoustic environments such as multispeaker scenarios. Increasing the audibility by means of hearing aids has not shown to provide sufficient improvement to the speech intelligibility. This is because the hearing aids are unable to estimate *apriori* to which speaker the user intends to listen. Hence, hearing aids amplify both the wanted signal (attended speaker) and interfering signals (ignored speakers). Recently, it has been shown that the cortical signals measured using EEG could infer the auditory attention by discriminating between the attended speaker and the ignored speaker in a dual-speaker scenario (O'Sullivan et al., 2014). Linear system analysis has been the commonly used methodology to analyse the EEG signals measured from a listener performing selective attention. However, in recent years, non-linear analyses based on neural networks have become prominent, thanks to the availability of customized hardware accelerators and associated software libraries.

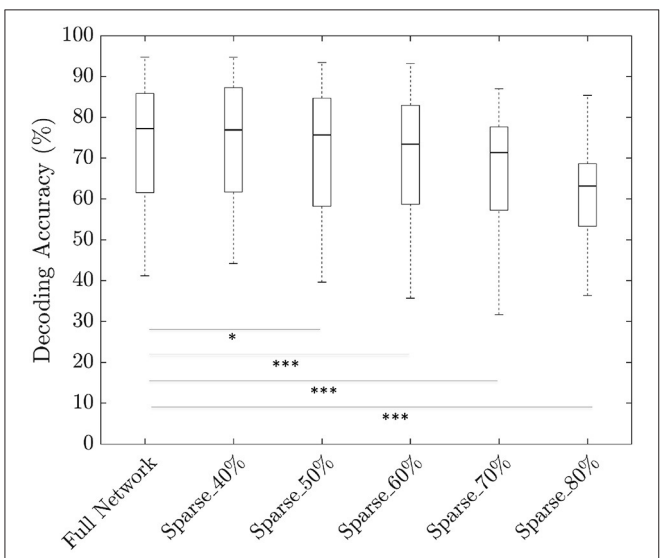


FIGURE 6 | Plots comparing the trade off between decoding accuracies and sparsity level (*$p < 0.05$; ***$p < 0.001$ based on paired Wilcoxon signed-rank test).

In this work, we developed a joint CNN-LSTM model to infer the auditory attention of a listener in a dual-speaker environment. CNNs take the EEG signal and spectrogram of the multiple speakers as inputs and extract features through successive convolutions. These convolutions generate an intermediate embeddings of the inputs which are then given to a BLSTM layer. As LSTMs fall under the category of recurrent neural networks, they can model the temporal relationship between the EEG embedding and the multiple spectrogram embeddings. Finally, the output of the BLSTM is processed through FC layers to infer the auditory attention. The effectiveness of the proposed neural network was evaluated with the help of three different EEG datasets collected from subjects who undertook dual-speaker experiment.

There are many choices for the acoustic cues of speech signal that could be given as input to the neural network. They are speech onsets (Howard and Poeppel, 2010), speech envelopes (Aiken and Picton, 2008), speech spectrograms (Pasley et al., 2012), or phonemes (Di Liberto et al., 2015). Due to the hierarchical processing of speech, all of the aforementioned cues could be tracked from the cortical signals measured using EEG (Hickok and Poeppel, 2007; Ding and Simon, 2014). Speech envelope is the most commonly used acoustic cues in the linear system analysis of EEG signal. However, we decided to use spectrogram due to its rich representational power of the corresponding speech signal and the ability of neural networks to index these multidimensional inputs efficiently.

5.1. Attention Decoding Accuracy

We analyzed the performance of our neural network in two different training scenarios. In the first scenario, individual dataset accuracy was found out by training the network using samples taken only from that particular dataset. In the second scenario, individual dataset accuracy was found out by training using samples combined from all three datasets together. The accuracies obtained in the second scenario were higher than the first scenario by 10.8% on average, which is in agreement with the premise of neural network learning that larger the amount of training data, the better the generalization. The decoding accuracies obtained for subjects belonging to the DTU_Dataset were markedly lower than the other two datasets similar to the observation made in Geirnaert et al. (2020). While the exact reason for the lower performance is unclear, a major difference of the DTU_Dataset compared to the other two datasets was that the former consisted of attention to male and female speakers whereas the latter consisted of attention to only male speakers. Therefore, training with additional EEG data that consist of attention to female speakers can provide more insights into

the lower performance. Additionally, we investigated the cross-set performance by training the model using one dataset and testing using the other two datasets. The accuracies obtained were all at chance level as seen in **Figure 3**. This is not against our expectation because if the underlying training set is not representative, neural networks will not generalize. Specifically, features in the training set and the test set are different since they were recorded in different audio settings, languages, and EEG devices. This further affirms the importance of having a large and diverse training set for the neural networks to function efficiently.

5.2. Decoding Accuracy vs. Trial Duration

One of the major challenges that AAD algorithms based on linear system theory faces is the deteriorated decoding performance when the trial duration is reduced. To this end, we calculated the accuracies using our neural network for different trial durations of 2, 3, 4, and 5 s. We observed a clear performance improvement when trial duration was increased from 2 to 3 s whereas for all other trial durations, accuracies did not improve substantially (**Figure 4**). However, increasing the trial duration will result in larger latency needed to infer the auditory attention that can adversely affect applications which require real-time operation. Hence, 3 s trial duration may be an optimal operation point as it is known from a previous study that human brain tracks the sentence phrases and phrases are normally not longer than 3 s (Vander Ghinst et al., 2019). Similarly, our analysis made use of 10 electrodes distributed all over the scalp but future work should investigate the effect of reducing the number of electrodes. This will help in integrating algorithms based on neural networks into devices such as hearing aids. We anticipate that the current network will require modifications with additional hyperparameter tuning in order to accommodate for the reduction in number of electrodes, as the fewer is the number of electrodes, the lower is the amount of data available for training.

5.3. Ablation Analysis

Performing ablation analysis gives the possibility to evaluate the contribution of different inputs and modules in a neural network. To our model, when only the speech features were given as input, the median decoding accuracy was 48.6% whereas only EEG features as input resulted in an accuracy of 53.6% (**Figure 5**). However, statistical analysis revealed that there is no significant difference between the two. This is contrary to our anticipation because we expected the model to learn more from the EEG features than from the audio features, as the EEG signal is unique to the subject while the audio stimulus was repeated among subjects per dataset. Nevertheless, in future care must be taken to design the experiment in such a way as to incorporate diverse speech stimuli. Further analysis ablating the BLSTM layer and the FC layers revealed that the BLSTM layer was more important than the FC layers. This is probably due to the ability of the LSTM layer to model the temporal delay between speech cues and the EEG. However, we anticipate that when the training datasets become larger and more dissimilar, FC layers will become more important due to the improved representation and optimization power of dense networks (Luo et al., 2017).

5.4. Sparse Neural Networks

Although neural networks achieve state-of-the-art performances for a wide range of applications, they have large memory footprint and require extremely high computation power. Over the years, neural networks were able to extend their scope of applications was by scaling up the network size. In 1998, the CNN model (LeNet) that was successful in recognizing handwritten digits consisted of under a million parameters (LeCun et al., 1998), whereas AlexNet that won the ImageNet challenge in 2012 consisted of 60 million parameters (Krizhevsky et al., 2017). Neural networks were further scaled up to the order of 10 billion parameters and efficient methods to train these extremely large networks were presented in Coates et al. (2013). While these large models are very powerful, running them on embedded devices poses huge challenges due to the large memory and computation requirements. Sparse neural networks are a novel architecture search where redundant weights are reinitialized to zero thereby reducing the computation load.

Investigation into the amount of sparsity that our neural network can tolerate revealed a tolerance of upto 50% sparsity without substantial loss of accuracy (**Figure 6**). However, standard benchmarking on sparsity has found that deep networks such as ResNet-50 can tolerate upto 90% sparsity (Gale et al., 2019). One of the potential reasons for the lower level of sparsity in our model is due to its shallow nature. That is, our model is comprised of less than half a million learnable parameters while deep networks such as ResNet-50 is comprised of over 25 million learnable parameters. It is also interesting to note that the accuracy obtained by removing the FC layer in our ablation analysis was 74.6% compared to the full network accuracy of 77.2%. And the ablated network consisted of 105,605 parameters which is approximately only a quarter of the total number of parameters (416,741) of the original network. This shows that by careful design choices, we can reduce the network size considerably compared to an automatic sparse network search using magnitude pruning.

Sparsification of neural network has also been investigated as a neural network architecture search rather than merely as an optimization procedure. In the lottery ticket hypothesis presented in Frankle and Carbin (2018), authors posit that, inside the structure of an over-parameterized network, there exist subnetworks (winning tickets) that when trained in isolation reaches accuracies comparable to the original network. The pre-requisite to achieve comparable accuracy is to initialize the sparse network using the original random weight initialization that was used to obtain the sparse architecture. However, it was shown that with careful choice of the learning rate, the stringent requirement on original weight initialization can be relaxed and the sparse network can be trained from scratch for any random initialization (Liu et al., 2018).

One of the assumptions that we have made throughout this paper is the availability of clean speech signal to obtain the spectrogram. In practice, only noisy mixtures are available and speech sources must be separated before the spectrogram can

be calculated. This is an active research field and algorithms are already available based on classical signal processing such as beamforming or based on deep neural networks (Wang and Chen, 2018). Another challenge in neural network learning and in particular, its application in EEG research is the scarcity of labeled data to train the network. This limits the ability of network to generalize well to unseen EEG data. To mitigate the aforementioned limitation, data augmentation techniques are widely used in neural network training. Data augmentation is a procedure to generate synthetic dataset that spans unexplored input signal space but corresponding to the true labels (Wen et al., 2020). In auditory attention paradigm, linear system analyses have shown that the TRF properties differ between attended and ignored speakers (Fiedler et al., 2019; Kuruvila et al., 2021). As a result, synthetic EEG can be generated by performing a linear convolution between TRFs and the corresponding speech signal cues (Miran et al., 2018). The signal-to-noise ratio of the synthesized EEG can be varied by adding appropriate noise to the convolved signal. The most commonly used speech cue is the signal envelope obtained using Hilbert transform. However, more sophisticated envelope extraction methods such as the computational models simulating the auditory system could improve the quality of synthesized EEG signals (Kates, 2013; Verhulst et al., 2018). It must be noted that the data augmentation techniques must only be used to train the network. The validation and the testing procedure must still be performed using real datasets.

6. CONCLUSION

Integrating EEG to track the cortical signals is one of the latest proposals to enhance the quality of service provided by hearing aids to the users. EEG is envisaged to provide neuro-feedback about the user's intention thereby enabling the hearing aid to infer and enhance the attended speech signals. In the present study, we propose a joint CNN-LSTM network to classify the attended speaker in order to infer the auditory attention of a listener. The proposed neural network uses speech spectrograms and EEG signals as inputs to infer the auditory attention. Results obtained by training the network using three different EEG datasets collected from multiple subjects who undertook a dual-speaker experiment showed that our network generalizes well to different scenarios. Investigation into the importance of different constituents of our network architecture revealed that adding an LSTM layer improved the performance of the model considerably. Evaluating sparsity on the proposed joint CNN-LSTM network demonstrates that the network can tolerate upto 50% sparsity without considerable deterioration in performance. These results could pave way to integrate algorithms based on neural networks into hearing aids that have constrained memory and computational power.

DATA AVAILABILITY STATEMENT

The original contributions presented in the study are included in the article/supplementary material, further inquiries can be directed to the corresponding author/s.

ETHICS STATEMENT

The studies involving human participants were reviewed and approved by Ethics committee, University of Erlangen-Nuremberg (Protocol Number: 314_18B) on 18th September 2018. The patients/participants provided their written informed consent to participate in this study.

AUTHOR CONTRIBUTIONS

UH and EF conceived the study and contributed to data analysis, supervision, and manuscript writing. JM contributed to the results interpretation and data analysis. IK is the main author and performed most of the data analysis. All authors contributed to the article and approved the submitted version.

FUNDING

This work was supported by a grant from *Johannes und Frieda Marohn-Stiftung, Erlangen*.

ACKNOWLEDGMENTS

We convey our gratitude to all participants who took part in the study and would like to thank the student Laura Rupprecht who helped us with data acquisition.

REFERENCES

Aiken, S. J., and Picton, T. W. (2008). Human cortical responses to the speech envelope. *Ear Hear.* 29, 139–157. doi: 10.1097/AUD.0b013e31816453dc

Biesmans, W., Das, N., Francart, T., and Bertrand, A. (2017). Auditory-inspired speech envelope extraction methods for improved eeg-based auditory attention detection in a cocktail party scenario. *IEEE Trans. Neural Syst. Rehabil. Eng.* 25, 402–412. doi: 10.1109/TNSRE.2016.2571900

Broderick, M. P., Anderson, A. J., and Lalor, E. C. (2019). Semantic context enhances the early auditory encoding of natural speech. *J. Neurosci.* 39, 7564–7575. doi: 10.1523/JNEUROSCI.0584-19.2019

Cherry, E. C. (1953). Some experiments on the recognition of speech, with one and with two ears. *J. Acoust. Soc. Am.* 25, 975–979. doi: 10.1121/1.1907229

Ciccarelli, G., Nolan, M., Perricone, J., Calamia, P. T., Haro, S., O'Sullivan, J., et al. (2019). Comparison of two-talker attention decoding from EEG with nonlinear neural networks and linear methods. *Sci. Rep.* 9, 1–10. doi: 10.1038/s41598-019-47795-0

Coates, A., Huval, B., Wang, T., Wu, D., Catanzaro, B., and Andrew, N. (2013). "Deep learning with COTS HPC systems," in *International Conference on Machine Learning* (Atlanta, GA), 1337–1345.

Craik, A., He, Y., and Contreras-Vidal, J. L. (2019). Deep learning for electroencephalogram (EEG) classification tasks: a review. *J. Neural Eng.* 16:031001. doi: 10.1088/1741-2552/ab0ab5

Das, N., Biesmans, W., Bertrand, A., and Francart, T. (2016). The effect of head-related filtering and ear-specific decoding bias on auditory attention detection. *J. Neural Eng.* 13:056014. doi: 10.1088/1741-2560/13/5/056014

Das, N., Francart, T., and Bertrand, A. (2019). Auditory attention detection dataset KULeuven (Version 1.0.0) [Data set]. *Zenodo*. doi: 10.5281/zenodo.3377911

de Taillez, T., Kollmeier, B., and Meyer, B. T. (2020). Machine learning for decoding listeners' attention from electroencephalography evoked by continuous speech. *Eur. J. Neurosci.* 51, 1234–1241. doi: 10.1111/ejn.13790

Di Liberto, G. M., O'Sullivan, J. A., and Lalor, E. C. (2015). Low-frequency cortical entrainment to speech reflects phoneme-level processing. *Curr. Biol.* 25, 2457–2465. doi: 10.1016/j.cub.2015.08.030

Ding, N., and Simon, J. Z. (2012). Neural coding of continuous speech in auditory cortex during monaural and dichotic listening. *J. Neurophysiol.* 107, 78–89. doi: 10.1152/jn.00297.2011

Ding, N., and Simon, J. Z. (2014). Cortical entrainment to continuous speech: functional roles and interpretations. *Front. Hum. Neurosci.* 8:311. doi: 10.3389/fnhum.2014.00311

Ephrat, A., Mosseri, I., Lang, O., Dekel, T., Wilson, K., Hassidim, A., et al. (2018). Looking to listen at the cocktail party: a speaker-independent audio-visual model for speech separation. *arXiv preprint arXiv:1804.03619*. doi: 10.1145/3197517.3201357

Feldman, J. (2003). What is a visual object? *Trends Cogn. Sci.* 7, 252–256. doi: 10.1016/S1364-6613(03)00111-6

Fiedler, L., Woestmann, M., Graversen, C., Brandmeyer, A., Lunner, T., and Obleser, J. (2017). Single-channel in-ear-EEG detects the focus of auditory attention to concurrent tone streams and mixed speech. *J. Neural Eng.* 14:036020. doi: 10.1088/1741-2552/aa66dd

Fiedler, L., Wöstmann, M., Herbst, S. K., and Obleser, J. (2019). Late cortical tracking of ignored speech facilitates neural selectivity in acoustically challenging conditions. *NeuroImage* 186, 33–42. doi: 10.1016/j.neuroimage.2018.10.057

Frankle, J., and Carbin, M. (2018). The lottery ticket hypothesis: finding sparse, TRainable neural networks. *arXiv preprint arXiv:1803.03635*.

Fuglsang, S. A., Dau, T., and Hjortkjær, J. (2017). Noise-robust cortical tracking of attended speech in real-world acoustic scenes. *NeuroImage* 156, 435–444. doi: 10.1016/j.neuroimage.2017.04.026

Fuglsang, S. A., Wong, D. D. E., and Hjortkjær, J. (2018). EEG and audio dataset for auditory attention decoding (Version 1) [Data set]. *Zenodo*. doi: 10.5281/zenodo.1199011

Gale, T., Elsen, E., and Hooker, S. (2019). The state of sparsity in deep neural networks. *arXiv preprint arXiv:1902.09574*.

Geirnaert, S., Francart, T., and Bertrand, A. (2019). An interpretable performance metric for auditory attention decoding algorithms in a context of neuro-steered gain control. *IEEE Trans. Neural Syst. Rehabil. Eng.* 28, 307–317. doi: 10.1101/745695

Geirnaert, S., Vandecappelle, S., Alickovic, E., de Cheveigné, A., Lalor, E., Meyer, B. T., et al. (2020). Neuro-steered hearing devices: decoding auditory attention from the brain. *arXiv preprint arXiv:2008.04569*. doi: 10.1109/MSP.2021.3075932

Goyal, P., Dollár, P., Girshick, R., Noordhuis, P., Wesolowski, L., Kyrola, A., et al. (2017). Accurate, large minibatch SGD: training imagenet in 1 hour. *arXiv preprint arXiv:1706.02677*.

Han, S., Pool, J., Tran, J., and Dally, W. (2015). Learning both weights and connections for efficient neural network. *Adv. Neural Inform. Process. Syst.* 28, 1135–1143.

Hickok, G., and Poeppel, D. (2007). The cortical organization of speech processing. *Nat. Rev. Neurosci.* 8, 393–402. doi: 10.1038/nrn2113

Hochreiter, S., and Schmidhuber, J. (1997). Long short-term memory. *Neural Comput.* 9, 1735–1780. doi: 10.1162/neco.1997.9.8.1735

Howard, M. F., and Poeppel, D. (2010). Discrimination of speech stimuli based on neuronal response phase patterns depends on acoustics but not comprehension. *J. Neurophysiol.* 104, 2500–2511. doi: 10.1152/jn.00251.2010

Ioffe, S., and Szegedy, C. (2015). "Batch normalization: accelerating deep network training by reducing internal covariate shift," in *International Conference on Machine Learning* (Lille: PMLR), 448–456.

Kates, J. (2013). "An auditory model for intelligibility and quality predictions," in *Proceedings of Meetings on Acoustics ICA2013, Vol. 19* (Montréal, QC: Acoustical Society of America), 050184. doi: 10.1121/1.4799223

Kingma, D. P., and Ba, J. (2014). Adam: a method for stochastic optimization. *arXiv preprint arXiv:1412.6980*.

Krizhevsky, A., Sutskever, I., and Hinton, G. E. (2017). ImageNet classification with deep convolutional neural networks. *Commun. ACM* 60, 84–90. doi: 10.1145/3065386

Kuruvila, I., Demir, K. C., Fischer, E., and Hoppe, U. (2021). Inference of the selective auditory attention using sequential LMMSE estimation. *IEEE Trans. Biomed. Eng.* doi: 10.1109/TBME.2021.3075337

Kuruvila, I., Fischer, E., and Hoppe, U. (2020). "An LMMSE-based estimation of temporal response function in auditory attention decoding," in *2020 42nd Annual International Conference of the IEEE Engineering in Medicine & Biology Society (EMBC)* (Montréal, QC), 2837–2840. doi: 10.1109/EMBC44109.2020.9175866

Lalor, E. C., and Foxe, J. J. (2010). Neural responses to uninterrupted natural speech can be extracted with precise temporal resolution. *Eur. J. Neurosci.* 31, 189–193. doi: 10.1111/j.1460-9568.2009.07055.x

LeCun, Y., Bottou, L., Bengio, Y., and Haffner, P. (1998). Gradient-based learning applied to document recognition. *Proc. IEEE* 86, 2278–2324. doi: 10.1109/5.726791

Liu, Z., Sun, M., Zhou, T., Huang, G., and Darrell, T. (2018). Rethinking the value of network pruning. *arXiv preprint arXiv:1810.05270*.

Louizos, C., Welling, M., and Kingma, D. P. (2017). Learning sparse neural networks through *L*_0 regularization. *arXiv preprint arXiv:1712.01312*.

Luo, J.-H., Wu, J., and Lin, W. (2017). "Thinet: a filter level pruning method for deep neural network compression," in *Proceedings of the IEEE International Conference on Computer Vision* (Venice), 5058–5066. doi: 10.1109/ICCV.2017.541

Mesgarani, N., and Chang, E. F. (2012). Selective cortical representation of attended speaker in multi-talker speech perception. *Nature* 485, 233–236. doi: 10.1038/nature11020

Miran, S., Akram, S., Sheikhattar, A., Simon, J. Z., Zhang, T., and Babadi, B. (2018). Real-time tracking of selective auditory attention from M/EEG: a Bayesian filtering approach. *Front. Neurosci.* 12:262. doi: 10.3389/fnins.2018.00262

Mirkovic, B., Debener, S., Jaeger, M., and De Vos, M. (2015). Decoding the attended speech stream with multi-channel EEG: implications for online, daily-life applications. *J. Neural Eng.* 12:046007. doi: 10.1088/1741-2560/12/4/046007

Molchanov, D., Ashukha, A., and Vetrov, D. (2017). Variational dropout sparsifies deep neural networks. *arXiv preprint arXiv:1701.05369*.

Monesi, M. J., Accou, B., Montoya-Martinez, J., Francart, T., and Van Hamme, H. (2020). "An LSTM based architecture to relate speech stimulus to EEG," in *ICASSP 2020-2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)* (Barcelona), 941–945. doi: 10.1109/ICASSP40776.2020.9054000

Narang, S., Elsen, E., Diamos, G., and Sengupta, S. (2017). Exploring sparsity in recurrent neural networks. *arXiv preprint arXiv:1704.05119*.

O'Sullivan, J. A., Power, A. J., Mesgarani, N., Rajaram, S., Foxe, J. J., Shinn-Cunningham, B. G., et al. (2014). Attentional selection in a cocktail party environment can be decoded from single-trial EEG. *Cereb. Cortex* 25, 1697–1706. doi: 10.1093/cercor/bht355

Pasley, B. N., David, S. V., Mesgarani, N., Flinker, A., Shamma, S. A., Crone, N. E., et al. (2012). Reconstructing speech from human auditory cortex. *PLoS Biol.* 10:e1001251. doi: 10.1371/journal.pbio.1001251

Schäfer, P. J., Corona-Strauss, F. I., Hannemann, R., Hillyard, S. A., and Strauss, D. J. (2018). Testing the limits of the stimulus reconstruction approach: auditory attention decoding in a four-speaker free field environment. *Trends Hear.* 22:2331216518816600. doi: 10.1177/2331216518816600

Schuster, M., and Paliwal, K. K. (1997). Bidirectional recurrent neural networks. *IEEE Trans. Signal Process.* 45, 2673–2681. doi: 10.1109/78.650093

Shinn-Cunningham, B. G. (2008). Object-based auditory and visual attention. *Trends Cogn. Sci.* 12, 182–186. doi: 10.1016/j.tics.2008.02.003

Srivastava, N., Hinton, G., Krizhevsky, A., Sutskever, I., and Salakhutdinov, R. (2014). Dropout: a simple way to prevent neural networks from overfitting. *J. Mach. Learn. Res.* 15, 1929–1958.

Tian, Y., and Ma, L. (2020). Auditory attention tracking states in a cocktail party environment can be decoded by deep convolutional neural networks. *J. Neural Eng.* doi: 10.1088/1741-2552/ab92b2

Vandecappelle, S., Deckers, L., Das, N., Ansari, A. H., Bertrand, A., and Francart, T. (2021). EEG-based detection of the locus of auditory attention with convolutional neural networks. *bioRxiv [Preprint]*. doi: 10.7554/eLife.56481.sa2

Vander Ghinst, M., Bourguignon, M., Niesen, M., Wens, V., Hassid, S., Choufani, G., et al. (2019). Cortical tracking of speech-in-noise develops from childhood to adulthood. *J. Neurosci*. 39, 2938–2950. doi: 10.1523/JNEUROSCI.1732-18.2019

Verhulst, S., Altoé, A., and Vasilkov, V. (2018). Computational modeling of the human auditory periphery: auditory-nerve responses, evoked potentials and hearing loss. *Hear. Res*. 360, 55–75. doi: 10.1016/j.heares.2017.12.018

Vogel, E. K., Woodman, G. F., and Luck, S. J. (2005). Pushing around the locus of selection: evidence for the flexible-selection hypothesis. *J. Cogn. Neurosci*. 17, 1907–1922. doi: 10.1162/089892905775008599

Wang, D., and Chen, J. (2018). Supervised speech separation based on deep learning: an overview. *IEEE/ACM Trans. Audio Speech Lang. Process*. 26, 1702–1726. doi: 10.1109/TASLP.2018.2842159

Wen, Q., Sun, L., Song, X., Gao, J., Wang, X., and Xu, H. (2020). Time series data augmentation for deep learning: a survey. *arXiv preprint arXiv:2002.12478*.

Zhu, M., and Gupta, S. (2017). To prune, or not to prune: exploring the efficacy of pruning for model compression. *arXiv preprint arXiv:1710.01878*.

Zwicker, E., and Fastl, H. (2013). *Psychoacoustics: Facts and Models, Vol. 22*. Berlin: Springer Science & Business Media.

Conflict of Interest: The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

frontiers in Physiology

ORIGINAL RESEARCH
published: 01 September 2021
doi: 10.3389/fphys.2021.732244

Estimation of Subglottal Pressure, Vocal Fold Collision Pressure, and Intrinsic Laryngeal Muscle Activation From Neck-Surface Vibration Using a Neural Network Framework and a Voice Production Model

*Emiro J. Ibarra[1,2], Jesús A. Parra[1], Gabriel A. Alzamendi[3], Juan P. Cortés[1,4], Víctor M. Espinoza[5], Daryush D. Mehta[4], Robert E. Hillman[4] and Matías Zañartu[1]**

[1] Department of Electronic Engineering, Universidad Técnica Federico Santa María, Valparaíso, Chile, [2] School of Electrical Engineering, University of the Andes, Mérida, Venezuela, [3] Institute for Research and Development on Bioengineering and Bioinformatics, Consejo Nacional de Investigaciones Científicas y Técnicas - Universidad Nacional de Entre Ríos, Oro Verde, Argentina, [4] Center for Laryngeal Surgery and Voice Rehabilitation Laboratory, Massachusetts General Hospital–Harvard Medical School, Boston, MA, United States, [5] Department of Sound, Faculty of Arts, University of Chile, Santiago, Chile

OPEN ACCESS

Edited by:
Michael Döllinger, University Hospital Erlangen, Germany

Reviewed by:
Wenjun Kou, Northwestern University, United States
Xudong Zheng, University of Maine, United States

***Correspondence:**
Matías Zañartu
matias.zanartu@usm.cl

Specialty section:
This article was submitted to Computational Physiology and Medicine, a section of the journal Frontiers in Physiology

Received: *28 June 2021*
Accepted: *09 August 2021*
Published: *01 September 2021*

Citation:
Ibarra EJ, Parra JA, Alzamendi GA, Cortés JP, Espinoza VM, Mehta DD, Hillman RE and Zañartu M (2021) Estimation of Subglottal Pressure, Vocal Fold Collision Pressure, and Intrinsic Laryngeal Muscle Activation From Neck-Surface Vibration Using a Neural Network Framework and a Voice Production Model. Front. Physiol. 12:732244. doi: 10.3389/fphys.2021.732244

The ambulatory assessment of vocal function can be significantly enhanced by having access to physiologically based features that describe underlying pathophysiological mechanisms in individuals with voice disorders. This type of enhancement can improve methods for the prevention, diagnosis, and treatment of behaviorally based voice disorders. Unfortunately, the direct measurement of important vocal features such as subglottal pressure, vocal fold collision pressure, and laryngeal muscle activation is impractical in laboratory and ambulatory settings. In this study, we introduce a method to estimate these features during phonation from a neck-surface vibration signal through a framework that integrates a physiologically relevant model of voice production and machine learning tools. The signal from a neck-surface accelerometer is first processed using subglottal impedance-based inverse filtering to yield an estimate of the unsteady glottal airflow. Seven aerodynamic and acoustic features are extracted from the neck surface accelerometer and an optional microphone signal. A neural network architecture is selected to provide a mapping between the seven input features and subglottal pressure, vocal fold collision pressure, and cricothyroid and thyroarytenoid muscle activation. This non-linear mapping is trained solely with 13,000 Monte Carlo simulations of a voice production model that utilizes a symmetric triangular body-cover model of the vocal folds. The performance of the method was compared against laboratory data from synchronous recordings of oral airflow, intraoral pressure, microphone, and neck-surface vibration in 79 vocally healthy female participants uttering consecutive /pæ/ syllable strings at comfortable, loud, and soft levels. The mean absolute error and root-mean-square error for estimating the mean subglottal pressure were 191 Pa (1.95 cm H_2O) and 243 Pa (2.48 cm H_2O), respectively, which are comparable with previous studies but with the key advantage of not requiring subject-specific training

and yielding more output measures. The validation of vocal fold collision pressure and laryngeal muscle activation was performed with synthetic values as reference. These initial results provide valuable insight for further vocal fold model refinement and constitute a proof of concept that the proposed machine learning method is a feasible option for providing physiologically relevant measures for laboratory and ambulatory assessment of vocal function.

Keywords: ambulatory monitoring, neck-surface accelerometer, subglottal pressure estimation, voice production model, neural networks, clinical voice assessment

1. INTRODUCTION

Laryngeal voice disorders have been estimated to affect approximately 30% of the adult population in the United States at some point in their lives (Bhattacharyya, 2014). Voice disorders can disrupt or preclude normal oral communication and thus have far-reaching social, professional, economic, and personal consequences for those affected. The most common voice disorders are associated with detrimental patterns of daily vocal behavior and voice use (often classified as vocal hyperfunction) for which there is limited understanding of the underlying etiological and pathophysiological mechanisms. The paucity of such information serves to hinder the effective prevention, diagnosis and treatment of these common voice disorders.

Ambulatory voice monitoring using a neck-placed accelerometer (ACC) provides the capability to quantitatively assess daily vocal function and has also been shown to have the potential to assist in modifying vocal behaviors via ambulatory biofeedback (Popolo et al., 2005; Hillman and Mehta, 2011; Mehta et al., 2012; Andreassen et al., 2017; Van Stan et al., 2017a). Numerous features have been extracted from the ambulatory recording of the ACC signal, including phonation duration, sound pressure level (SPL), fundamental frequency (f_o) (Ghassemi et al., 2014), vocal vibration-dose measures (Titze et al., 2003; Titze and Hunter, 2015), spectral and cepstral measures (Mehta et al., 2015, 2019), and aerodynamic measures (Llico et al., 2015; Cortés et al., 2018). These measures have been used to differentiate the daily voice use of patients with vocal hyperfunction from matched controls (Ghassemi et al., 2014; Cortés et al., 2018; Van Stan et al., 2021) and to track changes related to surgical and voice therapy treatment of hyperfunctional voice disorders (Van Stan et al., 2017b, 2020). Current classification accuracy using these parameters is in the range of 0.7–0.85.

We argue that the extraction of additional physiological measures from ambulatory ACC recordings, such as subglottal pressure, vocal fold collision pressure, and laryngeal muscle activation, would provide critical additional insights into the etiologic and pathophysiological mechanisms that underlie hyperfunctional voice disorders and thus significantly enhance the capability to identify the detrimental daily patterns of vocal behavior associated with these disorders (Espinoza et al., 2017; Galindo et al., 2017; Hillman et al., 2020). There have been recent efforts to develop subject-specific representations that can capture such physiologically relevant measures (e.g., subglottal pressure, contact pressure, muscle activation, and material properties of the vocal folds) that are difficult to obtain directly (Deng et al., 2019; Hadwin et al., 2019; Alzamendi et al., 2020; Drioli and Foresti, 2020). These approaches take advantage of the physiological relevance of lumped and finite element models of voice production, which have been shown to be useful tools for the investigation, diagnosis, and treatment of voice disorders (Erath et al., 2013). The most recent *in vivo* approach uses a Bayesian framework to estimate lumped-element vocal fold model parameters to predict subglottal pressure, vocal fold collision pressure, and laryngeal muscle activation along with their corresponding confidence intervals from observations obtained in clinical recordings, i.e., high-speed videoendoscopy (HSV) and oral airflow (Alzamendi et al., 2020).

Direct application of Bayesian subject-specific estimation from the ACC signal remains unsolved. There are challenges associated with the current extended Kalman filter approach for processing ambulatory data and using the ACC as the solely observation that remain to be addressed, including the large computational cost for the volume of data to be processed, the need for data fusion from different recording sessions, the need for an online estimation of model covariance, and the incorporation of a time-domain neck skin model for the ACC sensor within the voice production model.

On the other hand, machine learning and artificial intelligence are becoming relevant tools in biomedical engineering, as they can provide accurate predictions and efficient implementations. Numerical models are attractive alternatives for training purposes, suitable representing a significant range of conditions and providing access to relevant measures that are difficult to obtain experimentally. Voice assessment is starting to make use of these modeling advantages, where machine learning methods have been trained using simulated data from physiological numerical models to predict clinical parameters of interest. This approach was utilized by Gómez et al. (2019) to predict subglottal pressure from HSV in excised porcine vocal folds and by Zhang (2020) to predict vocal fold (geometric and mechanical) properties and subglottal pressure from a microphone signal. No machine learning method trained with a voice production model has been devised for the ACC signal in a laboratory or ambulatory context.

Although there are ongoing efforts to address the challenges of the Bayesian framework for the ambulatory monitoring, we propose in this study a more direct solution for the estimation of ambulatory physiologically-based features from

the ACC that uses machine learning and voice modeling tools. Thus, we propose a method to obtain a non-linear optimal mapping between ACC features and subglottal pressure, vocal fold collision pressure, and laryngeal muscle activation. We propose using the impedance based inverse filtering (IBIF) algorithm (Zañartu et al., 2013; Cortés et al., 2018), which yields an unsteady glottal airflow signal from the ACC signal, to provide aerodynamic features that are used as inputs to the non-linear mapping. At the same time, we propose using a neural network (NN) regression architecture trained from a physiologically relevant muscle-controlled voice synthesizer with a triangular body-cover vocal fold model (Alzamendi et al., 2019, 2021) that takes the aerodynamic features as input and provides subglottal pressure, collision pressure, and laryngeal muscle activation as output. Predictions obtained with this scheme are validated against numerical simulations and laboratory measurements of subglottal pressure. The contributions of this work are twofold: First, the proposed scheme provides access, for the first time, to various physiologically relevant model-based features from a neck-surface accelerometer signal. Then, the approach provides a comprehensive contrast of the selected voice production model against laboratory data.

2. MATERIALS AND METHODS

Figure 1 provides an overall schematic of the proposed method of estimating four vocal function measures from neck-surface vibration recorded using a neck-surface accelerometer (ACC) sensor. The first analysis block results in an estimate of the unsteady glottal airflow volume velocity signal using the IBIF model (Zañartu et al., 2013), which has been shown to provide aerodynamic features reliably for the classification of vocal hyperfunction in laboratory (Espinoza et al., 2020) and ambulatory (Cortés et al., 2018) settings. The second analysis block computes the following six features from the glottal airflow signal: amplitude of the unsteady glottal airflow (ACFL), maximum flow declination rate (MFDR), open quotient (OQ), speed quotient (SQ), spectral tilt measured as the log-magnitude difference between the first and second harmonics ($H_1 - H_2$), and fundamental frequency (f_o). A seventh feature—the sound pressure level (SPL)—can be estimated either directly using an acoustic microphone (MIC) in the laboratory setting or using a log-log mapping between the root-mean-square magnitude of the ACC signal and SPL (Švec et al., 2005). See **Table 1** for descriptions of each feature. These seven features are used as input into a NN to estimate four desirable measures of vocal function: subglottal pressure (P_s), vocal fold collision pressure (P_c), and normalized activation levels of the cricothyroid (a_{CT}) and thyroarytenoid (a_{TA}) muscles.

The NN was trained using 13,000 Monte Carlo simulations of a numerical voice production model. The design of the network architecture and overall training description are provided in section 2.1, and the details of the numerical voice production model are found in section 2.2. Validation of the estimated output features were performed using *in vivo* laboratory reference measures of P_s or numerical simulations of phonation for reference measures of P_c, a_{CT}, and a_{TA}. Details of the experimental validation are provided in section 2.3.

2.1. Neural Network Architecture and Training

A supervised machine learning framework for regression was implemented based on a multi-layer NN (Hagan et al., 2014). The network consisted of an input layer of the seven aerodynamic and acoustic features (ACFL, MFDR, OQ, SQ, $H_1 - H_2$, f_o, and SPL), an output layer composed of the four target vocal function measures (P_s, P_c, a_{TA}, and a_{CT}), and two interconnected hidden layers with a 10% dropout to avoid overfitting. Each neuron within the hidden layers had adjustable weight and bias parameters that combined with the outputs of the preceding layer to activate a rectified linear unit activation function; then, the resulting activation served as input for the next layer Bianco et al. (2019). The number of neurons for each layer was investigated as a function of the model performance against both numerical and experimental data. The training stage updates the weights and biases using the Adam optimization algorithm (Kingma and Ba, 2017) with a learning rate of 0.001. All the NNs involved in this work were implemented in a virtual machine from Google Colaboratory with two CPU models Intel(R) Xeon(R) CPU @ 2.00GHz, using Python 3.7.11. and the TensorFlow 2.5.0 library (Abadi et al., 2015). The runtime for the largest network (8 hidden layer with 128 neurons and 100 epoch) was less than 120 s.

The NN regression models were trained following the scheme shown in **Figure 2**. For this purpose, a synthetic voice dataset was obtained with a numerical voice production model described in section 2.2. Similar approaches were recently taken by other authors using different sensing modalities, i.e., high-speed videoendoscopy (Gómez et al., 2019) and MIC sensors (Zhang, 2020) in *ex vivo* experimental validation platforms (instead of *in vivo*). Using synthetic data for training helped addressing the lack of comprehensive and massive *in vivo* human datasets with thousands or even millions of conditions. Testing of the NN models was performed with both numerical and *in vivo* laboratory datasets. The laboratory dataset is described in section 2.3.

The voice production model described in section 2.2 was used to create 110,000 Monte Carlo simulations of sustained phonation. The simulations included a wide variation of the model control parameters such as lung pressure (P_L), activation levels for the cricothyroid (a_{CT}), thyroarytenoid (a_{TA}), lateral cricoarytenoid (a_{LCA}), interarytenoid (a_{IA}), and posterior cricoarytenoid (a_{PCA}) muscles. Control model parameters and their variation range are shown in **Table 2**. Each simulation lasted 800 ms, with the mean value of the seven input features taken for the last 50 ms to avoid transient artifacts. The glottal airflow was filtered using the same low- and high-pass filters utilized in the analysis of the laboratory recordings, as described in section 2.3.

As suggested by Gómez et al. (2019), the training data resembled the empirical distribution of the population-based aerodynamic and acoustic feature set. Thus, simulated data with ACFL less than 30 mL/s and f_o outside the range of 120–400 Hz were discarded, as these cases were not found in the

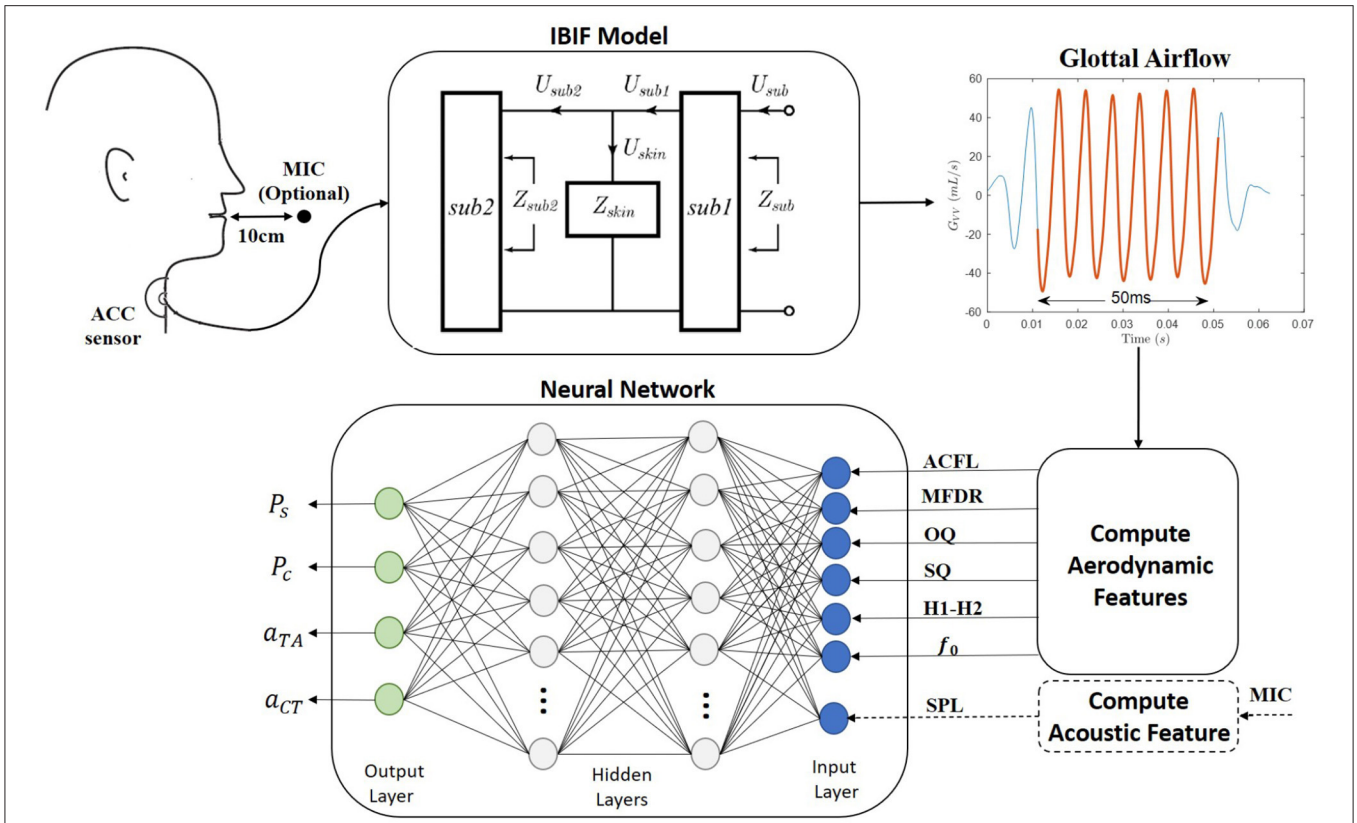


FIGURE 1 | A schematic of the proposed method for the ambulatory vocal assessment based on processing the neck skin acceleration signal and a regression neural network.

TABLE 1 | Description of aerodynamic features extracted from the glottal airflow signal and acoustic sound pressure level extracted from the microphone or accelerometer signal.

Feature	Description	Units
ACFL	The difference between the maximum and minimum amplitude of the AC glottal airflow (peak-to-peak) within each glottal cycle	mL/s
MFDR	Maximum flow declination rate: Negative peak of the first derivative of the glottal waveform	L/s^2
OQ	Open quotient: Ratio of the open time of the glottal vibratory cycle to the corresponding cycle period. Computed as in Cortés et al. (2018)	%
SQ	Speed quotient: Ratio of the opening time of the glottis to the closing time. Computed as in Cortés et al. (2018)	–
$H_1 - H_2$	Difference between the magnitude of the first two harmonics	dB
f_o	Fundamental frequency	Hz
SPL	Sound pressure level: dB from the RMS envelope of the acoustic signal	dB SPL

laboratory data used for testing the NN. As a result, the final synthetic dataset consisted of 13,000 samples. **Figure 3** shows the normalized histogram of features for the synthetic data (blue color) and laboratory data (red color).

Notice that feature ranges and distributions for both clinical and synthetic data sets agree, except for SPL and P_s, where ranges are noticeable dissimilar (see histograms for attenuated red color). Two bias corrections were considered for these components. First, as the SPL for the voice production model is obtained at the lips, the SPL value was corrected to match the 10 cm mouth-to-microphone recording distance considered in the clinical recordings, yielding a -28.5 dB correction factor (Švec and Granqvist, 2018). In addition, histograms of P_s suggest that the physiological voice synthesizer yields higher values for this measure. It is possible that sub and supra glottal tract propagation losses and the losses at the glottal boundary were not sufficiently high, thus amplifying source-filter interactions and raising up subglottal pressure. This bias has motivated subsequent exploration and model developments. However, to address the need to correct for the difference in P_s in this study,

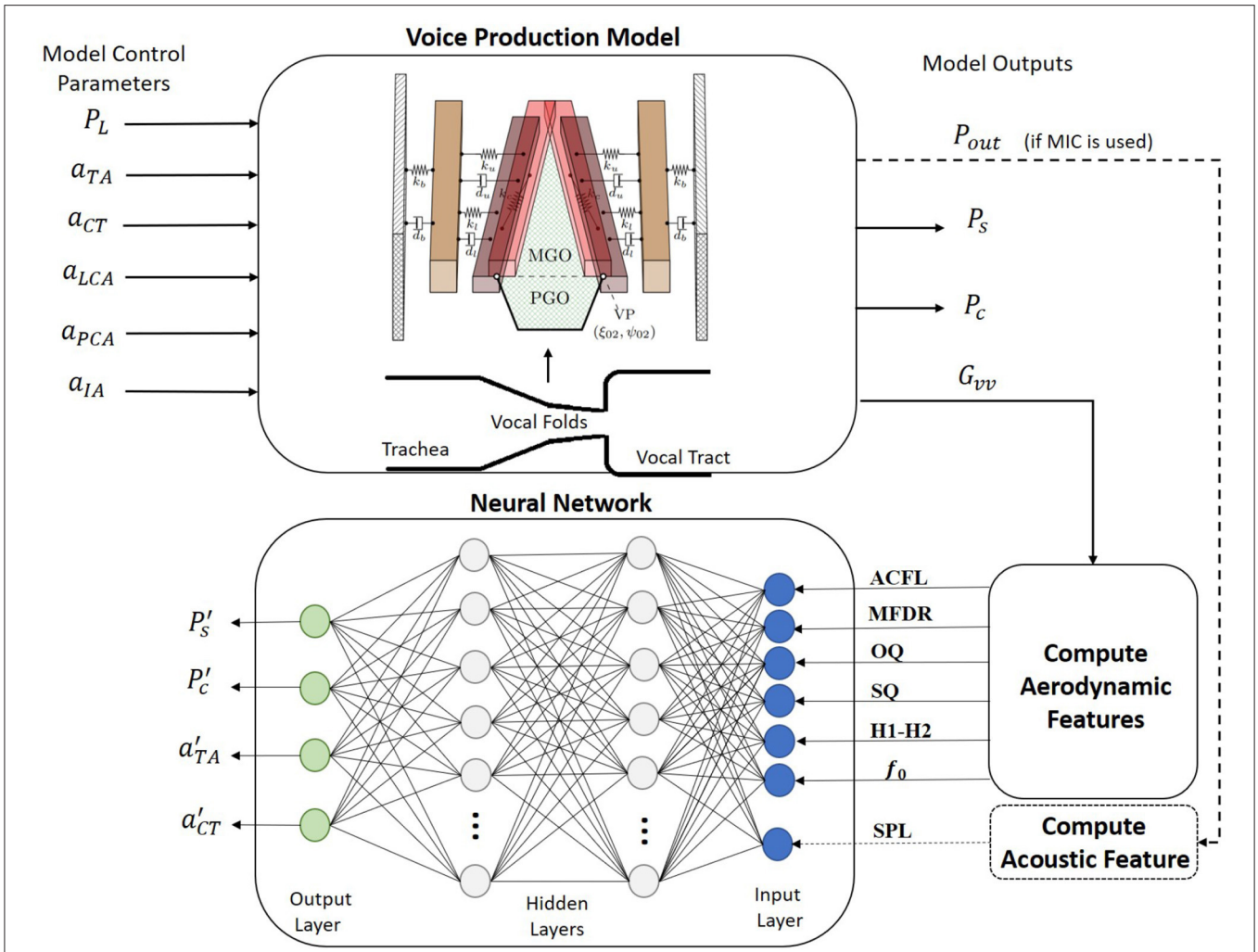


FIGURE 2 | A schematic for the proposed training procedure. A regression neural network is built for mapping accelerometer-based vocal features into clinically relevant estimates for subglottal pressure, subglottal collision pressure, and laryngeal muscle activation levels of the thyroarytenoid (TA) and cricothyroid (CT) muscles. Training data are produced from a numerical voice production model.

TABLE 2 | Range and increment step for control parameters in the numerical voice production model considered for building the synthetic dataset.

Parameter	Range	Step	Unit
a_{CT}	0-1	0.1	–
a_{TA}	0-1	0.1	–
a_{LCA}	0.2-0.8	0.1	–
a_{PCA}	0-0.1	0.1	–
a_{IA}	0.2-0.8	0.1	–
P_L	500 – 2000	150	Pa

a bias correction was applied by taking the differences between the mean of clinical and synthetic P_s values, thus leading to a -3.37 cm H_2O offset.

Synthetic training data were min-max normalized and selected randomly from 80% of the total simulations. Testing was performed in the remaining 20% of synthetic data and in the clinical data in order to identify the models providing the best estimation of subglottal pressure. To assess the regression performance during both training and validation stages and to compare with prior studies (Gómez et al., 2018; Lin et al., 2020; Zhang, 2020), the mean absolute error (MAE) and the root-mean-squared error (RMSE) metrics were utilized.

Several NN architectures with different numbers of neurons in the hidden layers were trained for two cases. Case I included six glottal aerodynamic features, described in **Table 1** (ACFL, MFDR, OQ, SQ, f_o, and $H_1 - H_2$) as input layer to the NNs, i.e., glottal measures extracted only from IBIF. Case II had the input layer of the NNs composed by all seven features in **Table 1**.

2.2. Voice Production Model

The selected voice production model for the training stage is a multi-physics scheme featuring a low-order model of the vocal

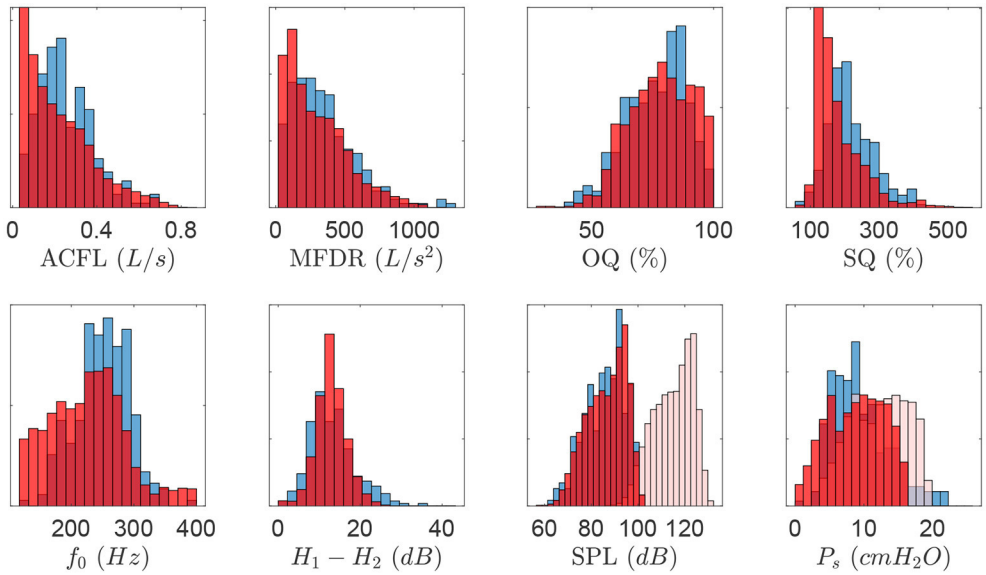


FIGURE 3 | Normalized histogram describing vocal features obtained for all measured quantities in the clinical data set (blue color). Resulting histograms for the synthetic dataset (red color) are superimposed to illustrate model matching. Bias correction for synthetic SPL and Ps are shown as additional histograms (light red color).

folds that allows for the coordinated activation of all five intrinsic laryngeal muscles (Alzamendi et al., 2019, 2021). The model was recently developed and was chosen due to its flexibility and physical and physiological relevant way to cover numerous (normal and disordered) phonatory conditions. The approach builds upon prior efforts that describe rules for controlling low-order models (Titze and Story, 2002), vocal fold posturing (Titze and Hunter, 2007), and a triangular body-cover vocal fold model (Galindo et al., 2017). The model also accounts for tissue-fluid-acoustic interactions at the glottis (Zañartu et al., 2014), sound wave propagation through the vocal tract followed by sound pressure radiated from mouth (Zañartu, 2006), and allows for describing sustained vowels and time-varying glottal gestures. Given that the model is fairly new, we describe its main components pertaining to the development of the NN regression model and training set.

The triangular body-cover model (TBCM) (Galindo et al., 2017) (see **Figure 2**) consists of paired three-mass body-cover systems interconnected with mechanical elements (Story and Titze, 1995) and configured in a triangular anatomical shape (Birkholz et al., 2011). Beside resembling the triangular glottis, the TBCM is physiologically relevant because it mimics the layered vocal fold structure and extends the vocal fold collision model with a gradual zipper-like incomplete glottal closure. The latter aided to describe the time-varying vocal fold collision pressure (P_c) during phonation. Similar to Galindo et al. (2017), the parameterization of the TBCM followed the original body-cover description (Story and Titze, 1995) and applied the empirical rules to change geometrical and viscoelastic vocal fold parameters developed by Titze and Story (2002). However, the major difference with (Galindo et al., 2017) resided in the computation of both the internal tension and elongation in the vocal folds. The remaining rules in Titze and Story (2002) were taken as originally proposed for deriving the lumped-element dynamical parameters.

Given the interest in estimating intrinsic laryngeal activity with the proposed method, a comprehensive description of muscle activity on the laryngeal configuration was considered. For this purpose, the contributions of all five intrinsic muscles and the passive response of connective tissue (i.e., the vocal ligament and vocal fold mucosa) were included in the model. Hence, simulated laryngeal muscle activations were the control variables governing the phonatory posture and vocal fold elongation. The incorporation of this muscle-controlled model of the larynx allowed to dynamically modify the glottal function during phonation, e.g., the vocal fold oscillatory dynamics, time-varying glottal resistance, and aerodynamic-acoustic coupling mechanisms. Following Titze and Hunter (2007), the five intrinsic muscles were modeled independently by using a modified Kelvin model (Hunter et al., 2004), which dynamically solves for the internal stress-strain response in one-dimensional fibrous tissues by integrating both active and passive properties. Passive stress was described as a non-linear function of longitudinal strain. Additionally, the active stress resulted from the maximum isometric active stress and the normalized activation level, in the range $0 \leq a \leq 1$, mapping from relaxed to strong muscle tension. The simulated muscle activation for each intrinsic muscle was thus adjusted trough the corresponding activation levels $\{a_{\mathrm{LCA}}, a_{\mathrm{IA}}, a_{\mathrm{PCA}}, a_{\mathrm{CT}}, a_{\mathrm{TA}}\}$. In the TBCM, an adducted glottal configuration is critical for setting the system into self-sustained oscillations, thus requiring higher activation

of the adductory intrinsic (LCA and IA) musculature than the abductory intrinsic (PCA) musculature. For simplicity, we did not consider the effects of elevated antagonistic muscles (Alzamendi et al., 2021) and only explored a small range of PCA activation to secure self-sustained oscillations in the TBCM. This approach allowed us to reduce the number of simulations to be discarded and to optimize the computational load. Future investigations will involve further scenarios for muscle control in typical and disordered phonation. Models for the vocal ligament and vocal fold mucosa were similarly implemented, except that the active component was set to zero for these cases (Titze, 2006).

Beside controlling intrinsic muscle activation, the voice production model also allowed for the adjustment of the aerodynamic lung pressure, P_L. The aerodynamic forces acting over the vocal fold cover layer were then computed from the resulting subglottal pressure, P_s, and supraglottal pressure, P_e, according to Titze (2002). The three-way interaction at the glottal level between sound, flow, and vocal fold tissue was included, whereas the glottal airflow was computed from the acoustic driving pressures impinging on the glottal (membranous plus posterior portions) area following (Titze, 2006; Zañartu et al., 2014; Lucero and Schoentgen, 2015). Simulation of the time-varying acoustic wave propagation was achieved by applying the wave reflection analog scheme, where the subglottal and supraglottal tracts were modeled as a discrete concatenation of short uniform acoustic cylinders with variable cross-sectional areas. Effects due to the boundary condition at the lips was approximated by including an inertive radiation impedance, that produces the reflected pressure wave and the radiated sound wave, P_{out}. Losses due to viscosity, moving walls, and other losses are described by an exponential attenuation factor in the propagation through the cylindrical sections (Zañartu, 2006; Zañartu et al., 2007). Vocal tract area functions that resemble a typical male (Story, 2008) and female (Story et al., 1998) that could match the *in vivo* experimental data were selected, i.e., vowels /æ/ and /ɑ, along with a representative subglottal tract (Zañartu et al., 2014).

2.3. Experimental Validation of NN-Estimated Subglottal Pressure

An *in vivo* laboratory dataset (Mehta et al., 2015; Espinoza et al., 2017, 2020) with synchronous recordings of intraoral pressure (IOP), oral airflow volume velocity (OVV), MIC, and ACC from vocally healthy subjects was utilized to provide a completely separate testing platform for the estimates obtained with the regression NN. This dataset was used to experimentally validate the NN estimates of subglottal pressure. Direct measurements of vocal fold collision pressure and laryngeal muscle activation are difficult to obtain in the laboratory and were not included in this experimental validation. Note that this dataset was not used to train the NN.

The data correspond to a group of participants composed of 79 adult females with no history of voice disorders. The mean (SD) age was 29.6 (13.0) years old. Their vocally healthy status was verified by a licensed speech-language pathologist via interview (subjects reported no difficulties with their voices in daily life), laryngeal videostroboscopic examination, and a clinician-administered Consensus Auditory-Perceptual Evaluation of Voice (CAPE-V) assessment (Kempster et al., 2009). Informed consent was obtained from all the participants in this study, and experimental and clinical protocols were approved by the institutional review board of Mass General Brigham (formerly Partners HealthCare) at the Massachusetts General Hospital. Data recordings were conducted in a sound-treated room where study staff instructed each participant to repeat strings of /pæ/ syllables in three loudness conditions (comfortable, loud, and soft). Although subjects were instructed to maintain a constant pitch and loudness within each syllable string, they were free to choose levels that were most natural for them without any prescribed levels of absolute pitch and loudness.

Recordings consisted of the simultaneous acquisition of acoustic pressure obtained with a condenser MIC (MKE104, Sennheiser, Electronic GmbH, Wedemark, Germany) placed 10 cm from the lips and having full bandwidth in the range of 0–6 kHz, OVV sensed by using a circumferentially vented pneumotachograph mask (PT-2E, Glottal Enterprises, Syracuse, NY) with a bandwidth of approximately 1.1 kHz, IOP measured with an oral catheter passed between the lips and connected to a low-bandwidth pressure sensor with an effective bandwidth of approximately 80 Hz (Espinoza et al., 2017), and ACC (BU-27135; Knowles Corp., Itasca, IL, USA) placed on the anterior neck surface halfway between the thyroid prominence and the suprasternal notch (Zañartu et al., 2013). All signals were sampled at 20 kHz/16 bits (Digidata 1440A, Axon Instruments, Inc.), low-pass filtered at 8-kHz cutoff frequency (CyberAmp Model 380, Axon Instruments, Inc.), and calibrated to physical units (Espinoza et al., 2017).

Signals obtained from the ACC and pneumotachograph mask were low-pass filtered at 1,100 Hz with a 10th-order Chebyshev Type II filter and decimated to 8,192 Hz. Then, a fourth-order Butterworth high-pass filter with cutoff frequency at 60 Hz was used to remove low-frequency components. The IOP signal was low-pass filtered at 80 Hz with a fifth-order Butterworth filter and then decimated to 256 Hz sample rate. All filters were applied with phase removal to avoid phase distortion (Perkell et al., 1994).

Reference values for subglottal pressure were obtained from IOP signals following (Espinoza et al., 2017). Driving pressure was extrapolated as the mean value of the two consecutive IOP plateaus produced by the combined lip closure and glottis opening prior to the /p/ sounds, that produced just before and after each vowel segment. The three middle syllables in each /pæ/ string were selected for the analysis, so that the initial and final portions were disregarded to avoid any evident transient dynamics. The estimated subglottal pressure was the average of these three-syllable values. Three reference measures per participant for comfortable, loud, and soft loudness conditions were obtained. Thus, a total of 237 /pæ/ tokens were used in this study.

The OVV-based glottal airflow was obtained through a common inverse filtering technique based on a single-notch filter with a conjugate pair of zeros and unity gain at DC at first vocal

tract resonance (Perkell et al., 1991; Cheyne, 2006). Each single-notch filter was applied to a 50 ms stable portion of the middle /pæ/ string. The center frequency of the filter was determined following an optimization procedure developed by Espinoza et al. (2017).

The ACC-based glottal airflow was estimated using the IBIF scheme (Zañartu et al., 2013; Cortés et al., 2018). This method uses an acoustic transmission line model and a calibration step to obtain a set of subject-specific parameters corresponding to the neck-skin surface, length of the trachea, and accelerometer position (Zañartu, 2010; Zañartu et al., 2013; Cortés et al., 2018). These parameters are determined by minimizing the waveform error between the OVV-based glottal airflow (reference signal described previously) and the inverse filtered neck-skin ACC signal via a particle swarm optimization Kennedy and Eberhart (1995). The middle 50 ms of the glottal airflow signal estimated from IBIF was selected to compute the six acceleration-based aerodynamic feature (see **Table 1**). Even though SPL can be computed from the ACC signal using regression methods Švec et al. (2005), the synchronous microphone signal was used in this study to avoid introducing any additional estimation error at this point. Future work can be devoted to enhance current linear mapping between ACC and SPL.

Validation with human data is the gold standard to assess the ability of the NN regression scheme to represent *in vivo* data; but direct measurement of certain physiological measures of vocal function is not feasible. An advantage of using a voice production model to train a neural network is that we can estimate vocal function measures that are difficult to measure in practice, which is the case for vocal fold collision pressure and intrinsic muscle activation. Thus, the assessment of the estimates of subglottal pressure is described in terms of test sets from numerical simulations and laboratory data, whereas the estimates of vocal fold collision pressure and laryngeal muscle activation are only evaluated using a synthetic data test set.

TABLE 3 | MAE and RMSE between the estimated P_s with the proposed NN regression model and the reference measures from synthetic and laboratory test data.

Neurons in each hidden layer	Number of hidden layers	Synthetic Data MAE (cm H_2O)	Synthetic Data RMSE (cm H_2O)	Laboratory Data MAE (cm H_2O)	Laboratory Data RMSE (cm H_2O)
Case I:					
4	2	1.98	2.51	2.23	2.82
8	2	1.81	2.34	2.28	2.86
16	2	1.35	1.83	2.56	3.13
32	2	1.18	1.64	2.82	3.43
64	2	1.02	1.48	2.89	3.50
128	2	0.99	1.68	2.94	3.58
128	4	0.93	1.33	3.17	3.87
128	6	0.97	1.38	3.14	3.85
128	8	1.01	1.45	3.12	3.76
Case II:					
4	2	1.84	2.42	1.95	2.48
8	2	1.87	2.43	1.97	2.52
16	2	1.27	1.74	2.42	2.98
32	2	1.13	1.58	2.55	3.17
64	2	0.99	1.42	2.88	3.45
128	2	0.90	1.30	2.98	3.58
128	4	0.78	1.12	3.23	3.87
128	6	0.87	1.21	3.04	3.71
128	8	1.00	1.38	3.08	3.70

Errors are reported for different NN architecture (different number of neurons and hidden layers). Case I: Input aerodynamic features of ACFL, MFDR, OQ, SQ, f_o, and $H_1 - H_2$. Case II: Input aerodynamic features in Case I and acoustic SPL.

3. RESULTS

3.1. Subglottal Pressure Estimation

The MAE and RMSE describing P_s estimates for the different architectures are reported in **Table 3** for both synthetic and clinical test data. For both cases I and II, additional hidden layers and neurons per layer yielded an improvement in subglottal pressure estimation when tested against the synthetic data. For example, in case I, MAE decreased from 1.98 cm H_2O to 0.93 cm H_2O from the simplest (2 hidden layers with 4 neurons) to a more complex (4 hidden layers with 128 neurons) architecture, respectively. In case II, MAE decreased from 1.84 cm H_2O to 0.78 cm H_2O for the same prior complexity in the NN architecture. This represents a reduction of more than 50% in MAE in both cases. A similar trend is observed with RMSE. An explanation for the improvement comes from the fact that the training and testing data were obtained from the same voice production model. Therefore, more complex NN models appear to capture efficiently the non-linear mechanisms of the model, which has been suggested by Zhang (2020), when training and testing with synthetic data from the same model. However, for the NN architectures composed over the six hidden layer with 128 neurons, the MAE and RMSE for synthetic data increase, showing that a deeper NN does not improve the estimation of subglottal pressure in this context.

It is important to highlight that all NNs were trained using 100 epochs. This criterion was selected to ensure the convergence of models. **Figure 4** shows mean squared error vs. the epochs for training and validation of the simplest and the most complex architecture models. The curves illustrate the convergence of the training procedure, where the simplest regression model exhibits a more rapid convergence. However, at around 100 epochs, the error remains constant, as the training progresses for both architectures. A similar trend was observed for all tested configurations. Another element to highlight is the absence of overfitting, since the training and validation error monotonically decrease at the same time. This shows that the network learns the structure of the observed data and is able to infer the validation data. An indication of overfitting would be a training error that decreases while the validation error remains the same or increases.

On the other hand, for the laboratory validation of subglottal pressure, we found the opposite trend for MAE as a function of the NN architecture complexity. In Case I, MAE increased from 2.23 cm H_2O to 3.17 cm H_2O for an increasing complexity

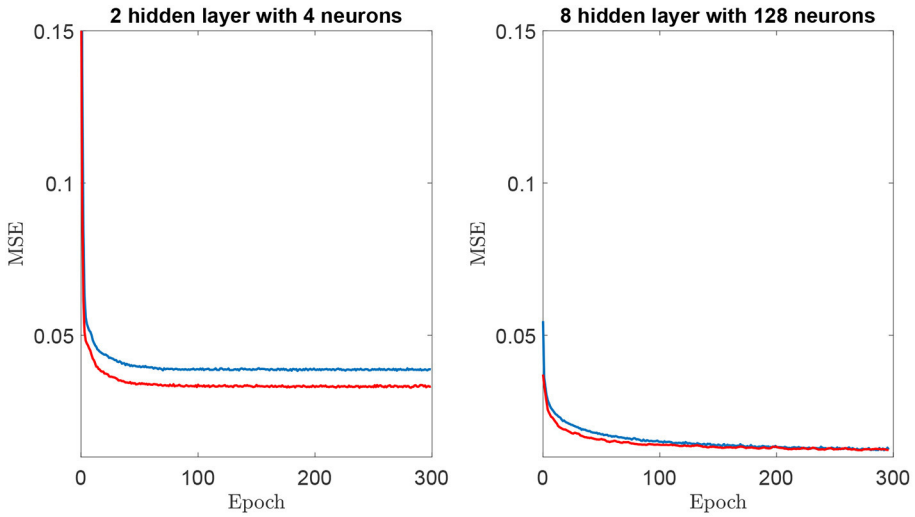


FIGURE 4 | Mean Squared Error (MSE) vs. epoch for training (blue color) and validation (red color) for two neural networks architectures. **(Left)** 2 hidden layers with 4 neurons. **(Right)** 8 hidden layers and 128 neurons.

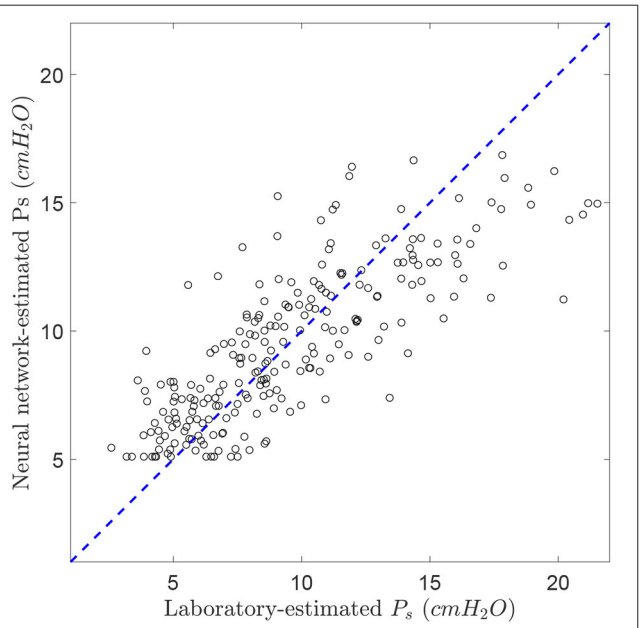


FIGURE 5 | Comparison between laboratory-estimated subglottal pressure and the corresponding estimates from the trained neural network (2-hidden layer, 4 neurons in each layer, and 7 voice features). $R^2 = 0.65$. The dashed line represents the theoretical 1:1 perfect matching.

from the 2 hidden layers with 4 neurons to 4 hidden layers with 128 neurons model. Case II also exhibited MAE increases from 1.95 cm H_2O to 3.23 cm H_2O for the same increasing complexity in the NN architecture. These results represented an increase of 42% and 66% in MAE for Case I and II, respectively, with similar trends for RMSE. Therefore, higher NN complexity was not adequate to represent sample distribution from the laboratory dataset.

Table 3 also illustrates that the inclusion of SPL in the input feature vector improves the estimation of subglottal pressure for all tested NN architectures. Using the best architecture for the laboratory validation, we found a 12% reduction in MAE and RMSE. The best architecture for the synthetic validation exhibited a 16% reduction in MAE and RMSE when SPL was added. These results are in agreement with previous studies (Titze et al., 2003; Björklund and Sundberg, 2016; Espinoza et al., 2017) that reported a strong correlation between subglottal pressure and acoustic SPL. Although not reported, no significant error differences were observed when estimating SPL from either the MIC or ACC sensor.

Therefore, the NN model with lowest error in the validation set from the laboratory data was selected from 4 neurons in the hidden layers and all seven input features. **Figure 5** shows a scatter plot of the NN-estimated subglottal pressure vs. the reference subglottal pressure from the laboratory data. The dashed line represents a 1:1 correspondence between the estimated and reference subglottal pressure. The coefficient of determination R^2 is 0.65 and the mean absolute percentage error is 24.9%. We highlight that even though the IOP data was used as ground truth for this assessment, differences in the subglottal pressure estimates from IOP and direct measurement of subglottal pressure via tracheal puncture has been reported in the range of 5% (Hertegård et al., 1995), although interpolation between the peaks of the pulses can lead to a 12% error (Rothenberg, 2013).

3.2. Vocal Fold Collision Pressure and Laryngeal Muscle Activation Estimation

Table 4 reports the coefficient of determination R^2, MAE (in physical units and in percentage of range) using synthetic data for the four outputs (P_s, P_c, a_{CT}, a_{TA}) obtained using the NN for the 2 hidden layers with 4-neuron and 4 the hidden layers with

TABLE 4 | Assessment of estimated vocal measures P_s, P_c, a_{TA}, and a_{CT} using the proposed NN regression method.

Parameters	Units	R^2	MAE (Units)	MAE (%)
2-HL and 4-N architecture:				
P_s	cm H_2O	0.64	1.84	11.4
P_c	cm H_2O	0.70	3.33	8.2
a_{TA}	-	0.07	0.21	21.1
a_{CT}	-	0.53	0.15	14.6
4-HL and 128-N architecture:				
P_s	cm H_2O	0.93	0.74	4.7
P_c	cm H_2O	0.92	1.70	4.2
a_{TA}	–	0.52	0.13	13.3
a_{CT}	–	0.84	0.07	7.1

Reported values for R^2 and MAE (in physical units and in percentage of range) for two NN architectures with different hidden layers (HL) and neurons (N). The input vector includes the seven aerodynamic measures.

128-neuron architectures. The architectures with more layers exhibited the best performance for estimating subglottal pressure for the synthetic data.

As seen before for the synthetic validation of subglottal pressure, increasing the complexity of the NN architecture increases the accuracy of the estimates. This performance holds true for the estimates of vocal fold collision pressure and laryngeal muscle activation. However, there is a significantly smaller R^2 of 0.52 for a_{TA} estimation when compared with $R^2 > 0.8$ for estimation of the other measures using the 4 hidden layer with 128 neurons NN. This finding suggests that certain measures, such as a_{TA}, require deeper, more complex NN architectures to achieve similar performance.

4. DISCUSSION

The purpose of this study was to explore the combination of neural network regression networks with a voice production model to estimate physiologically relevant vocal measures, i.e., subglottal pressure, vocal fold collision pressure, and (TA and CT) laryngeal muscle activation from a neck-surface vibration signal. Validation for this study was done both numerically and experimentally. Given that some of the predicted measures are difficult to obtain experimentally, only the estimates of subglottal pressure could be compared with reference estimates of mean subglottal pressure derived from the standard airflow interruption technique in the laboratory.

Both numerical and experimental validation experiments yielded reasonable accuracy. The robust and reliable estimates of the proposed method are dependent on the capacity of the selected voice production model to mimic the observed distributions in the laboratory data. As the architecture complexity of the NN increased, the estimation error decreased for the synthetic data but increases for the laboratory data. We argue that this is a result of the way the model was utilized, i.e., model parameters were swept across a large range of values, but no anatomical changes were considered; thus the model simply described a single subject for a range of conditions. This may be playing a role in the accuracy for the estimation of subglottal pressure because no inter-subject variability was considered. At the same time, we discarded cases that differed from the laboratory distributions during the training process and corrected for a bias in the estimation of subglottal pressure. In addition, it is possible that subglottal and supraglottal tract propagation losses and the losses at the glottal boundary were not high enough, thus amplifying the source-filter interactions and resulting subglottal pressure. Future efforts will be devoted to improve model development, better reflect population behaviors, and assess these effects in the predicted accuracy of the proposed approach. In spite of its simplicity and the aforementioned limitations, we still conclude that the triangular body cover model provides a good general representation of typical sustained phonation for a large range of subjects and conditions.

The predicted subglottal pressure in this study are comparable with those obtained in previous studies. Our lowest mean absolute error for estimated subglottal pressure from clinical data was 191 Pa (1.95 cm H_2O), whereas two relevant studies reported mean absolute errors of 194 Pa (Gómez et al., 2019) and 115 Pa (Zhang, 2020). However, it is important to highlight that our predictions are obtained from a neck-surface accelerometer and that we tested our predictions against *in vivo* human data, whereas these studies used porcine and human excised larynx experiments. Lin et al. (2020) estimated subglottal pressure from a neck-surface accelerometer using a subject-specific step-wise factorial regression model in 26 normal subjects. The investigators obtained an average root-mean-square error in the range of 2.4–2.5 cm H_2O, which is comparable with the root-mean-square error of 2.48 cm H_2O in the current study. The linear regression models in Lin et al. (2020) included cepstral peak prominence and fundamental frequency, along with ACC-based aerodynamic measures and were constructed on an individual basis for every subject across multiple elicited voice qualities. The main advantages of the NN model approach are the fact that a single, general non-linear regression mapping is utilized and that our mapping also provides estimates of other clinically relevant measures of vocal function. It is acknowledged that future work is required to experimentally validate these other measures of vocal fold collision pressure and laryngeal muscle activation.

Titze et al. (2003) put forth a simple, empirically derived formula (Equation 15) that computed subglottal pressure using only measurements of SPL and f_o. Applying this formula to the laboratory data (237 tokens) in the current study to estimate subglottal pressure resulted in a root-mean-square error of 2.86 cm H_2O and mean absolute error of 2.11 cm H_2O. The relatively good performance for such a simple formula supports the idea that simple regression architectures are adequate for predicting subglottal pressure in vocally typical conditions; estimation accuracy of linear regression models reduces when non-modal voice qualities are included (Marks et al., 2019, 2020; Lin et al., 2020). The model-based approach of the current work allows for the estimation of additional measures of vocal function (e.g., vocal fold collision pressure, laryngeal muscle activation).

On the other hand, the accuracy of estimates of muscle activation and collision pressure that was assessed against synthetic data was hampered by the simplicity of the rather shallow NN architecture that resulted from matching clinical data for subglottal pressure. When the complexity of the network is increased, the estimation of muscle activation and collision pressure improves. This result is encouraging to investigate the development of subject-specific models that can handle more complex neural network architectures without losing the ability to predict subglottal pressure.

These initial results constitute a proof of concept that the proposed NN method is a feasible option for estimating clinically relevant vocal measures that are difficult to directly measure in laboratory and ambulatory settings. Current results could be significantly improved by exploring different NN architectures, improving model development, training with ACC signal features directly (vs. model features), using subject-specific tuning with transfer learning instead of a generic training for all subjects, and including experimental validation of all predicted values. This study delineates a path for various subsequent research efforts in this direction.

The neck-surface accelerometer sensor can be worn by a speaker for laboratory, clinical, and ambulatory assessments of vocal function. The estimation of subglottal pressure was validated using sustained phonation datasets from numerical modeling and laboratory recordings. There is potential to translate this method into ambulatory settings due to the input of the network only needing accelerometer-based features for short-time windows of 50 ms in duration. We hypothesize that the physiologically relevant measures that are obtained with the proposed approach will yield salient measures of vocal function in real-world environments. We expect that the physiologically relevant measures that are obtained with the proposed approach will provide unique quantitatively based insights into the etiologic and pathophysiological mechanisms associated with daily voice use in patients with hyperfunctional voice disorders. The capability to link model outputs with clinical data is expected to produce more comprehensive and specific descriptions of aberrant phonatory mechanisms that will lead to better subclassification (phenotyping) of hyperfunctional voice disorders and ultimately improve the prevention, diagnosis, and treatment of these disorders.

5. CONCLUSION

A framework to estimate subglottal pressure, collision pressure, and muscle activation from a neck surface accelerometer is developed integrating machine learning tools and a numerical model of voice production. Aerodynamic measures estimated from the neck surface accelerometer are combined with a sound pressure level estimate obtained from either an accelerometer or a microphone, and are selected as inputs to a neural regression network. The non-linear mapping is trained solely with a low-order voice production model featuring a symmetric triangular body-cover model of the vocal folds. When compared with clinical recordings from 79 female vocally healthy participants, the mean absolute error and root mean square error for the subglottal pressure were 1.95 cm H_2O and 2.48 cm H_2O. These results are comparable with previous studies but with the advantage of having a general mapping for all patients and providing simultaneous estimates of collision pressure and muscle activation. However, given that clinical validation for these latter features is cumbersome, only synthetic data were used for that purpose, and experimental validation is left for future efforts. At the same time, relevant insights are gained by comparing the numerical model with the clinical data that will lead to further model refinements. The initial results constitute a proof of concept that the proposed machine learning method is a feasible option for providing highly relevant physical measures for the ambulatory assessment of voice. Future efforts will be focused on creating individualized mappings for normal and disordered voices with transfer learning and validating all the estimated features with *in vivo* recordings.

DATA AVAILABILITY STATEMENT

The datasets presented in this study can be found in online repositories. The names of the repository/repositories and accession number(s) can be found in the article/supplementary material.

ETHICS STATEMENT

The studies involving human participants were reviewed and approved by Institutional Review Board of Mass General Brigham. The patients/participants provided their written informed consent to participate in this study.

AUTHOR CONTRIBUTIONS

MZ conceived the study and contributed to data analysis and interpretation, supervision, and manuscript writing. EI and JP conducted numerical simulations, neural network design, training and testing, data analysis and interpretation, and manuscript writing. GA developed the triangular body-cover model and contributed to data analysis and interpretation, supervision, and manuscript writing. JC and VE contributed to data analysis and interpretation and manuscript writing. DM and RH contributed to laboratory data collection, results interpretation, clinical input, and review editing. All authors contributed to the article and approved submitted version.

FUNDING

This research was supported by the National Institutes of Health (NIH) National Institute on Deafness and Other Communication Disorders grant P50 DC015446, ANID grants BASAL FB0008, FONDECYT 1191369 and FONDECYT 11200665, Becas de Doctorado Nacional 21190074 and 21202490, and STIC AmSud ASPMLM-Voice 21-STIC-05, and the Voice Health Institute.

REFERENCES

Abadi, M., Agarwal, A., Barham, P., Brevdo, E., Chen, Z., Citro, C., et al. (2015). *TensorFlow: Large-Scale Machine Learning on Heterogeneous Systems*. Available online at: tensorflow.org

Alzamendi, G., Manríquez, R., Hadwin, P., Deng, J., Peterson, S., Erath, B., et al. (2020). Bayesian estimation of vocal function measures using laryngeal high-speed videoendoscopy and glottal airflow estimates: an *in vivo* case study. *J. Acoust. Soc. Am.* 147, EL434–EL439. doi: 10.1121/10.0001276

Alzamendi, G., Peterson, S., Erath, B., and Zaartu, M. (2019). "Updated rules for constructing a triangular body-cover model of the vocal folds from intrinsic laryngeal muscle activation," in *The 13th International Conference on Advances in Quantitative Laryngology, Voice and Speech Research* (Montreal, QC).

Alzamendi, G. A., Peterson, S. D., Erath, B. D., Hillman, R. E., and Zaartu, M. (2021). Triangular body-cover model of the vocal folds with coordinated activation of five intrinsic laryngeal muscles with applications to vocal hyperfunction. *arXiv preprint arXiv:2108.01115*.

Andreassen, M. L., Litts, J. K., and Randall, D. R. (2017). Emerging techniques in assessment and treatment of muscle tension dysphonia. *Curr. Opin. Otolaryngol. Head Neck Surg.* 25, 447–452. doi: 10.1097/MOO.0000000000000405

Bhattacharyya, N. (2014). The prevalence of voice problems among adults in the united states. *Laryngoscope* 124, 2359–2362. doi: 10.1002/lary.24740

Bianco, M. J., Gerstoft, P., Traer, J., Ozanich, E., Roch, M. A., Gannot, S., et al. (2019). Machine learning in acoustics: theory and applications. *J. Acoust. Soc. Am.* 146, 3590–3628. doi: 10.1121/1.5133944

Birkholz, P., Kröger, B. J., and Neuschaefer-Rube, C. (2011). "Synthesis of breathy, normal, and pressed phonation using a two-mass model with a triangular glottis," in *Interspeech 2011: 12th Annual Conference ofthe International Speech Communi- cation Association* (Florence), 2681–2684.

Björklund, S., and Sundberg, J. (2016). Relationship between subglottal pressure and sound pressure level in untrained voices. *J. Voice* 30, 15–20. doi: 10.1016/j.jvoice.2015.03.006

Cheyne, H. A. (2006). "Estimating glottal voicing source characteristics by measuring and modeling the acceleration of the skin on the neck," in *2006 3rd IEEE/EMBS International Summer School on Medical Devices and Biosensors*, Boston, MA, 118–121.

Cortés, J. P., Espinoza, V. M., Ghassemi, M., Mehta, D. D., Van Stan, J. H., Hillman, R. E., et al. (2018). Ambulatory assessment of phonotraumatic vocal hyperfunction using glottal airflow measures estimated from neck-surface acceleration. *PLoS ONE* 13:e0209017. doi: 10.1371/journal.pone.0209017

Deng, J. J., Hadwin, P. J., and Peterson, S. D. (2019). The effect of high-speed videoendoscopy configuration on reduced-order model parameter estimates by bayesian inference. *J. Acoust. Soc. Am.* 146, 1492–1502. doi: 10.1121/1.5124256

Drioli, C., and Foresti, G. L. (2020). Fitting a biomechanical model of the folds to high-speed video data through bayesian estimation. *Inform. Med. Unlocked* 20:100373. doi: 10.1016/j.imu.2020.100373

Erath, B. D., Zañartu, M., Stewart, K. C., Plesniak, M. W., Sommer, D. E., and Peterson, S. D. (2013). A review of lumped-element models of voiced speech. *Speech Commun.* 55, 667–690. doi: 10.1016/j.specom.2013.02.002

Espinoza, V. M., Mehta, D. D., Stan, J. H. V., Hillman, R. E., and Zañartu, M. (2020). Glottal aerodynamics estimated from neck-surface vibration in women with phonotraumatic and nonphonotraumatic vocal hyperfunction. *J. Speech Lang. Hear. Res.* 63, 2861–2869. doi: 10.1044/2020_JSLHR-20-00189

Espinoza, V. M., Zañartu, M., Stan, J. H. V., Mehta, D. D., and Hillman, R. E. (2017). Glottal aerodynamic measures in women with phonotraumatic and nonphonotraumatic vocal hyperfunction. *J. Speech Lang. Hear Res.* 60, 2159–2169. doi: 10.1044/2017_JSLHR-S-16-0337

Galindo, G. E., Peterson, S. D., Erath, B. D., Castro, C., Hillman, R. E., and Zañartu, M. (2017). Modeling the pathophysiology of phonotraumatic vocal hyperfunction with a triangular glottal model of the vocal folds. *J. Speech Lang. Hear. Res.* 60, 2452–2471. doi: 10.1044/2017_JSLHR-S-16-0412

Ghassemi, M., Van Stan, J. H., Mehta, D. D., Zañartu, M., Cheyne, H. A. II, Hillman, R. E., et al. (2014). Learning to detect vocal hyperfunction from ambulatory neck-surface acceleration features: initial results for vocal fold nodules. *IEEE Trans. Biomed. Eng.* 61, 1668–1675. doi: 10.1109/TBME.2013.2297372

Gómez, P., Schützenberger, A., Kniesburges, S., Bohr, C., and Dllinger, M. (2018). Physical parameter estimation from porcine *ex vivo* vocal fold dynamics in an inverse problem framework. *Biomech. Model Mechanobiol.* 17, 777–792. doi: 10.1007/s10237-017-0992-5

Gómez, P., Schützenberger, A., Semmler, M., and Döllinger, M. (2019). Laryngeal pressure estimation with a recurrent neural network. *IEEE J. Transl. Eng. Health Med.* 7, 1–11. doi: 10.1109/JTEHM.2018.2886021

Hadwin, P. J., Motie-Shirazi, M., Erath, B. D., and Peterson, S. D. (2019). Bayesian inference of vocal fold material properties from glottal area waveforms using a 2D finite element model. *Appl. Sci.* 9:2735. doi: 10.3390/app9132735

Hagan, M., Demuth, H., Beale, M., and De Jesús, O. (2014). *Neural Network Design*. Stillwater, OK: Martin Hagan.

Hertegård, S., Gauffin, J., and Åke Lindestad, P. (1995). A comparison of subglottal and intraoral pressure measurements during phonation. *J. Voice* 9, 149–155.

Hillman, R. E., and Mehta, D. D. (2011). Ambulatory monitoring of daily voice use. *Perspect. Voice Disord.* 21, 56–61. doi: 10.1044/vvd21.2.56

Hillman, R. E., Stepp, C. E., Stan, J. H. V., Zañartu, M., and Mehta, D. D. (2020). An updated theoretical framework for vocal hyperfunction. *Am. J. Speech Lang. Pathol.* 29, 2254–2260. doi: 10.1044/2020_AJSLP-20-00104

Hunter, E. J., Titze, I. R., and Alipour, F. (2004). A three-dimensional model of vocal fold abduction/adduction. *J. Acoust. Soc. Am.* 115, 1747–1759. doi: 10.1121/1.1652033

Kempster, G. B., Gerratt, B. R., Abbott, K. V., Barkmeier-Kraemer, J., and Hillman, R. E. (2009). Consensus auditory-perceptual evaluation of voice: development of a standardized clinical protocol. *Am. J. Speech Lang. Pathol.* 18, 124–132. doi: 10.1044/1058-0360(2008/08-0017)

Kennedy, J., and Eberhart, R. C. (1995). "Particle swarm optimization," in *Proceedings of the IEEE International Conference on Neural Networks*, Perth, WA, 1942–1948.

Kingma, D. P., Ba, J. (2017). Adam: a method for stochastic optimization. *arXiv preprint arXiv:1412.6980*.

Lin, J. Z., Espinoza, V. M., Marks, K. L., Zañartu, M., and Mehta, D. D. (2020). Improved subglottal pressure estimation from neck-surface vibration in healthy speakers producing non-modal phonation. *IEEE J. Select. Top. Signal Process.* 14, 449–460. doi: 10.1109/jstsp.2019.2959267

Llico, A. F., Zañartu, M., Gonzalez, A. J., Wodicka, G. R., Mehta, D. D., Van Stan, J. H., et al. (2015). Real-time estimation of aerodynamic features for ambulatory voice biofeedback. *J. Acoust. Soc. Am.* 138, EL14–EL19. doi: 10.1121/1.4922364

Lucero, J. C., and Schoentgen, J. (2015). Smoothness of an equation for the glottal flow rate versus the glottal area. *J. Acoust. Soc. Am.* 137, 2970–2973. doi: 10.1121/1.4919297

Marks, K. L., Lin, J. Z., Burns, J. A., Hron, T. A., Hillman, R. E., and Mehta, D. D. (2020). Estimation of subglottal pressure from neck surface vibration in patients with voice disorders. *J. Speech Lang. Hear. Res.* 63, 2202–2218. doi: 10.1044/2020_JSLHR-19-00409

Marks, K. L., Lin, J. Z., Fox, A. B., Toles, L. E., and Mehta, D. D. (2019). Impact of nonmodal phonation on estimates of subglottal pressure from neck-surface acceleration in healthy speakers. *J. Speech Lang. Hear. Res.* 62, 3339–3358. doi: 10.1044/2019_JSLHR-S-19-0067

Mehta, D. D., Espinoza, V. M., Van Stan, J., Zañartu, M., and Hillman, R. (2019). The difference between first and second harmonic amplitudes correlates between glottal airflow and neck-surface accelerometer signals during phonation. *J. Acoust. Soc. Am.* 145, EL386–EL392. doi: 10.1121/1.5100909

Mehta, D. D., Van Stan, J. H., Zañartu, M., Ghassemi, M., Guttag, J. V., Espinoza, V. M., et al. (2015). Using ambulatory voice monitoring to investigate common voice disorders: research update. *Front. Bioeng. Biotechnol.* 3:155. doi: 10.3389/fbioe.2015.00155

Mehta, D. D., Zañartu, M., Feng, S. W., Cheyne, H. A. II, and Hillman, R. E. (2012). Mobile voice health monitoring using a wearable accelerometer sensor and a smartphone platform. *IEEE Trans. Biomed. Eng.* 59, 3090–3096. doi: 10.1109/TBME.2012.2207896

Perkell, J. S., Hillman, R. E., and Holmberg, E. B. (1994). Group differences in measures of voice production and revised values of maximum airflow declination rate. *J. Acoust. Soc. Am.* 96, 695–698. doi: 10.1121/1.410307

Perkell, J. S., Holmberg, E. B., and Hillman, R. E. (1991). A system for signal processing and data extraction from aerodynamic, acoustic, and electroglottographic signals in the study of voice production. *J. Acoust. Soc. Am.* 89, 1777–1781.

Popolo, P. S., Svec, J. G., and Titze, I. R. (2005). Adaptation of a pocket PC for use as a wearable voice dosimeter. *J. Speech Lang. Hear. Res.* 48, 780–791. doi: 10.1044/1092-4388(2005/054)

Rothenberg, M. (2013). "Rethinking the interpolation method for estimating subglottal pressure," in *Proceedings of the 10th International Conference on Advances in Quantitative Laryngology, Voice and Speech Research* (Cincinnati, OH: AQL Press), 111–112.

Story, B. H. (2008). Comparison of magnetic resonance imaging-based vocal tract area functions obtained from the same speaker in 1994 and 2002. *J. Acoust. Soc. Am.* 123, 327–335. doi: 10.1121/1.2805683

Story, B. H., and Titze, I. R. (1995). Voice simulation with a body-cover model of the vocal folds. *J. Acoust. Soc. Am.* 97, 1249–1260.

Story, B. H., Titze, I. R., and Hoffman, E. A. (1998). Vocal tract area functions for an adult female speaker based on volumetric imaging. *J. Acoust. Soc. Am.* 104, 471–487.

Švec, J. G., and Granqvist, S. (2018). Tutorial and guidelines on measurement of sound pressure level in voice and speech. *J. Speech Lang. Hear. Res.* 61, 441–461. doi: 10.1044/2017_JSLHR-S-17-0095

Švec, J. G., Titze, I. R., and Popolo, P. S. (2005). Estimation of sound pressure levels of voiced speech from skin vibration of the neck. *J. Acoust. Soc. Am.* 117, 1386–1394. doi: 10.1121/1.1850074

Titze, I. R. (2002). Regulating glottal airflow in phonation: application of the maximum power transfer theorem to a low dimensional phonation model. *J. Acoust. Soc. Am.* 111, 367–376. doi: 10.1121/1.1417526

Titze, I. R. (2006). *The Myoelastic Aerodynamic Theory of Phonation, 1st Edn*. Iowa, IA: National Center for Voice and Speech.

Titze, I. R., and Hunter, E. J. (2007). A two-dimensional biomechanical model of vocal fold posturing. *J. Acoust. Soc. Am.* 121, 2254–2260. doi: 10.1121/1.2697573

Titze, I. R., and Hunter, E. J. (2015). Comparison of vocal vibration-dose measures for potential-damage risk criteria. *J. Speech Lang. Hear. Res.* 58, 1425–1439. doi: 10.1044/2015_JSLHR-S-13-0128

Titze, I. R., and Story, B. H. (2002). Rules for controlling low-dimensional vocal fold models with muscle activation. *J. Acoust. Soc. Am.* 112(3 Pt 1), 1064–1074. doi: 10.1121/1.1496080

Titze, I. R., Svec, J. G., and Popolo, P. S. (2003). Vocal dose measures: quantifying accumulated vibration exposure in vocal fold tissues. *J Speech Lang. Hear. Res.* 46, 919–932. doi: 10.1044/1092-4388(2003/072)

Van Stan, J. H., Mehta, D. D., and Hillman, R. E. (2017a). Recent innovations in voice assessment expected to impact the clinical management of voice disorders. *Perspect. ASHA Spcl. Interest Groups* 2, 4–13. doi: 10.1044/persp2.SIG3.4

Van Stan, J. H., Mehta, D. D., Ortiz, A. J., Burns, J. A., Marks, K. L., Toles, L. E., et al.. (2020). Changes in a daily phonotrauma index after laryngeal surgery and voice therapy: implications for the role of daily voice use in the etiology and pathophysiology of phonotraumatic vocal hyperfunction. *J. Speech Lang. Hear. Res.* 63, 3934–3944. doi: 10.1044/2020_JSLHR-20-00168

Van Stan, J. H., Mehta, D. D., Sternad, D., Petit, R., and Hillman, R. E. (2017b). Ambulatory voice biofeedback: relative frequency and summary feedback effects on performance and retention of reduced vocal intensity in the daily lives of participants with normal voices. *J. Speech Lang. Hear. Res.* 60, 853–864. doi: 10.1044/2016_JSLHR-S-16-0164

Van Stan, J. H., Ortiz, A. J., Cortes, J. P., Marks, K. L., Toles, L. E., Mehta, D. D., et al. (2021). Differences in daily voice use measures between female patients with nonphonotraumatic vocal hyperfunction and matched controls. *J. Speech Lang. Hear. Res.* 64, 1457–1470. doi: 10.1044/2021_JSLHR-20-00538

Zañartu, M. (2006). *Influence of acoustic loading on the flow-induced oscillations of single mass models of the human larynx* (Master's thesis). School of Electrical and Computer Engineering, Purdue University, West Lafayette, IN, United States.

Zañartu, M. (2010). *Acoustic coupling in phonation and its effect on inverse filtering of oral airflow and neck surface acceleration* (Ph.D. thesis). School of Electrical and Computer Engineering, Purdue University, West Lafayette, IN, United States.

Zañartu, M., Galindo, G. E., Erath, B. D., Peterson, S. D., Wodicka, G. R., and Hillman, R. E. (2014). Modeling the effects of a posterior glottal opening on vocal fold dynamics with implications for vocal hyperfunction. *J. Acoust. Soc. Am.* 136, 3262–3271. doi: 10.1121/1.4901714

Zañartu, M., Ho, J. C., Mehta, D. D., Hillman, R. E., and Wodicka, G. R. (2013). Subglottal impedance-based inverse filtering of voiced sounds using neck surface acceleration. *IEEE Trans. Audio Speech Lang. Process.* 21, 1929–1939. doi: 10.1109/TASL.2013.2263138

Zañartu, M., Mongeau, L., and Wodicka, G. R. (2007). Influence of acoustic loading on an effective single mass model of the vocal folds. *J. Acoust. Soc. Am.* 121, 1119–1129. doi: 10.1121/1.2409491

Zhang, Z. (2020). Estimation of vocal fold physiology from voice acoustics using machine learning. *J. Acoust. Soc. Am.* 147, EL264–EL270. doi: 10.1121/10.0000927

Author Disclaimer: The content is solely the responsibility of the authors and does not necessarily represent the official views of the National Institutes of Health.

Conflict of Interest: MZ has a financial interest in Lanek SPA, a company focused on developing and commercializing biomedical devices and technologies. MZ's interests were reviewed and are managed by Universidad Técnica Federico Santa María in accordance with its conflict-of-interest-policies. RH and DM have a financial interest in InnoVoyce LLC, a company focused on developing and commercializing technologies for the prevention, diagnosis, and treatment of voice-related disorders. RH's and DM's interests were reviewed and are managed by Massachusetts General Hospital and Mass General Brigham in accordance with their conflict-of-interest policies.

The remaining authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

Publisher's Note: All claims expressed in this article are solely those of the authors and do not necessarily represent those of their affiliated organizations, or those of the publisher, the editors and the reviewers. Any product that may be evaluated in this article, or claim that may be made by its manufacturer, is not guaranteed or endorsed by the publisher.

frontiers in Physiology

ORIGINAL RESEARCH
published: 06 September 2021
doi: 10.3389/fphys.2021.692667

Artificial Intelligence May Predict Early Sepsis After Liver Transplantation

Rishikesan Kamaleswaran[1,2], *Sanjaya K. Sataphaty*[3], *Valeria R. Mas*[4], *James D. Eason*[5] *and Daniel G. Maluf*[4]*

[1] Emory University School of Medicine, Atlanta, GA, United States, [2] Georgia Institute of Technology, Atlanta, GA, United States, [3] Sandra Atlas Bass Center for Liver Diseases & Transplantation, Northshore University Hospital, Northwell Health, Manhasset, NY, United States, [4] University of Maryland School of Medicine, Baltimore, MD, United States, [5] Transplant Institute, University of Tennessee, Memphis, TN, United States

OPEN ACCESS

Edited by:
Michael Döllinger,
University Hospital Erlangen, Germany

Reviewed by:
Jinzhi Lei,
Tianjin Polytechnic University, China
Kranthi Kolli,
Abbott, United States
Yan Wang,
Amazon, United States

***Correspondence:**
Daniel G. Maluf
dmaluf@som.umaryland.edu

Specialty section:
This article was submitted to Computational Physiology and Medicine, a section of the journal Frontiers in Physiology

Received: 08 April 2021
Accepted: 29 July 2021
Published: 06 September 2021

Citation:
Kamaleswaran R, Sataphaty SK, Mas VR, Eason JD and Maluf DG (2021) Artificial Intelligence May Predict Early Sepsis After Liver Transplantation. Front. Physiol. 12:692667. doi: 10.3389/fphys.2021.692667

Background: Sepsis, post-liver transplantation, is a frequent challenge that impacts patient outcomes. We aimed to develop an artificial intelligence method to predict the onset of post-operative sepsis earlier.

Methods: This pilot study aimed to identify "physiomarkers" in continuous minute-by-minute physiologic data streams, such as heart rate, respiratory rate, oxygen saturation (SpO2), and blood pressure, to predict the onset of sepsis. The model was derived from a cohort of 5,748 transplant and non-transplant patients across intensive care units (ICUs) over 36 months, with 92 post-liver transplant patients who developed sepsis.

Results: Using an alert timestamp generated with the Third International Consensus Definition of Sepsis (Sepsis-3) definition as a reference point, we studied up to 24 h of continuous physiologic data prior to the event, totaling to 8.35 million data points. One hundred fifty-five features were generated using signal processing and statistical methods. Feature selection identified 52 highly ranked features, many of which included blood pressures. An eXtreme Gradient Boost (XGB) classifier was then trained on the ranked features by 5-fold cross validation on all patients (n = 5,748). We identified that the average sensitivity, specificity, positive predictive value (PPV), and area under the receiver-operator curve (AUC) of the model after 100 iterations was 0.94 ± 0.02, 0.9 ± 0.02, 0.89 ± 0.01, respectively, and 0.97 ± 0.01 for predicting sepsis 12 h before meeting criteria.

Conclusion: The data suggest that machine learning/deep learning can be applied to continuous streaming data in the transplant ICU to monitor patients and possibly predict sepsis.

Keywords: machine learning, liver transplant, surgery, physiological data streams, artificial intelligence, sepsis

INTRODUCTION

Liver transplantation continues to be the optimal and more successful therapy for end-stage liver disease and cirrhosis (Kim et al., 2019). One of the biggest challenges in the transplant community is the discrepancy of donor availability and the need of the recipients. Transplant centers frequently appeal to the use of marginal or suboptimal donors to decrease this gap, while, at the same time, increasing the chances for post-transplant complications associated with organ dysfunction (Kim et al., 2019).

Moreover, in recent years, the increased availability of more potent immunosuppressive agents, along with sicker and older recipients needing transplantation, has increased the incidence of opportunistic infections (OIs) affecting patient survival after liver transplantation (LT) (Haidar et al., 2019; He et al., 2019). Post-transplant infections with or without surgical complications are the leading cause of morbidity and mortality post-LT (Kim et al., 2019). Overall, infections and sepsis are estimated to occur in more than half of LT recipients, and are the main cause of post-LT death between days 21 and 180 (Sun et al., 2011; Fischer et al., 2013; Martin et al., 2014; Haidar et al., 2019; He et al., 2019). Bacterial infections are the most common post-transplant infections (>70%), followed by viral and fungal infections (Sun et al., 2011; Haidar et al., 2019; He et al., 2019). Fortunately, due to intensive screening practices to detect latent infections in liver transplant candidates, and with the implementation of appropriate prophylactic protocols and therapy, mortality associated with post-LT infections is still low (<10%) (Sun et al., 2011; Martin et al., 2014; He et al., 2019). Known risk factors associated with infection after LT include a high model for end-stage liver disease (MELD) score, re-transplantation, advanced age of the recipient, number of blood transfusions, renal replacement therapy (RRT), and a long intensive care unit (ICU) stay, among others (Haidar et al., 2019; He et al., 2019). Several steps in the physical examination and laboratory assessment allow a clinician to identify active infections that would prompt therapy to prevent complications. It is known, however, that delays in diagnosis and therapy implementations would carry higher mortality in this population (Kumar et al., 2006; Dombrovskiy et al., 2007). Because of the scarce resource of liver grafts and the associated mortality of post-transplant infections, biomarkers or markers capable of accurately expediting diagnosis would be of significant clinical significance in a transplant unit.

Sepsis is a common event, with more than a million Americans getting hospitalized each year (Dombrovskiy et al., 2007; Liu et al., 2014). Sepsis is caused by a heightened inflammatory response to an infection, and can quickly progress to multi-organ failure and death (Liu et al., 2014). In the septic shock phase of the disease, every hour that treatment is delayed can lead to a 7.6% increase in mortality (Kumar et al., 2006). In liver transplantation, this phenomenon is not different in the general population, and an early infection due to surgical complications, such as bleeding, bile leak, or rejection, may trigger infections and sepsis with severe consequences in recipients (Kumar et al., 2006; Elkholy et al., 2019).

A number of recent studies have applied artificial intelligence (AI) and machine learning to identify patients at risk for sepsis earlier, thereby potentially reducing mortality and morbidity (Kumar et al., 2006; Nemati et al., 2018; Elkholy et al., 2019). These methods have typically used an array of clinical and laboratory variables in the electronic medical record (EMR) to predict the risk of sepsis. While such methods have achieved a significant performance in retrospective studies, they are limited by the aperiodic and unstructured nature of EMR data. Alternative methods for developing predictive models for sepsis have used high-frequency data streams captured from the medical monitor, such as heart rate, blood pressures, respiratory rate, and oxygen saturation (Kamaleswaran et al., 2018; van Wyk et al., 2019). The use of such biosensor data may identify physiomarkers that present hours before the clinical manifestation of the disease or event, thereby allowing for earlier recognition and the initiation of therapy. In this study, we evaluated the effectiveness of high-frequency physiological data stream analysis in predicting the onset of sepsis in liver transplant patients. We developed and tested a number of machine learning methods using features derived from the physiological time series to generate predictions at various time intervals before the Third International Consensus Definition of Sepsis (Sepsis-3) clinical definition (Singer et al., 2016).

MATERIALS AND METHODS

Data Collection Environment

This observational retrospective study was approved by the Institutional Review Board (IRB) of the University of Tennessee Health Science Center. We collected continuous physiological data streams from bed-side monitors using the Cerner iBus (Cerner Corporation, Kansas City, MO, United States) (Cerner Corporation, 2014). The Cerner iBus generated minute-by-minute heart rate (HR), respiratory rate (RR), blood pressure (mean, systolic, and diastolic), and oxygen saturation (SpO_2) data streams; however, continuous temperature was not available and was, therefore, excluded from the analysis. We captured non-invasive blood pressure (NIBP), which was sampled at least once an hour, and, in some cases where clinical deterioration was suspected, the NIBP was sampled more frequently.

Case Definition

Patients admitted to the intensive care unit across the Methodist University Hospital and Transplant Institute (UTHSC) between January 2017 and January 2020, with continuous minute-by-minute physiological monitoring data, were included in the study. In this study, we utilized the Sepsis-3 definition [SHAP (SHapley Additive exPlanations), 2021]; patients who met Sepsis-3 criteria but did not have high-frequency data recorded within the prior 24 h were excluded. Sepsis-3 definitions were applied serially using the method described by Nemati et al. (2018) in order to identify the time of sepsis onset (event time) (Nemati et al., 2018). We identified controls as those who had never met sepsis criteria during their encounter. To identify a control event time (for supervised learning), we used a randomly generated timestamp that occurred between admission and discharge,

provided that the 24-h data availability criterion prior to the random event time was met. All the data were then temporally aligned to the event time, identified as t_{sepsis}

Feature Extraction and Feature Selection

For each of the six physiological data streams [heart rate, respiratory rate, oxygen saturation, systolic blood pressure (SBP), diastolic blood pressure (DBP), and mean femoral artery blood pressure (MAP)], features were extracted using eight statistical and two time-frequency domain methods, namely, mean, sum, minimum, maximum, frequency of the measurement (length), standard deviation, variance, kurtosis, fast Fourier transform (FFT), and continuous wavelet transform (CWT) (Christ et al., 2018). A number of parameters were included for evaluating the FFT coefficients (0–100, with a step of 4); then, we extracted the absolute coefficient values for each parameter. For CWT features, we evaluated width values of 0–20 at a step of 2. These features were extracted for each hour within the 3-h window across six data streams, for a total of 774 features per window; 24 FFT features consistently returned null and subsequently removed, resulting in a total of 750. Missing data were imputed if there was a previous record; otherwise, we used the population median value. These features were then concatenated into a single feature vector that incorporated temporal dynamics over the 3-h period.

We then applied a variety of feature selection methods, including statistical, and thus performed both non-parametric Mann–Whitney-U and parametric independent sample *t*-tests, ridge, lasso, recursive feature elimination (RFE), and random forest-based variable importance utilizing information gain and gini impurity. These feature selection methods were performed in order to reduce data dimensionality to a limited set of markers that predict the onset of sepsis.

The dataset was then segmented into two cohorts; the first included all patients who were admitted to the intensive care unit without having received a liver transplant at least 31 days prior to admission, and the second cohort included all patients who underwent transplantation. For the training of the model, we implemented a subsampling strategy where we randomly selected an equal number of controls to cases. In order to control for over-fitting, we implemented a 5-fold cross validation on each iteration to derive training and test performances. We then iterated this training 100 times to generate unique model performances from each run and reported the averaged performance measure overall runs. Hyperparameters were evaluated using a grid-search approach, with which we predefined the upper and lower limits of the hyperparameters and generated a series of models and recorded their performance. The hyperparameters that achieved the most stable model performance, with minimal variance over the 100 runs, were selected and used to train the entire first cohort data. We selected the optimal hyperparameter for each of the algorithms that were explored, namely, eXtreme Gradient Boosting (XGB), logistic regression (LR), support vector machine (SVM), and random forest (RF). The remaining selected models were then validated on the transplant cohort.

Machine Learning Pipeline

Prior to the modeling of high-dimensional data streams, we applied an unsupervised cluster visualization technique called t-distributed stochastic neighbor embedding (tSNE) (Van der Maaten and Hinton, 2008). This method converts similarities between data points to joint probabilities and tries to minimize the divergence between these joint probabilities in a low-dimensional manner to illustrate possible clusters and separation. Then, in the binary classification, we applied a number of machine learning classifiers to generate complementary but competing models. We investigated supervised learning methods such as eXtreme Gradient XGB, LR, SVM, and RF, with both XGB and RF being non-linear ensemble-based learning methods. In particular, XGB is unique in incorporating sequential boosting to improve classification performance, but it may also be sensitive to overfitting. Furthermore, SVM is a classical machine learning method that utilizes hyperplanes to optimize separation among features and has been successfully used for binary classification tasks. In addition, ALR is a statistical learning method and often serves as a benchmark for machine learning model comparison. We utilized the above algorithms to compare performance across unique learning strategies to select an optimal algorithm that performs best for this dataset.

In order to generate explainable feature importance, we used the SHapley Additive exPlanations (SHAP) package (Lundberg and Lee, 2017). The SHAP algorithm uses methods from game theory to explain the output of machine learning models; it has been noted to be state-of-the-art in terms of generating reliable explanations of predictive model outputs.

Model benchmarks were generated by computing area under the receiver-operator curve (AUC), area under the precision-recall curve (AUPRC), sensitivity, specificity, and positive predictive value (PPV). In particular, AUC is a traditional benchmarking tool for determining performance over a range of possible model-estimated probability thresholds; however, it assumes a balanced distribution of samples. Conversely, AUPRC is more useful for measuring performance across imbalanced and low-PPV scenarios; a higher AUPRC indicates that the model can accurately identify all positive examples without compromising specificity.

We utilized Python 3.6 and the XGBoost package (XGBoost Documentation, 2021) for developing the XGB model and the sci-kit learn (Scikit-Learn: Machine Learning in Python, 2021) package for developing the remaining machine learning and statistical analysis code base. We utilized the SHAP library to derive explainable interpretations and summary plots [SHAP (SHapley Additive exPlanations), 2021].

RESULTS

Data Missingness

In the derivation dataset, the rate of missing value was highest between MAP and DBP, with an average of 16% patients having at least one missing value in the 3-h observational window. Oxygen saturation was the most often recorded, with only 0.1% of the patients missing this measure, followed by HR with a missing value of up to 0.6%, RR in 2.7% of patients, and

TABLE 1 | Characteristics of the study population.

Characteristics	Sepsis (Non-Transplant)			Sepsis (Transplant)		
	Overall	Yes	No	Overall	Yes	No
Patient, n (%)	5,748 (100)	604 (10.5)	5,144 (89.5)	252 (100)	92 (27)	160 (73)
Male, n (%)	2,932 (49.2)	299 (48.5)	2,633 (49.3)	160 (63)	50 (54)	110 (69)
Mechanical ventilation, n (%)	1,356 (22.8)	490 (74.6)	896 (16.8)	252 (100)	92 (100)	160 (100)
In hospital deaths, n (%)	439 (7.4)	176 (28.5)**	263 (4.9)	23 (9)	8 (9)	15 (9)
Age (yr.) median (IQR)	62 (50–72)	63 (52–72)	61.5 (50–72)	57 (48–66)	61 (46–67)	57 (50–65)
ICU LOS (d), median (IQR)	5 (3–8)	11 (6–20)**	4 (3–7)	2 (2–5)	4 (2–5)	2 (2–5)
ICU LOS > = 7d, n (%)	1,917 (32.2)	435 (70.5)**	1,482 (27.7)	58 (22.9)	21 (23)	37 (22.9)
Self-reported race, n (% row-wise)						
Black or African American	3,453 (58.0)	387 (62.7)*	3,066 (57.4)	44 (17)	7 (16)	37 (84)
White	2,360 (39.6)	214 (34.7)	2,146 (40.2)	178 (71)	14 (8)	164 (80)
Other/Unknown	104 (1.8)	13 (2.1)	91 (1.7)	20 (8)	2	18 (6)
Multiple	21 (0.4)	1 (0.2)	20 (0.4)	0	0	0
Asian	19 (0.3)	2 (0.3)	17 (0.3)	1	0	1
Self-reported Ethnicity, n (%)						
Not Hispanic or Latino	5,847 (98.2)	605 (98.0)	5,242 (98.2)	219 (87)	11 (1)	208 (99)
Hispanic or Latino	69 (1.2)	8 (1.4)	69 (1.3)	19 (8)	2 (10)	17 (90)
Unknown or Declined	33 (0.6)	4 (0.6)	29 (0.5)	0	0	0

*Significant at a = 0.01; **significant at a = 0.001.

SBP in 4.8%. **Supplementary Figure S1** illustrates the correlation between the missing variables, and suggests that when MAP is missing, DBP is also missing and vice versa. In 60% of the cases, SBP is associated with missing MAP and DBP. In cases where HR is missing, in 30% of the patients, MAP and DBP are also missing.

We identified a total of 5,748 non-transplant patients who were admitted to the intensive care unit over an 8-month period, 604 (10.5%) of whom met the "Sepsis-3" criteria defined as suspicion of infection in the presence of organ failure (Singer et al., 2016). Furthermore, another 252 patients were separately identified to have undergone a liver transplant, 92 (36%) of whom met Sepsis-3 criteria during their stay in the ICU.

Age and gender differences were not statistically significant in the general cohort (**Table 1**). Model for end-stage liver disease scores was also not statistically different between the cohorts, with scores consistently ranging from 22 to 28 across both cohorts. As expected, in the transplant program, a greater portion of the transplant cohort consisted of male Caucasians. In-hospital mortality in the transplantation cohort was 9%, which is less than the in-hospital mortality in the general cohort. The incidence of sepsis in the transplant cohort was significantly higher than in the general cohort. The median age of the transplant cohort was 57 years, with the sepsis patients being, on average, 4 years older than the non-sepsis liver transplant patients across each group similar to the general cohort. All patients in the transplant cohort were temporarily mechanically ventilated, while only 23% was in the general cohort.

An Unsupervised clustering, using the tSNE method, of the raw data up to 12 h prior to sepsis onset suggests that clusters can be distinguishable (**Figure 1A**, tSNE plot). The cluster to the left largely consists of patients without sepsis, while the cluster to the top and to the right contains a significant portion of patients with sepsis, indicating that further analysis of the data may reveal useful predictive markers for sepsis. We found a number of overlapping distinguishing physiomarkers when we utilized the gradient boosting method (**Figures 1B,C**), and the SHAP output (**Figure 1D**). Notably, HR, RR, and SBPs were significant explainers for patients who developed sepsis early in the clinical course.

Figure 2 illustrates an example patient with sepsis where the physiological data streams were available up to 16 h before onset. In this figure, dynamic shifts are seen in the HR, RR, and blood pressure data streams during the time leading to sepsis. Moreover, interventional response *via* fluid resuscitation is also observed shortly thereafter.

Statistical Analysis

A total of 750 features were generated from all the physiological data streams using statistical and time-frequency domain methods (described in the methods section); these represent features generated in the observational window at 12 h (prediction horizon) prior to sepsis onset. By Student's *t*-test against these continuous measures to identify distinguishing features, we found that the statistical significance for the transplant cohort at $p < 0.05$ was observed in 311 features, of which 106 were various time-frequency abstractions of DBP and 79 features were related to SBP, 73 to RR, 38 to MAP, and 15 belonged to HR. None of the SpO_2 features figured as statistically significant. Among the signal processing features, at $p < 0.001$, FFT of DBP and SBP, and CWT of RR were significant (**Figure 3**). The box plots illustrated

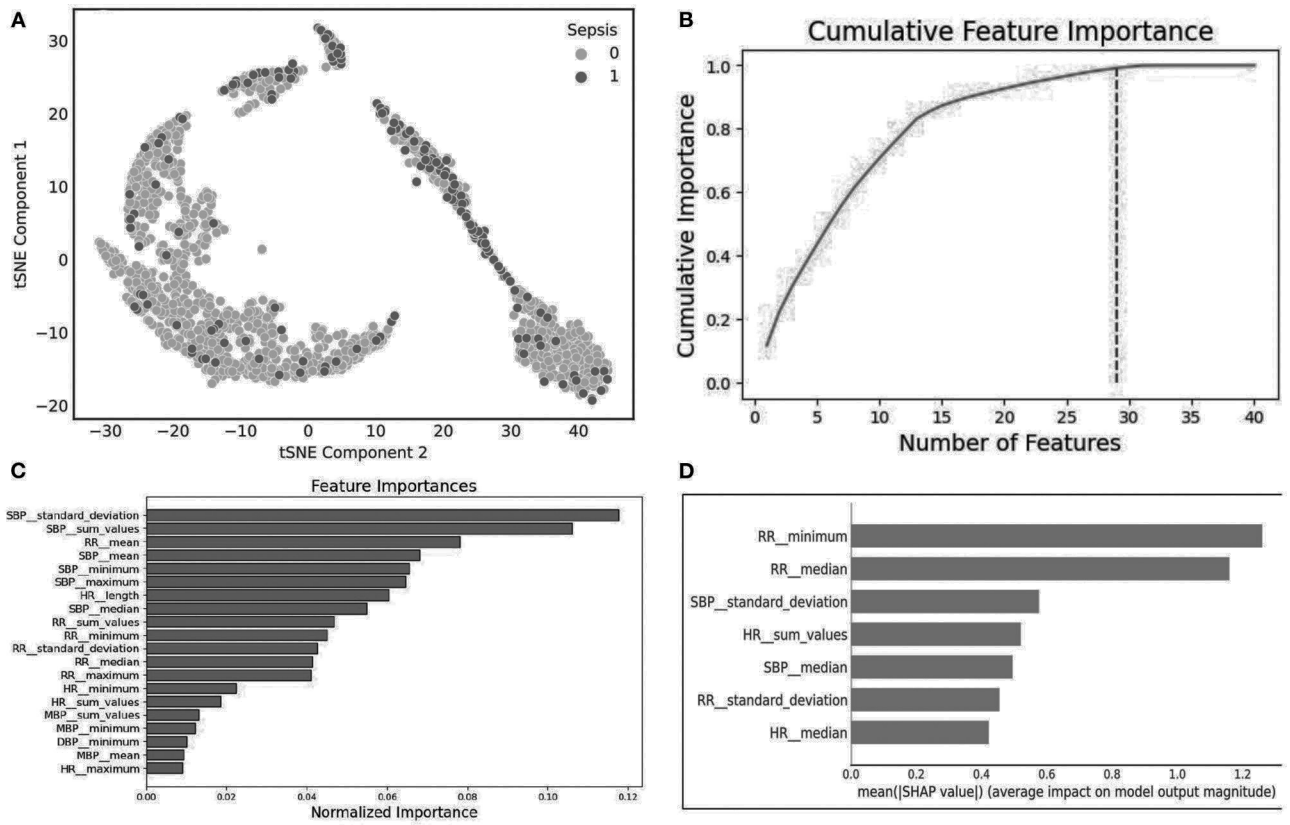


FIGURE 1 | (A) t-Distributed stochastic neighbor embedding (tSNE) plot characterizing sepsis (dark purple) and non-sepsis clusters (pink); **(B)** cumulative feature importance identifies 29 features that explain 99% of the data; **(C)** feature importance generated using light gradient boosted machines; **(D)** explainable importance with SHapley Additive exPlanations (SHAP) after feature selection.

in **Figure 3** show that frequency-domain characteristics were meaningfully distinguishable among blood pressures, while more complex dynamics that spanned the time-frequency domain were apparent among respiratory rates.

Machine Learning

Utilizing the statistically significant features ($n = 311$), we applied feature selection techniques, namely, the RFE method, which generated 22 features that were highly predictive. All of these 22 features were derived from statistical and continuous wavelet transform methods, and indicated that SBP characteristics are the top predictor of sepsis (**Figure 1D**). Separately, using ridge and lasso feature selection, we applied a defined coefficient threshold of 0.5 to select the most predictive features. The lasso method selected 12 features, which consisted exclusively of statistics from respiratory rate. The ridge method selected 52 highly ranked features, of which the top feature was SBP, with various temporal permutations of SBP appearing a total of nine times. The second most important feature was DBP, which appeared a total of 10 times, followed by RR, which appeared 12 times. The models were developed using both the RFE and ridge methods, and the ridge-based feature selection was identified as the optimal feature set because of its improved performance across the 5-fold cross-validation benchmarks. While we evaluated XGB, LR, SVM, RF, and MLP, XGB was identified as the optimal model after averaging 10 randomized runs of the 5-fold cross-validation. Because of the significant overfitting that occurred in the MLP pipeline early in the analysis, we did not pursue it for further hyperparameterization. **Figure 4** illustrates model performance, such as AUC and AUPRC, for the machine learning methods evaluated. In the figure, both XGB and RF are consistently shown to have the highest performance across both benchmarks, with XGB slightly outperforming RF.

Table 2 lists the performances of the logistic regression, support vector machine, random forest, and eXtreme Gradient Boost models. The XGB model was identified as the optimal model because of generally improved performance across all metrics, with a mean sensitivity of 0.94, specificity of 0.90, and an AUC of 0.97, as shown in **Figure 4**. The SVM model performed worst with respect to AUC (0.63) but had the highest specificity (0.94). The RF model performed relatively close to the XGB model but with a lower sensitivity (0.92) and specificity (0.88). The LR model had the lowest overall PPV (0.76). The XGB model outperformed all the other models in terms of each metric except for specificity. The optimal hyperparameters used in the XGB model were as follows: max depth of 6, subsample parameter

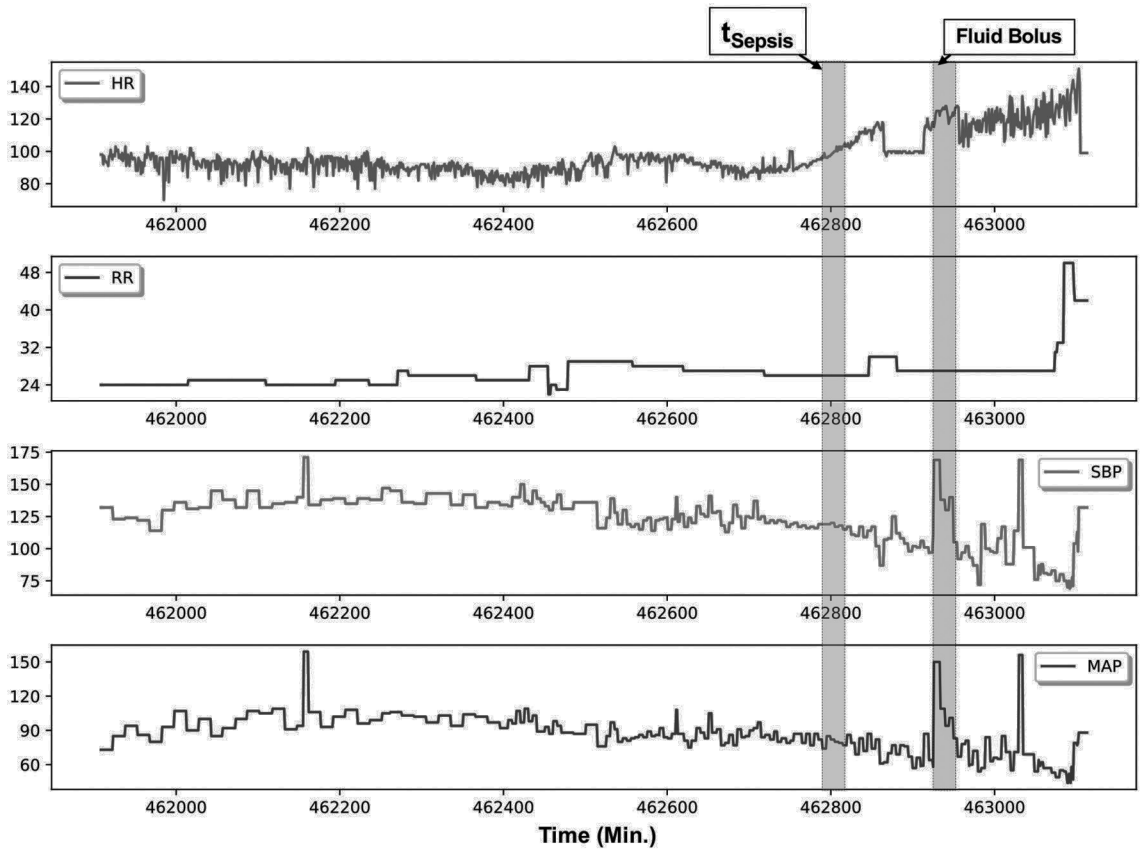


FIGURE 2 | An example patient with sepsis is illustrated in this figure; continuous physiological data were captured over an 18-h post-transplantation period. The patient met Sepsis-3 criteria (t_{Sepsis}) 13 h post transplantation (retrospectively identified), and fluid resuscitation (fluid bolus) was initiated 1.5 h thereafter. Several elements are of note within this patient, namely, in the preceding hours before meeting criteria, heart rate (HR) variability is noticeably reduced, accompanied by increased dynamics in the systolic blood pressure (SBP) and mean femoral artery blood pressure (MAP) data streams.

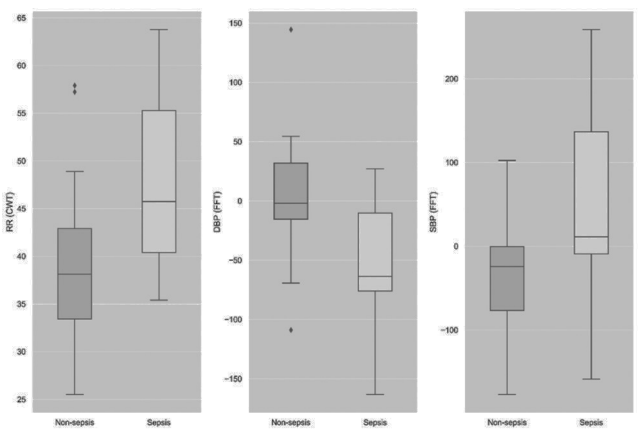


FIGURE 3 | Statistical significance ($p < 0.001$) was observed among the three physiological signals, namely, respiratory rate, diastolic blood pressure, and systolic blood pressure at 12 h prior to sepsis onset.

of 1, the minimum sum of instance weight for child of 1, and a learning rate of 0.1. The optimal SVM kernel function was linear. The threshold used for binary classification was 0.5.

DISCUSSION

Liver transplantation is a life-saving therapy for patients with liver cancer and end-stage liver disease. In the United States in 2017, more than 7,000 LTs were performed (Kim et al., 2019).

Transplant recipients are, however, at high risk for complications such as infections due to advanced age, obesity, comorbidities, and issues associated with the transplant event that may be related to surgical complications or organ dysfunction (Pedersen and Seetharam, 2014). Furthermore, systemic immunosuppression has rendered liver recipients susceptible to *de novo* infections and the reactivation of preexisting latent infections such as viral infections. Infections occurring during the first month post-LT are usually nosocomial or donor-derived or the result of a perioperative complication, such as a surgical complication, or organ dysfunction (Hernandez Mdel et al., 2015). A recent review of the Organ Procurement and Transplantation Network (OPTN) data from 64,977 patients who underwent liver transplantation identified the incidence of 90-day and 1-year mortalities at 5 and 10%, respectively. Although death associated with cardiovascular/cerebrovascular/pulmonary/hemorrhage was the most common cause of death within the first 21 days (7-day:

53%), only 20% of patients who underwent liver transplantation died from these causes after 180 days. Importantly, infections were the most frequent cause of death 30–180 days after liver transplantation. In contrast, after roughly 200 days from the time of liver transplantation, other causes were the most frequent cause of death (Baganate et al., 2018).

Severe sepsis, or infection with systemic inflammation, poses a substantial burden on the United States healthcare system, leading to >7,50,000 hospitalizations and 2,00,000 deaths annually (Moore et al., 2016). Severe sepsis remains a leading cause of death in the United States, with in-hospital mortality ranging from 12 to 26% (Donnelly et al., 2016). In solid organ transplantation and in contrast to the general belief, infection, and sepsis are more frequent in the general population, but the mortality associated with sepsis is lower, as also demonstrated in the analysis and results (Donnelly et al., 2016).

A big challenge is, however, early diagnosis, as the syndrome of sepsis has a wider range of causative organisms and differing presentations among immunosuppressed individuals such as patients who underwent liver transplant (Oriol et al., 2015). Furthermore, in the septic shock phase of the disease, every hour that treatment is delayed can lead to a 7.6% increase in mortality (Kumar et al., 2006).

Traditional markers of systemic inflammatory response syndrome and clinical presentation may not be present among the immunosuppressed, despite active overwhelming infection (Gauer, 2013).

Hereby, the analysis of patients who underwent liver transplantation patients who were admitted to the intensive care unit post-surgery revealed novel physiomarkers that can predict the onset of sepsis earlier and may have an impact on clinical decisions. An illustration using the tSNE visualization method indicated that there are unique clusters that emerge with separation between sepsis and non-sepsis cohorts. This indicates that the source data, comprising physiological data streams, may indeed be useful to predict the onset of sepsis within this cohort. We further found that these physiomarkers existed at least 12 h before a clinical definition was made. Among the important features, we noted that, when compared across two different explainability methods, we saw a consistent trend in the statistical markers of RR and SBP, along with HR, dominating the list of signals that predicted sepsis early in the clinical course. These vital sign measures have been previously described using EMR data. However, they have not been discussed in the context of continuous bedside monitoring for patients who received liver transplants in the past (Desautels et al., 2016; Bloch et al., 2019). While signal processing methods, FFT and CWT, were both statistically significant between the cohorts, they were outranked by the statistical features derived from the same physiological data streams.

We also found that, while several models may be useful as optimal candidates, the eXtreme Gradient Boost model specifically showed higher performance. In the selection criteria for the optimal model, we ensured that a specificity value of at least 0.6 would be required, as to not overwhelm nursing staff with false alarms. Therefore, these results indicate a value in the

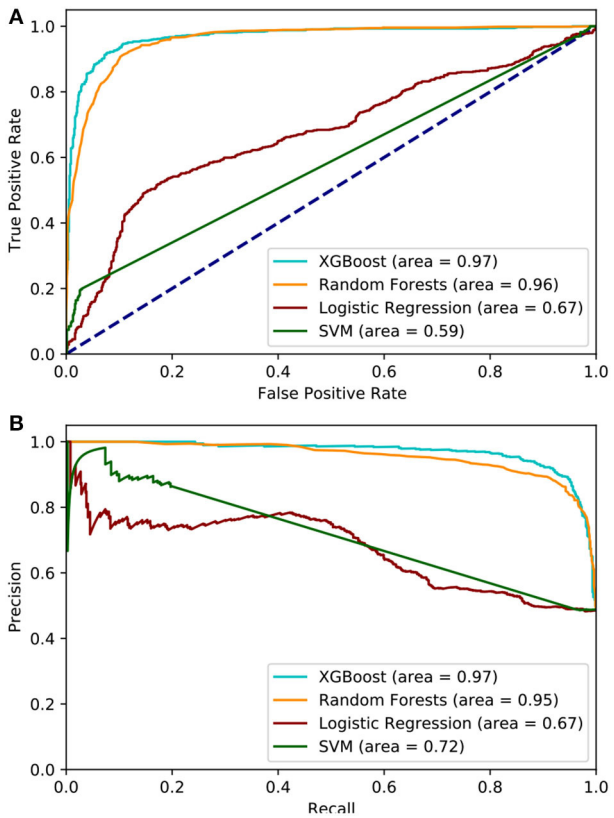


FIGURE 4 | (A) Area under the Receiver Operator Curve, the XGBoost model achieves the highest AUC while SVM performed the worst. **(B)** area under the Precision Recall curve, XGBoost maintains the best AUPRC but Logistic Regression is ranked the worst.

TABLE 2 | Comparison of model performance.

	LR	SVM	RF	XGB
	mean [95% CI]			
Sensitivity	0.49 [0.44–0.49]	0.28 [0.27–0.33]	0.92 [0.90–0.92]	**0.94 [0.93–0.95]**
Specificity	0.85 [0.83–0.86]	0.94 [0.93–0.95]	0.88 [0.83–0.89]	**0.90 [0.89–0.90]**
PPV	0.76 [0.74–0.78]	0.83 [0.82–0.84]	0.88 [0.87–0.89]	**0.89 [0.88–0.91]**
AUC	0.67 [0.66–0.70]	0.63 [0.62–0.65]	0.96[0.93–0.96]	**0.97 [0.95–0.97]**

use of bedside monitoring data streams for informed clinical decision-making and potential treatment plans for patients who received liver transplants in the past, and may serve as useful alternatives to existing clinical monitoring.

Specific features derived from time-frequency domain extractions revealed the useful characteristics of the continuous physiological data streams that can highly predict sepsis. Specifically, in the results, we found that SBP and DBP, along with changes in RR, were among the dominant features (top 10) in the model. We noted that SpO_2 was not observed to be a significant predictor. While significant literature has been proposed around the utility of HR and HR variability (Ahmad et al., 2009), we noted that, in the model, these appear only seven times in the 52 features that were included in model development.

We sought to develop a minimal physiologic model of sepsis because of the unpredictable nature of orders and their results. The development, therefore, of a minimalistic predictive model may allow for wider use. However, we note that there may be significant improvements in the performance of the model by incorporating clinical- and laboratory-based findings. We expect this to improve the model performance in prospective deployment.

Limitations associated with this study are being derived from a single site and incomplete data analysis due to incomplete clinical data collection (e.g., IMS, surgical complications, HAT, among others). As for the future study, we seek to incorporate data across multiple sites. We were unable to compute standard severity of illness scores because of the limited clinical data, such as the composite sepsis risk score, D-MELD (donor age recipient MELD), donor risk index, Euro-transplant donor risk index, or survival outcome following liver transplantation (SOFT) score, to perform benchmark comparisons. We also reported a small sample size of patients who received liver transplants in the past, which could limit the generalizability of the model; thus, larger datasets from multi-site transplantation units could improve the external generalization.

Clinical Translation and Future Study

In this pilot study, we demonstrated that continuous physiological data streams can be used for informed clinical decision-making related to the risk of sepsis among patients who received liver transplants in the past. While the model proposed in this study can be directly applied, clinical translation has been a major challenge for machine learning algorithms. We have previously demonstrated that, while clinical data may be useful by themselves, machine learning algorithms are also influenced by measurement indicators (e.g., practice patterns), such as specific applications of sepsis bundles that may indicate increased clinical suspicion (Futoma et al., 2021). In order to control for these confounding variables, a clinical translation of such machine learning models needs to be carefully managed, for instance, by enacting benchmark methods that include silent prospective pilots and clinical adjudication of alerts. These efforts form the basis for the future study.

CONCLUSION

Artificial intelligence is becoming an important tool to assist many areas in the field, such as inpatient and outpatient monitoring, including the setting of solid organ transplantation (Woldaregay et al., 2019). In this context, this is one of the first studies that aim to demonstrate that the use of machine learning and AI tools may accurately assess a large amount of continuous data streams from the bedside of patients and help to make earlier diagnoses or event recognition, allowing for faster and more accurate clinical decisions.

DATA AVAILABILITY STATEMENT

The data analyzed in this study is subject to the following licenses/restrictions: Restricted due to institutional IRB policy. Requests to access these datasets should be directed to rkamaleswaran@emory.edu.

ETHICS STATEMENT

The studies involving human participants were reviewed and approved by University of Tennessee Health Science Center. Written informed consent for participation was not required for this study in accordance with the national legislation and the institutional requirements.

AUTHOR CONTRIBUTIONS

RK and DM participated in research design. DM, RK, and VM participated in the writing of the article. RK, DM, VM, and SS participated in data analysis. RK, DM, and JE participated in performing the research. All authors contributed to the article and approved the submitted version.

FUNDING

This work was supported by YOLT (You Only Live Twice) Foundation.

SUPPLEMENTARY MATERIAL

The Supplementary Material for this article can be found online at: https://www.frontiersin.org/articles/10.3389/fphys.2021.692667/full#supplementary-material

REFERENCES

Ahmad, S., Ramsay, T., Huebsch, L., Flanagan, S., McDiarmid, S., Batkin, I., et al. (2009). Continuous multi-parameter heart rate variability analysis heralds onset of sepsis in adults. *PLoS ONE* 4:e6642. doi: 10.1371/journal.pone.0006642

Baganate, F., Beal, E. W., Tumin, D., Azoulay, D., Mumtaz, K., Black, S. M., et al. (2018). Early mortality after liver transplantation: Defining the course and the cause. *Surgery* 164, 694–704. doi: 10.1016/j.surg.2018.04.039

Bloch, E., Rotem, T., Cohen, J., Singer, P., and Aperstein, Y. (2019). Machine learning models for analysis of vital signs dynamics: a case for

sepsis onset prediction. *J. Healthc. Eng.* 2019:5930379. doi: 10.1155/2019/5930379

Cerner Corporation (2014). *CareAware iBus: Increasing Efficiency and Eliminating Error*.

Christ, M., Braun, N., Neuffer, J., *and* Kempa-Liehr, A. (2018). "Time series feature extraction on basis of scalable hypothesis tests (tsfresh – A Python package). *Neurocomputing* 307, 72–77. doi: 10.1016/j.neucom.2018.03.067

Desautels, T., Calvert, J., Hoffman, J., Jay, M., Kerem, Y., Shieh, L., et al. (2016). Prediction of sepsis in the intensive care unit with minimal electronic health record data: a machine learning approach. *JMIR Med. Inform.* 4:e5909. doi: 10.2196/medinform.5909

Dombrovskiy, V. Y., Martin, A. A., Sunderram, J., and Paz, H. L. (2007). Rapid increase in hospitalization and mortality rates for severe sepsis in the United States: a trend analysis from 1993 to 2003. *Crit. Care Med.* 35, 1244–1250. doi: 10.1097/01.CCM.0000261890.41311.E9

Donnelly, J. P., Locke, J. E., MacLennan, P. A., McGwin, G. Jr, Mannon, R. B., Safford, M. M., et al. (2016). Inpatient mortality among solid organ transplant recipients hospitalized for sepsis and severe sepsis. *Clin. Infect. Dis.* 63, 186–194. doi: 10.1093/cid/ciw295

Elkholy, S., Mansour, D. A., El-Hamid, S. A., Al-Jarhi, U. M., El-Nahaas, S. M., and Mogawer, S. (2019). Risk index for early infections following living donor liver transplantation. *Arch. Med. Sci. AMS.* 15:656. doi: 10.5114/aoms.2019.84736

Fischer, S. A., Lu, K., and AST Infectious Diseases Community of Practice (2013). Screening of donor and recipient in solid organ transplantation. *Am. J. Transplant.* 13, 9–21. doi: 10.1111/ajt.12094

Futoma, J., Simons, M., Doshi-Velez, F., and Kamaleswaran, R. (2021). Generalization in clinical prediction models: the blessing and curse of measurement indicator variables. *Crit. Care Explor.* 3:e0453. doi: 10.1097/CCE.0000000000000453

Gauer, R. L. (2013). Early recognition and management of sepsis in adults: the first six hours. *Am. Fam. Phys.* 88, 44–53.

Haidar, G., Green, M., and American Society of Transplantation Infectious Diseases Community of Practice (2019). Intra-abdominal infections in solid organ transplant recipients: guidelines from the American society of transplantation infectious diseases community of practice. *Clin. Transplant.* 33:e13595. doi: 10.1111/ctr.13595

He, Q., Liu, P., Li, X., Su, K., Peng, D., Zhang, Z., et al. (2019). Risk factors of bloodstream infections in recipients after liver transplantation: a meta-analysis. *Infection* 47, 77–85. doi: 10.1007/s15010-018-1230-5

Hernandez Mdel, P., Martin, P., and Simkins, J. (2015). Infectious complications after liver transplantation. *Gastroenterol. Hepatol. (N. Y.)* 11, 741–753.

Kamaleswaran, R., Akbilgic, O., Hallman, M. A., West, A. N., Davis, R. L., and Shah, S. H. (2018). Applying artificial intelligence to identify physiomarkers predicting severe sepsis in the PICU. *Pediatr. Crit. Care Med.* 19, e495–e503. doi: 10.1097/PCC.0000000000001666

Kim, W. R., Lake, J. R., Smith, J. M., Schladt, D. P., Skeans, M. A., Noreen, S. M., et al. (2019). OPTN/SRTR 2017 annual data report: liver. *Am. J. Transplant.* 19, 184–283. doi: 10.1111/ajt.15276

Kumar, A., Roberts, D., Wood, K. E., Light, B., Parrillo, J. E., Sharma, S., et al. (2006). Duration of hypotension before initiation of effective antimicrobial therapy is the critical determinant of survival in human septic shock*. *Crit. Care Med.* 34, 1589–1596. doi: 10.1097/01.CCM.0000217961.75225.E9

Liu, V., Escobar, G. J., Greene, J. D., Soule, J., Whippy, A., Angus, D. C., et al. (2014). Hospital deaths in patients with sepsis from 2 independent cohorts. *JAMA* 312, 90–92. doi: 10.1001/jama.2014.5804

Lundberg, S. M., and Lee, S. I. (2017). "A unified approach to interpreting model predictions". in *Proceedings of the 31st International Conference on Neural Information Processing Systems* (Red Hook, NY), 4768–4777.

Martin, P., DiMartini, A., Feng, S., Brown, R. Jr., and Fallon, M. (2014). "Evaluation for liver transplantation in adults: 2013 practice guideline by the American association for the study of liver diseases and the American society of transplantation. *Hepatology* 59, 1144–1165. doi: 10.1002/hep.26972

Moore, J. X., Donnelly, J. P., Griffin, R., Howard, G., Safford, M. M., and Wang, H. E. (2016). Defining sepsis mortality clusters in the United States. *Crit. Care Med.* 44, 1380–1387. doi: 10.1097/CCM.0000000000001665

Nemati, S., Holder, A., Razmi, F., Stanley, M. D., Clifford, G. D., and Buchman, T. G. (2018). An interpretable machine learning model for accurate prediction of sepsis in the ICU. *Crit. Care Med.* 46, 547–553. doi: 10.1097/CCM.0000000000002936

Oriol, I., Sabé, N., Melilli, E., Lladó, L., González-Costello, J., Soldevila, L., et al. (2015). Factors influencing mortality in solid organ transplant recipient with bloodstream infection. *Clin. Microbiol. Infect.* 21, 1104.e9–1104.e14. doi: 10.1016/j.cmi.2015.07.021

Pedersen, M., and Seetharam, A. (2014). Infections after orthotopic liver transplantation. *J. Clin. Exp. Hepatol.* 4, 347–360. doi: 10.1016/j.jceh.2014.07.004

Scikit-Learn: Machine Learning in Python (2021). Scikit-learn 0.24.2 Documentation. Available online at: https://scikit-learn.org/stable/ (accessed May 22, 2021).

SHAP (SHapley Additive exPlanations) (2021). SHAP ProgramRepository. Available online at: https://shap.readthedocs.io/en/latest/ (accessed July 15, 2021).

Singer, M., Deutschman, C. S., Seymour, C. W., Shankar-Hari, M., Annane, D., Bauer, M., et al. (2016). The third international consensus definitions for sepsis and septic shock (Sepsis-3). *JAMA* 315:801. doi: 10.1001/jama.2016.0287

Sun, H. Y., Cacciarelli, T. V., and Singh, N. (2011). Identifying a targeted population at high risk for infections after liver transplantation in the MELD era. *Clin. Transplant.* 25, 420–425. doi: 10.1111/j.1399-0012.2010.01262.x

Van der Maaten, L., and Hinton, G. (2008). Visualizing data using t-SNE. *J. Mach. Learn. Res.* 9, 2579–2605.

van Wyk, F., Khojandi, A., Akram, M., Begoli, E., Davis, R. L., and Kamaleswaran, R. (2019). "A minimal set of physiomarkers in continuous high frequency data streams predict adult sepsis onset earlier. *Int. J. Med. Inform.* 122, 55–62. doi: 10.1016/j.ijmedinf.2018.12.002

Woldaregay, A. Z., Årsand, E., Walderhaug, S., Albers, D., Mamykina, L., Botsis, T., et al. (2019). Data-driven modeling and prediction of blood glucose dynamics: machine learning applications in type 1 diabetes. *Artif. Intell. Med.* 98, 109–134. doi: 10.1016/j.artmed.2019.07.007

XGBoost Documentation (2021). XGBoost 1.5.0-SNAPSHOT Documentation. Available online at: https://xgboost.readthedocs.io/en/latest/ (accessed May 22, 2021)

Conflict of Interest: The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

Publisher's Note: All claims expressed in this article are solely those of the authors and do not necessarily represent those of their affiliated organizations, or those of the publisher, the editors and the reviewers. Any product that may be evaluated in this article, or claim that may be made by its manufacturer, is not guaranteed or endorsed by the publisher.

frontiers in Physiology

ORIGINAL RESEARCH
published: 14 October 2021
doi: 10.3389/fphys.2021.716434

The Use of Artificial Neural Networks to Forecast the Behavior of Agent-Based Models of Pathophysiology: An Example Utilizing an Agent-Based Model of Sepsis

*Dale Larie, Gary An and R. Chase Cockrell**

Department of Surgery, Larner College of Medicine, University of Vermont, Burlington, VT, United States

OPEN ACCESS

Edited by:
Carlos D. Maciel,
University of São Paulo, Brazil

Reviewed by:
Erol Eğrioğlu,
Giresun University, Turkey
Rafael Naime Ruggiero,
University of São Paulo, Brazil

****Correspondence:***
R. Chase Cockrell
robert.cockrell@uvm.edu

Specialty section:
This article was submitted to Computational Physiology and Medicine, a section of the journal Frontiers in Physiology

Received: *28 May 2021*
Accepted: *24 September 2021*
Published: *14 October 2021*

Citation:
Larie D, An G and Cockrell RC (2021) The Use of Artificial Neural Networks to Forecast the Behavior of Agent-Based Models of Pathophysiology: An Example Utilizing an Agent-Based Model of Sepsis. Front. Physiol. 12:716434. doi: 10.3389/fphys.2021.716434

Introduction: Disease states are being characterized at finer and finer levels of resolution *via* biomarker or gene expression profiles, while at the same time. Machine learning (ML) is increasingly used to analyze and potentially classify or predict the behavior of biological systems based on such characterization. As ML applications are extremely data-intensive, given the relative sparsity of biomedical data sets ML training of artificial neural networks (ANNs) often require the use of synthetic training data. Agent-based models (ABMs) that incorporate known biological mechanisms and their associated stochastic properties are a potential means of generating synthetic data. Herein we present an example of ML used to train an artificial neural network (ANN) as a surrogate system used to predict the time evolution of an ABM focusing on the clinical condition of sepsis.

Methods: The disease trajectories for clinical sepsis, in terms of temporal cytokine and phenotypic dynamics, can be interpreted as a random dynamical system. The Innate Immune Response Agent-based Model (IIRABM) is a well-established model that utilizes known cellular and molecular rules to simulate disease trajectories corresponding to clinical sepsis. We have utilized two distinct neural network architectures, Long Short-Term Memory and Multi-Layer Perceptron, to take a time sequence of five measurements of eleven IIRABM simulated serum cytokine concentrations as input and to return both the future cytokine trajectories as well as an aggregate metric representing the patient's state of health.

Results: The ANNs predicted model trajectories with the expected amount of error, due to stochasticity in the simulation, and recognizing that the mapping from a specific cytokine profile to a state-of-health is not unique. The Multi-Layer Perceptron neural network, generated predictions with a more accurate forecasted trajectory cone.

Discussion: This work serves as a proof-of-concept for the use of ANNs to predict disease progression in sepsis as represented by an ABM. The findings demonstrate that multicellular systems with intrinsic stochasticity can be approximated with an ANN, but that forecasting a specific trajectory of the system requires sequential updating of the system state to provide a rolling forecast horizon.

Keywords: agent-based model (ABM), machine learning, sepsis, neural networks, time series

BACKGROUND

The characterization of the gene expression or protein level patterns associated with clinical disease, which generally manifest as physiological derangements, has led to attempts to use this type of fine-grained, detailed information to forecast clinical outcomes. This approach underlies the concepts of personalized and precision medicine, where disease characterization in terms of molecular-level features (microstates) are intended to more finely define and distinguish patients who might have otherwise similar physiology (macrostates). Increasingly, machine learning (ML) has been investigated as a means of aiding in the ability to predict and forecast the course of disease. Modern ML generally involves training an artificial neural network (ANN) on a given data set such that the ANN "learns" an underlying function that generates the data. While a powerful method, ML-trained ANNs can be brittle and prone to overfitting, which can lead to their failure when applied in real-world situations (Ross and Swetlitz, 2017; Strickland, 2019; D'Amour et al., 2020). As ML is extremely data-intensive, training is often augmented by the use of synthetic data; however, it is crucial that the generated surrogate/synthetic data effectively replicates the underlying generative process of the real-world system being learned. This issue is less important for such static tasks such as image recognition/classification but takes on considerable importance if time-series/dynamic processes (and therefore functions) are being analyzed. The need to effectively generate synthetic data is accentuated in biomedical applications, where, in general, biomedical data sets are relatively sparse, particularly in terms of time series data needed to predict or forecast the dynamic course of disease. This sparsity is further accentuated when molecular-level biomarker panels are proposed as the means of disease characterization, as currently this information can only be acquired through invasive blood sampling. Thus, there is an inherent tension between the desire for a finer-grained characterization of disease state and limitations in terms of both availability of such data and the ability to correlate these detailed representations to the physiological derangements present clinically.

Multi-scale simulation models that represent cellular and molecular mechanisms can reproduce the dynamics of tissue or system level physiology and pathophysiology have potential as a means of generating synthetic training data, but these models come with their own challenges and limitations. Specifically, dealing with the high-dimensional parameter spaces of such complex mechanism-based models presents computational challenges in terms of calibration and validation. Additional ML methods have been proposed as an adjunct to the exploration of these models' high dimensional parameter spaces (Cockrell et al., 2019; Ozik et al., 2019; Wang et al., 2019), including the training of artificial neural networks (ANNs) as surrogates for the mechanism-based model (Wang et al., 2019). However, to our knowledge, the application of ML to train ANN surrogates for agent-based models (ABMs), a prevalent method for multi-scale computational modeling, has not been previously reported in the biomedical literature. This is a potentially significant capability, as ABMs structurally share many of the features of biological systems (Bonabeau, 2002; An et al., 2009; Metzcar et al., 2019) and exhibit behaviors not necessarily represented by other types of modeling methods, particularly in terms of their stochastic behavior and reflection of biological heterogeneity (Cockrell and An, 2017). The ability of ABMs to generate "emergent" phenomena (Bonabeau, 2002), i.e., where populations of components and their interactions lead to system-level phenomenon that cannot be directly inferred from the behavioral rules governing the components is particularly relevant to being able to translate cellular and molecular mechanisms and data into system-level behavior manifesting as physiology. Cell-based ABMs explicitly represent existing knowledge about cellular and molecular mechanisms, which is the level at which modern medicine strives to characterize patients in a precise and personalized fashion (e.g., biomarker or -omics panels), and through their simulation are able to generate aggregated, system-level output corresponding to the physiology at which disease primarily manifests. Given the prevailing interest in characterizing disease states through molecular-level profiling and the application of ML methods to forecasting the physiological trajectories of disease, we believe that it is important to examine the capabilities and limitations of applying ML to forecast trajectories that bridge microstate (mediator/molecular profiles) and macrostate (system-level/physiological output). Toward that end we present herein an investigation of the ability of trained ANNs to forecast the dynamic behavior of a complex biomedical ABM used to simulate acute systemic inflammation and the clinical condition of sepsis.

The Multi-Scale Challenge of Sepsis

Sepsis is a complex physiological and clinically significant problem with approximately 1 million cases in the United States each year, with a mortality rate between 28–50% (Wood and Angus, 2004). Sepsis is a highly dynamic process with multi-scale features, ranging from clinical phenotypes characterized by features such as multi-system organ failure, down to the molecular level with dysregulation of the body's internal cytokine signaling network (Cockrell and An, 2017, 2019, 2021). While care process improvements in the treatment of sepsis, such as

the development of treatment bundles and practice guidelines, have improved clinical outcomes in the past few decades, the search for new drugs to treat the biological-basis of sepsis has been marked by complete failure: there is currently not a single drug approved by the U.S. Food and Drug Administration that targets the underlying pathophysiology of sepsis (Angus, 2011; Buchman et al., 2016). One of the major challenges in designing therapies for sepsis is an inability to effectively forecast the disease trajectories of individual patients, thereby limiting the effective sub-stratification of this heterogeneous population into those biologically similar enough to control. Existing means of classifying sepsis patients, such as with the Sequential Organ Failure Score (SOFA; Vincent et al., 1996) or various biomarker panels (Gibot et al., 2012; Riedel, 2012; Samraj et al., 2013), while potentially useful for coarse-grained outcome risk stratification, are only able to provide population-level projections that cannot effectively be updated to an individual patient's disease course. Adding to the limitations of data-centric population-based scoring systems is the inherent stochasticity of the biological processes driving sepsis. The presence of stochasticity in the system governing inflammation makes accurately predicting the entire trajectory of the disease, or accurately predicting the patient state 30 days into the future, given one point of assessment, an impossibility (see description of Stochastic Trajectory Analysis regarding sepsis in (Cockrell and An, 2017)).

Biological Heterogeneity, Stochasticity, and Forecasting

Ultimately, the biological heterogeneity seen clinically is generated from a combination of inter-patient (genetic variability) and intra-patient (stochastic processes) effects. The result is that it is not tractable to comprehensively enumerate all possible biomarker states and configurations (i.e., phenotypes) that can be generated from a specific systemic perturbation or injury. The challenge (and solution) is similar to that faced by Q-Learning (Watkins and Dayan, 1992) (now Deep Reinforcement Learning); Q-learning is a type of reinforcement learning in which agents determine what action to take (a) by looking up their current state (s) in the lookup table, $Q(s,a)$ that lists the probability of a desirable outcome based on that decision. Because the lookup table needs to provide this probability to guide the decision process it requires a finite (and computationally tractable) state space. In order to work effectively in continuous (infinite states) search spaces Q-learning utilizes the Universal Approximation Theorem (Barron, 1993), which states that a feed-forward neural network can approximate, to arbitrary fidelity, a real and continuous function. In the case of Q-learning, it is the lookup table that is being approximated; we note that the lookup table does not necessarily meet the strict mathematical definition for continuity, however, the technique works in practice as long as the density/granularity of the lookup table is sufficiently fine. Acquiring time series data of this granularity is often not logistically feasible, therefore we pose that mechanism-based simulations can serve as means of generating such surrogate data. In particular, given their structural similarity to biological systems, ABMs are appealing candidates for this task.

In previous work, we have demonstrated that the cytokine signaling network which controls the inflammatory process can be modeled as a random dynamical system (Cockrell and An, 2017, 2018), which is a system that evolves in time according to fixed rules, but also incorporates stochasticity (Bhattacharya and Majumdar, 2003; Arnold, 2013). Knowledge of the underlying cellular and molecular processes of acute inflammation has been used to create a dynamic model, the Innate Immune Response Agent-based Model (IIRABM; An, 2004), that can serve as a proxy model for the development of more advanced prediction and forecasting methods. The IIRABM is an ABM of the innate immune response that represents the endothelial-blood interface (e.g., the inside of blood vessels in the tissue region of interest) and the response of that system to either injury or infection. The IIRABM simulation is initiated with the application of a simulated injury or infection to the endothelium. The injury is defined by five parameters: injury size, microbial virulence, microbial toxigenesis, environmental toxicity, and host resilience. The simulated inflammatory response is generated by the damaged endothelium, and recruits a variety of inflammatory cells, including neutrophils, macrophages, and a suite of T-lymphocytes, to respond to, contain, and heal the injury/infection. The simulation then proceeds until it reaches a terminal state – either complete healing or death, which is triggered when the aggregate system damage exceeds 80%. This threshold has been chosen to represent the ability of supportive medical technology (i.e., a ventilator) to keep people alive in situations in which they would otherwise die.

Despite its acknowledged abstraction the IIRABM has proven useful in examining the complexity of sepsis and the challenges associated with trying to treat the syndrome. The IIRABM has been used to demonstrate the use of *in silico* clinical trials as a means of evaluating the plausibility of planned potential interventions (An, 2004), provided fundamental insights into the mathematical and dynamic properties of sepsis that account for patient heterogeneity (Cockrell and An, 2017), demonstrating the futility of standard biomarker-based outcome prediction (Cockrell and An, 2017), and served as a proxy model (An et al., 2017) for control discovery for sepsis. This most recent control discovery work has employed advanced computational methods such as genetic algorithms/evolutionary computing (Cockrell and An, 2018) and deep reinforcement learning/artificial intelligence (Petersen et al., 2019) to describe what would be required for multi-modal treatment of sepsis. While the IIRABM is nearly 20 years old its central component structure remains valid and has predicted a series of behaviors associated with sepsis that have since been recognized in the subsequent years, specifically the temporal concurrence of pro- and anti-inflammatory cytokine responses (as opposed to sequential pro- and compensatory responses) (Osuchowski et al., 2006; Tamayo et al., 2011) and the importance of the immunoparalyzed recovery phase of sepsis, particularly with respect to its prolonged duration (Ferguson et al., 1999; Boomer et al., 2011; Hotchkiss et al., 2013a,b). Key to all these studies is the recognition that even though

the IIRABM is an abstract representation far less complex than the "real" immune system, it has structural properties that mimic a multicellular biological system (i.e., composed of semiautonomous components that harbor the system's stochastic potential), and generates the type of system dynamics that challenges traditional methods of biomedical analysis. As such we consider the IIRABM a useful surrogate for generating biology-like synthetic data that bridges mediator-level microstate and system-level microstate output that can be used to examine the ability of ML to capture its behavior. The current work aims to train an ANN on simulated data generated from the IIRABM and evaluate its sufficiency as a surrogate for the IIRABM by assessing the ability of the trained ANN for dynamic trajectory prediction of the IIRABM.

MATERIALS AND METHODS

The foresting procedure is divided into two principal tasks: (1) predict future cytokine trajectories in an 11-dimensional space (microstate characterization); and (2) regress the overall "health" of the simulation as a function of its current cytokine profile (predicting system macrostate). Training and validation data was generated using the IIRABM (Cockrell and An, 2017). The training/validation set was composed of cytokine measurements for 11 unique cytokines over 10,000 time-steps in 66,000 *in silico* patients. Networks were constructed Using Keras (Gulli and Pal, 2017), a TensorFlow based deep learning library for Python.

Trajectory Forecasting

In order to forecast future values in the cytokine time series, we utilized long short-term memory (LSTM) recursive neural networks (RNN). RNNs are different from standard multi-layer-perceptron networks because they have a neural network contained within a cell which takes information from the current input to help determine the adjusted state of the cell based on its current cell state. This adjusted cell state becomes the new cell state, and an output is determined for the network.

Long short-term memory networks' memory cells have a unique structure, characterized by an input gate, two update layers, and an output gate to determine the adjusted cell state (Hochreiter and Schmidhuber, 1997). The memory cells in LSTM networks allow for more long term memory than typical RNNs which make them well suited for time-series analysis and prediction (Nelson et al., 2017). Noting this, an LSTM network will likely be able to predict future cytokine levels, given that they are continuous and previous cytokine levels will likely have a large impact on near-future values.

We constructed a unique network for each cytokine that was to be predicted; each LSTM network takes five sequential 11-dimensional cytokine profiles as input and predicts the subsequent value(s). The first three layers of the network are 100-node LSTM layers; the output from these layers are fed into two fully connected layers of 300 and 200 nodes, respectively, then to a single output node, resulting in 296,301 trainable parameters. Training data was arranged into five sequential 11-dimensional points as training input features and the next 11-dimensional point as the training label. The data was then shuffled to avoid biasing the training. After data preprocessing, 8,576,100 data sequences and labels were used to train the network. The loss metric used to train this network is mean absolute error (MAE), and the Adam optimizer (Kingma and Ba, 2014). Each network was trained until loss converged to a minimum.

For the ultimate utilization of this network, 11 cytokine values are observed for five time steps, then a prediction for each of the next values is made using its own LSTM network. This set of 11 observations is combined into one 11-dimensional point, which is then added to the original five samples as the next sample. Predictions are made recursively in this manner for 100 time steps after the initial observation. Accuracy of this algorithm was measured using the average MSE across the 11 cytokine values at 1, 2, 3, 4, 5, 10, 25, 50, and 100 time steps after the initial observation. Prediction variance and error bars were calculated through stochastic variations to the dropout layer (Baldi and Sadowski, 2013), as demonstrated with regards to Active Learning for regression in (Tsymbalov et al., 2018).

As a comparison of the efficacy of LSTM neural networks, MLP prediction networks for each cytokine were also created. Each network functionally acts the same as the LSTM networks, accepting five sequential 11-dimensional points in cytokine space and predicting the future value for a single cytokine. Each network has a structure beginning with a fully connected layer of 1000 nodes, followed by a function to flatten the output shape from a 5 by 1000 array to a single vector of length 5000. Next is another fully connected layer of 1000 nodes, then a 1% permanent dropout layer, feeding into a fully connected layer of 500 nodes, then another fully connected layer of 500 nodes, then finally to a single output node. These networks were trained using a loss function to minimize MSE. Eleven-dimensional cytokine profiles, predicted either from the LSTM network the MLP network are then fed into an MLP-regressor (described below) in order to translate the cytokine profile into an aggregate measure of patient health or disease state.

Health Metric Regression

The IIRABM uses the "Oxygen Deficit" metric as a measure of health, where a low oxygen deficit is good, and a high oxygen deficit is bad; "Oxygen Deficit" is therefore the system-level output that corresponds to the macrostate of the IIRABM. We note that, both *in silico* and *in vivo*, cytokine profiles provide a non-unique mapping to state-of-health (a concept which is more nebulously defined *in vivo* than in our *in silico* model). As such, error is expected when attempting to regress from an 11-dimensional cytokine profile to a single health metric.

This regression was performed by using a fully connected deep network that takes an 11-dimensional cytokine vector as input, feeding into two fully connected layers with 1,500 nodes each, then into a layer with 150 nodes, and finally to a single output node. The loss metric used to train this algorithm is MSE. Using the regression network, a prediction of oxygen deficit trajectory can be made from the 11-dimensional matrix created by the LSTM network. Overall accuracy was measured by comparing the oxygen deficit path to the predicted path and calculating the MSE.

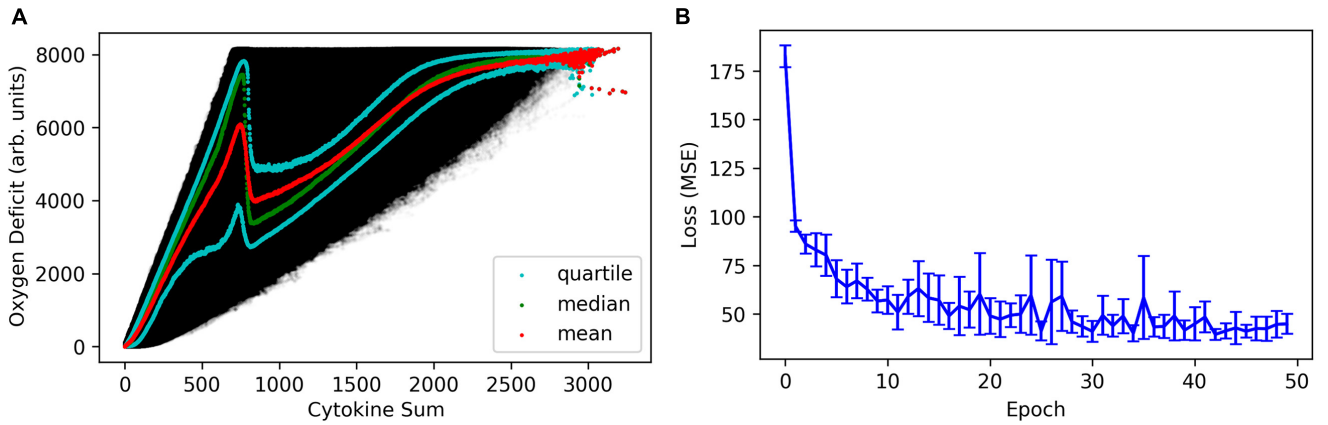


FIGURE 1 | In panel **(A)**, we present the variance in oxygen deficit as a function of the sum of cytokine concentrations in the whole area of simulated tissue, effectively compressing an 11-dimensinal vector into a scalar quantity indicative of total biological activity (e.g., cells performing functions) in the simulated tissue. In panel **(B)**, we show the mean absolute error in the regression of the oxygen deficit as a function of cytokine profile, as a function of training epoch.

The prediction of whether an *in silico* patient will live or die, or the decision on whether or not pharmacologic therapeutics are more likely to be beneficial than detrimental, is ultimately based on the temporal trajectory of the patient's state of health (in this case, a measure of systemic oxygen deficit). The predicted trajectory in 11-dimensional cytokine space is then fed into the health-metric regression network to forecast the most-likely outcome, time-to-outcome, and time-horizon for potentially effective therapeutic interventions.

Additionally, we created a multi-layer perceptron (MLP) to predict the future oxygen deficit trajectory as a function of past values only, effectively treating the simulation output as a Markov chain. This network expects an input of five sequential oxygen deficit values and will return a single future oxygen deficit value predicted for the next time step. The structure of this MLP begins with a fully connected layer of 1000 nodes, followed by a 1% permanent dropout layer, then two fully connected layers of 150 nodes each, followed by a single output node. This network was trained using a loss function to minimize MSE. Trajectory prediction for this network is made in the same recursive manner as the cytokine prediction networks.

RESULTS

It is important to note that the map which translates a cytokine profile (microstate) into its associated oxygen deficit (macrostate; and vice-versa) is non-unique, and therefore some amount of error is expected and unavoidable. In **Figure 1A**, we present the variance in oxygen deficit as a function of the sum of cytokine concentrations in the whole area of simulated tissue. The sum of cytokine concentrations is a coarse metric that roughly represents the amount of inflammation (no distinction is made between pro- and anti-inflammatory signals) and inflammatory signaling present in the model. This is analogous to what is seen clinically – patients that see ostensibly identical insults/infections/injuries will invariably present a range of responses, in terms of temporal cytokine profiles or other clinical physiological observables (heart-rate, blood pressure, temperature, etc.).

This figure also illustrates a key difference between the structure of the noise in the IIRABM and the stochastic structure in a stochastic differential equation; the noise present in the IIRABM varies spatio-temporally and cannot be represented with a closed-form analytical expression. Very generally speaking, the reason for this is that when cell-signaling is high, there is lots of activity in the model, and therefore lots of opportunities for stochastic events. This can be illustrated with a simple thought experiment: consider two system states, one with a single infected cell and one with 10 infected cells, and each infected cell has a probability of infecting a single neighbor, and some probability of healing. If we evolve the simulation a single time step, system 1 can have 0, 1, or 2 infected cells, while the range of infected cells in system 2 can vary from 0 to 20 (depending on the spatial configuration of the infected cells). In **Figure 1B**, we show the mean absolute error as a function of training epoch when training the health regression neural net. The error quickly converges to a minimum with a relatively constant error of approximately 200 units (on a scale of 8160), with the caveat that the predicted error would be lower when the true oxygen deficit is lower, and higher when the true oxygen deficit is higher.

Cytokine trajectories present similar stochastic properties as the oxygen deficit: when levels are high, the plausible range of cytokine expression for the subsequent time step is larger than when levels are low. We present the mean squared error (in arbitrary units) as a function of training epoch for TNFα, which is representative of the full cytokine set in **Figure 2**. Once again, the network quickly converges to a low and constant level of error. We note that the total error quickly and significantly increases as we extend the time-prediction horizon past 100 time-steps. This distinguishes this methodology from that of ML-augmented surrogate modeling (Cicchese et al., 2017) because we do not claim the ability to accurately represent the entire course of a sepsis disease trajectory (up to 90 days in our computational model) using neural-network approximations.

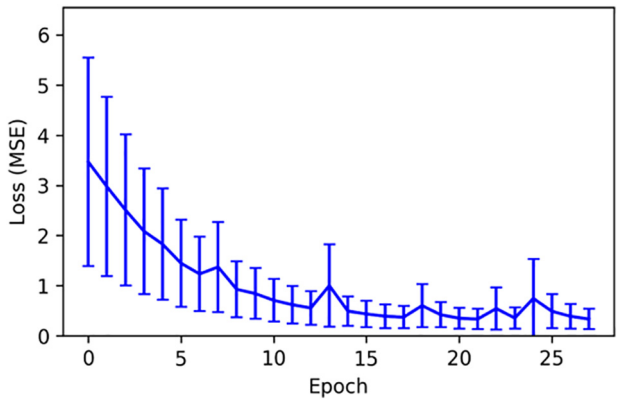


FIGURE 2 | The mean squared error (in arbitrary units) as a function of training epoch for TNFα. The training of this cytokine prediction network was representative of all simulated cytokine prediction networks.

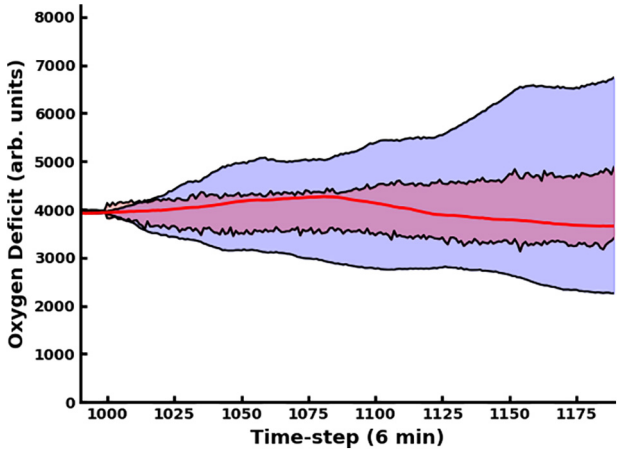


FIGURE 3 | In this figure, we compare the dual-network predictor, which uses an MLP to predict the 11-dimensional cytokine profile trajectory and then another MLP to regress the oxygen deficit value (shaded in red), with an MLP which is informed solely by prior values for the oxygen deficit (shaded in blue).

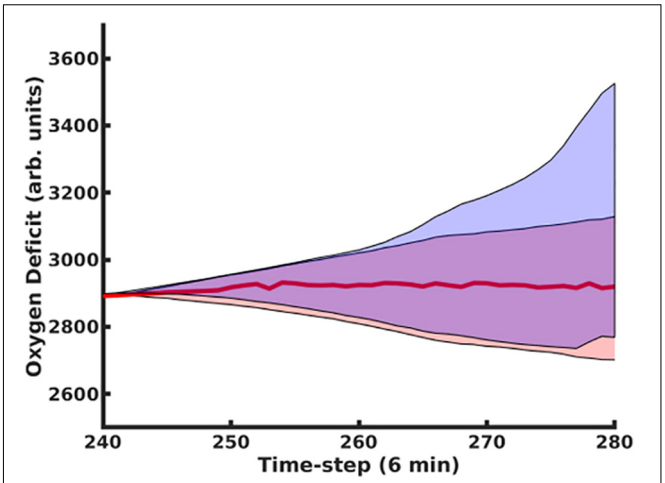


FIGURE 4 | Here we present trajectory clouds for the MLP and LSTM-generated predictions. In both cases, 11-dimensional cytokine trajectories were predicted. The predicted values were then fed into the MLP regressor network, which regressed the future predicted oxygen deficit as a function of the predicted cytokine profile trajectories. The future health-trajectory probability cloud predicted using the MLP network is shaded in red; the future health-trajectory probability cloud predicted using the LSTM network is shaded in blue; the true trajectory is plotted in red.

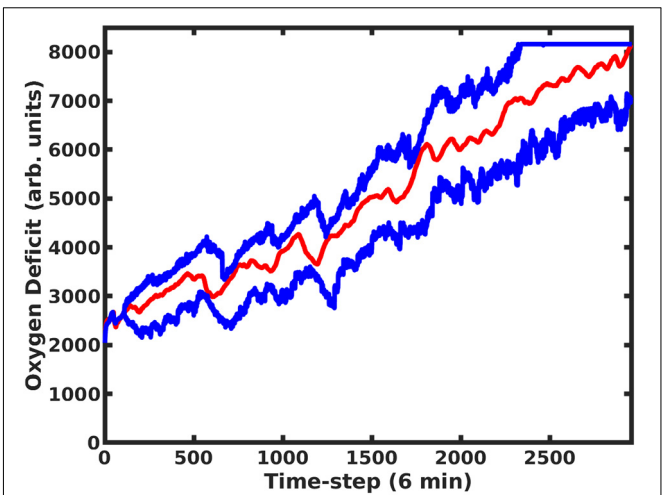


FIGURE 5 | Using the MLP Trajectory Forecast Model combined with the MLP Oxygen-Deficit Regressor, we determined the upper and lower boundaries for the future health-trajectory probability cone, which are indicated by the blue lines; the actual trajectory is plotted with a red line. Predictions began at t = 200 and were update upon every time-step.

The use of the dropout layer allows for the simple creation of an ensemble of predictive networks by stochastically varying the specific node(s) in the layer that are dropped out, allowing us to visualize probability clouds for future trajectories. In **Figure 3**, we compare the dual-network predictor, which uses an MLP to predict the 11-dimensional cytokine profile trajectory and then another MLP to regress the oxygen deficit value (shaded in red), with an MLP which is informed solely by prior values for the oxygen deficit (shaded in blue); the performance of the dual-network model is significantly more stable and accurate when compared with the MLP using only the oxygen deficit for prediction. In **Figure 4**, we have visualized the probability cloud for future health trajectories generated using the MLP network (shaded in red), future health trajectories generated using the LSTM network (shaded in blue) and plotted the true trajectory (red line). This figure visualizes a single prediction iteration (predict future cytokine trajectory, regress state of health) for the above-described workflow. As new data is fed into the model about the true trajectory of the system, the forecast cloud is updated. The actual health trajectory typically lies in the center of the probability cloud, which is a clear benefit of the ensemble approach. In **Figure 5**, we display the probability cone for the entire simulation run, starting at the 240th time step, and then updating the trajectory cone on every subsequent time step.

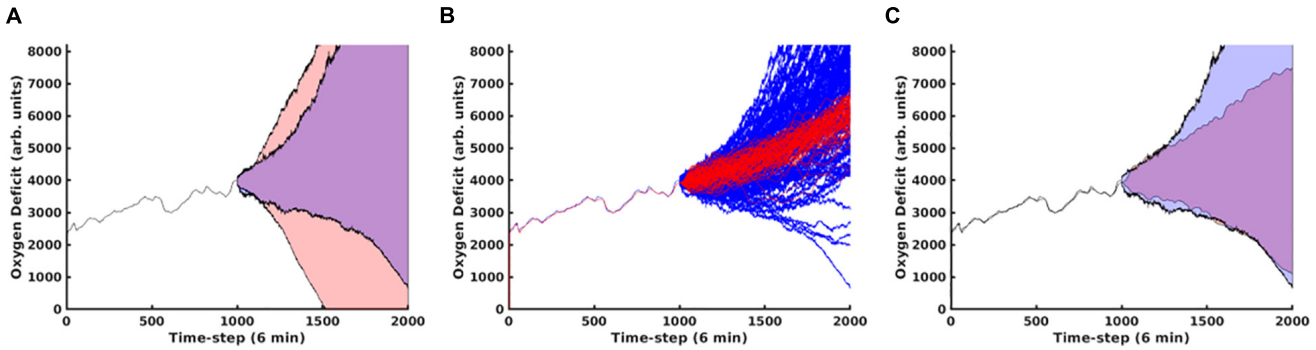


FIGURE 6 | In this figure, we compare predictions generated using the MLP Cytokine Trajectory Prediction/MLP Oxygen-Deficit Regressor with a set of stochastic replicates generated by the simulation. In panel **(A)**, we contrast the future health-trajectory probability cones with networks that used data collected in the first 24 h (shaded in red) and networks that excluded the first-day data (shaded in blue); we note that the blue area appears purple as it is entirely contained within the red area. In panel **(B)**, we present 100 stochastic replicates of the actual simulation health-trajectory, reseeded at the time of the first prediction (red) compared with the predicted probability cone trajectories (blue). In panel **(C)**, we generated the simulation trajectory cone through reseeding the simulation's random number generator at the upper and lower boundaries of the trajectory cone every 100 time steps from t = 1100 to t = 1800.

In **Figure 6**, Panel A, we contrast predictions that considered the full time evolution of the system when training the neural network model, shaded in red, with predictions that only used training data collected after the 240th time step, representing approximately 1 day. The network that only utilizes data collected more than 24 h post-injury performs substantially better. This is primarily due to the massive amount of stochasticity introduced at the time of injury; the degree of this stochasticity is significantly larger in magnitude than later in the simulation, as discussed below. In Panel B, we display the same oxygen-deficit probability cone as in Panel A, however, also reseed the simulation's random number generator at this time step to generate 100 stochastic replicates of the time evolution of that specific instantiation of the IIRABM. We see that the predicted probability cone has a greater spread than the actual probability cone, however, we note that the MLP predictor is constantly updating its trajectory predictions: the set of observations $\{t^a_{-5}, t^a_{-4}, t^a_{-3}, t^a_{-2}, t^a_{-1}\}$, where t^a_{-5} has the superscript, "a", representing an actual observation, and the subscript, "−5" to denote that the time point is five points prior to the starting reference point, is used to predict t^p_0, where the superscript, "p", indicates a predicted observation. Eventually, the set of points used to generate the prediction will consist entirely of previously predicted points, allowing for the compounding of any errors. In Panel C, we show the same probability cone as in Panel A, however, this time, we have re-seeded the simulation every 100 time steps at t = 1100 to t = 1800, for 100 stochastic replicates each. This is a more direct comparison since the MLP predictor effectively reseeds itself every time step. We observe that the actual probability cone is significantly wider than in Panel B, but still not as wide as the predicted cone. This is discussed in detail below.

DISCUSSION

The MLP predictor which predicts cytokine trajectories and uses those to regress the oxygen deficit performs better than using an LSTM to predict future state-of-health trajectories, however, this does *not* represent a failure of the LSTM method (or indicate superiority over the MLP). This is best illustrated in **Figure 1**, which illustrates the non-unique mapping between a specific cytokine profile and a physiological state of health, which is well-known clinically (Bergquist et al., 2019). The accuracy of the cytokine trajectory predictions, shown in **Figure 2**, is high, but even with an accurate prediction of the future cytokine profile, turning that profile into an informative state of health prediction is not possible. Additionally, this predictor also outperforms the NN model which used an MLP to predict future oxygen deficit from previous values of oxygen deficit, not incorporating any cytokine data (see **Figure 3**). This indicates that, while a static cytokine profile can be correlated with a wide range of health-state, the dynamics/trajectories of these mediators provide actional information regarding system state.

Additionally, we note that the predictor performs better when using training data starting 1 day after the simulated injury perturbs the system, and the reasoning for this is similar to that above, in that the cytokine and spatial dynamics are dominated by stochasticity. When the simulated injury occurs, a large, contiguous, area of tissue is injured with a homogenous injury, representing a significant perturbation to the system; thus, early simulation behaviors contain a significant amount of stochasticity, leading the training data to be less informative as to the true mechanisms which underlie the dynamics of the simulation.

Recognizing that there are configurations in which the system is more or less strongly influenced by randomness can also help to explain why treating the global simulation output as a Markov chain (or full cytokine trajectory output as a Markov random field, as we have described in (Cockrell and An, 2017)) is only an approximation. The full simulation, which takes place on a two-dimensional grid, is memoryless, and begins with a homogenous injury. However, as the injury evolves, the spatial distribution of damage or of various cytokine concentrations begins to vary, due to both stochastic and deterministic influences. All

this information about spatial heterogeneity is lost when it is collapsed into a single trajectory. A Markov Transition Matrix (or kernel) could be constructed for the simulation output that would be true in the comprehensive sense, e.g., when considering all possible trajectories and model configurations, however, the utility of this information becomes more limited the farther out the prediction goes, as seen in **Figure 5**.

The model reseedings in **Figure 6** indicate to use that the model is in a very deterministic configuration. Due to the spatial distribution of the injury, there is essentially no chance that it will heal the *in silico* patient back to full health, while also being in no danger of an immediate/near-term death. Essentially, the simulations are entirely under the influence of a single Probabilistic Basin of Attraction; this is discussed in detail in Cockrell and An (2017). Therefore, while the fully spatially realized simulation does not have a memory (and can safely be treated as a Markov process), the historical paths of the aggregate cytokine/health trajectories do provide some predictive ability; while information regarding the spatial distribution of tissue damage and systemic response is washed away when considering the trajectory of the system as a whole, features that describe the time evolution of the trajectory (i.e., temporal derivatives) play a role in the future predictions.

The failure to identify effective drugs to treat sepsis is due in significant part to a failure to account for the heterogeneity of the state space for sepsis and the non-uniqueness of mapping from state space to trajectory space: without understanding the potential future histories of an individual patient from any point in time there can be no rationally justified attempt at controlling or steering that patient's eventual outcome. We have proposed that mechanism-based multi-scale computational models (as defined by the National Institutes of Health Interagency Modeling and Analysis Group[1]) can serve as proxy systems that can address the "Denominator Problem" that arises out of the non-uniqueness of the mapping between system state and behavior and the inevitable sparsity of biological data (An, 2018); we pose that the IIRABM represents one example of a proxy model for sepsis. However, for multiscale modeling and simulation to be deployed in clinical practice, it must be practical to utilize the models in a clinical setting. As we have shown in previous work (Cockrell and An, 2017), this requires an immense amount of computational power as the simulation must be repeated for many stochastic replicates. Compressing/approximating the information and dynamics contained within the computational model using an ANN allows for a computationally cheap and tractable method of rapidly updating predictions about patient disease trajectory as new information becomes available. Therefore, ability to predict requires an additional layer of surrogate models to render such prediction clinically tractable, and the complexity of the dynamic structure of inflammation/sepsis calls for the use of ANNs for this purpose. While there have been some attempts to use ANNs to serve as surrogates for multiscale models (Lagaris et al., 1998; Wang et al., 2019), these approaches involve the approximation of models that are based on known and explicitly described functions. Given that the Universal Approximation Theorem states that an ANN can be trained to reproduce any function, knowing the target function beforehand provides a greater likelihood of success. This is in direct contrast to ABM, which are generally explicitly used because they have no equivalent equation-based formulation. In particular, it is the ability to generate more biologically realistic probability distributions of behavior, as seen in **Figure 1A** and Ref (Cockrell and An, 2017). We posit that this is due to the nature of the noise in ABM's compared to stochastic differential equation methods. In contrast to a differential equation, an ABM does not typically have closed form expression describing the randomness in the simulation; rather, randomness in the execution of the ABM is biologically motivated and incorporates aspects of observed biological heterogeneity (i.e., the spatial distribution of tissue resident macrophages in our prior sepsis simulations). Therefore, ANNs trained on ABMs inherently have a forecast horizon and prediction/forecasting applications of such ANNs need to account for updates of system state in order to provide a "rolling" forecasting cone. The concept is similar to that as seen in weather prediction, with the notable difference that in weather models the future uncertainty is due to deterministic chaos whereas in the ABM/biological system it is due to intrinsic aleatory stochasticity.

This current work the first step of the development of a workflow that integrates mechanism-based ABMs with ML in order to train predictive ANNs that can inform what sort of sensing technology and capabilities need to be developed in the real world. The demonstration of the non-unique mapping between system-microstate (in terms of cytokine profiles) and an overall metric of system macrostate (e.g., system health) suggests that data-centric attempts to develop predictive models, which at their root involve reverse engineering causal relationships between microstate and macrostate, are futile. We assert that it is only through mechanism-based, generative simulations that the sufficient density of time series data can be made available to parse the multiple trajectories that can arise from a particular system state. The basis for this assertion lies in the fact that the computational mechanistic model (as opposed to a data-driven statistical model) limits the future possibility space to one that can evolve from experimentally validated biological (microstate) mechanisms, whereas a statistical model with a sufficient number of terms can fit any data set arbitrarily well. While this space or the number of potential configurations a biological system can exist in is combinatorially/astronomically large, it is not infinite.

Our simulations also demonstrate the crucial role that aleatory stochasticity plays in the time evolution of these multi-agent/cellular systems, thereby necessitating a "rolling forecast" approach in which the update interval is informed by the simulations. Integrating mechanism-based, multi-scale simulations with ML provides a means of training predictive ANNs "off-line," circumventing the need to run high-fidelity simulations in real time. Freed from the constraint of execution time, this in turn allows for more detailed mechanistic simulation

[1]https://www.imagwiki.nibib.nih.gov/content/multiscale-modeling-msm-consortium

models able to generate synthetic data that more closely matches that produced by the real world system, and, crucially, performed in a fashion that more comprehensively captures the range of biological heterogeneity seen in the clinical setting and has the ability to potentially falsify the model's underlying structure (for a full description of this process see Ref (Cockrell and An, 2021)). We hope that this work can provide a starting point for additional investigations into the integration of ML and agent-based models.

DATA AVAILABILITY STATEMENT

The original contributions presented in the study are included in the article/supplementary material, further inquiries can be directed to the corresponding author.

AUTHOR CONTRIBUTIONS

RC and GA conceived the model and the experiments. DL performed the machine-learning work and analysis. All authors contributed to the manuscript.

FUNDING

This work was funded by the Defense Advanced Research Projects Agency through contract HR00111950027, and by the National Institutes of Health through Grant No. U01EB025825. Additionally, this research used high performance computing resources provided by the Vermont Advanced Computing Core (VACC).

REFERENCES

An, G. (2004). In silico experiments of existing and hypothetical cytokine-directed clinical trials using agent-based modeling. *Crit. Care Med.* 32, 2050–2060. doi: 10.1097/01.ccm.0000139707.13729.7d

An, G. (2018). The crisis of reproducibility, the denominator problem and the scientific role of multi-scale modeling. *Bull. Math. Biol.* 80, 3071–3080. doi: 10.1007/s11538-018-0497-0

An, G., Fitzpatrick, B. G., Christley, S., Federico, P., Kanarek, A., Miller Neilan, R., et al. (2017). Optimization and control of agent-based models in biology: a perspective. *Bull. Math. Biol.* 79, 63–87. doi: 10.1007/s11538-016-0225-6

An, G., Mi, Q., Dutta-Moscato, J., and Vodovotz, Y. (2009). Agent-based models in translational systems biology. *Wiley Interdiscip. Rev. Syst. Biol. Med.* 1, 159–171. doi: 10.1002/wsbm.45

Angus, D. C. (2011). The search for effective therapy for sepsis: back to the drawing board? *JAMA* 306, 2614–2615. doi: 10.1001/jama.2011.1853

Arnold, L. (2013). *Random Dynamical Systems*. Berlin: Springer Science & Business Media.

Baldi, P., and Sadowski, P. J. (2013). Understanding dropout. *Adv. Neural Inform. Proc. Syst.* 26, 2814–2822.

Barron, A. R. (1993). Universal approximation bounds for superpositions of a sigmoidal function. *IEEE Trans. Inf. Theory* 39, 930–945. doi: 10.1109/18.256500

Bergquist, M., Hästbacka, J., Glaumann, C., Freden, F., Huss, F., and Lipcsey, M. (2019). The time-course of the inflammatory response to major burn injury and its relation to organ failure and outcome. *Burns* 45, 354–363. doi: 10.1016/j.burns.2018.09.001

Bhattacharya, R., and Majumdar, M. (2003). Random dynamical systems: a review. *Econ. Theory* 23, 13–38.

Bonabeau, E. (2002). Agent-based modeling: methods and techniques for simulating human systems. *Proc. Natl. Acad. Sci.* 99(Suppl 3), 7280–7287. doi: 10.1073/pnas.082080899

Boomer, J. S., To, K., Chang, K. C., Takasu, O., Osborne, D. F., Walton, A. H., et al. (2011). Immunosuppression in patients who die of sepsis and multiple organ failure. *JAMA* 306, 2594–2605. doi: 10.1001/jama.2011.1829

Buchman, T. G., Billiar, T. R., Elster, E., Kirk, A. D., Rimawi, R. H., Vodovotz, Y., et al. (2016). Precision medicine for critical illness and injury. *Crit. Care Med.* 44, 1635–1638. doi: 10.1097/ccm.0000000000002028

Cicchese, J. M., Pienaar, E., Kirschner, D. E., and Linderman, J. J. (2017). Applying optimization algorithms to tuberculosis antibiotic treatment regimens. *Cell. Mol. Bioeng.* 10, 523–535. doi: 10.1007/s12195-017-0507-6

Cockrell, C., and An, G. (2017). Sepsis reconsidered: identifying novel metrics for behavioral landscape characterization with a high-performance computing implementation of an agent-based model. *J. Theor. Biol.* 430, 157–168. doi: 10.1016/j.jtbi.2017.07.016

Cockrell, C., and An, G. (2019). Genetic algorithms for model refinement and rule discovery in a high-dimensional agent-based model of inflammation. *bioRxiv* [Preprint]. doi: 10.1101/790394

Cockrell, C., and An, G. (2021). Utilizing the heterogeneity of clinical data for model refinement and rule discovery through the application of genetic algorithms to calibrate a high-dimensional agent-based model of systemic inflammation. *Front. Physiol.* 12:662845. doi: 10.3389/fphys.2021.662845

Cockrell, C., Ozik, J., Collier, N., and An, G. (2019). Nested active learning for efficient model contextualization and parameterization: pathway to generating simulated populations using multi-scale computational models. *Simulation* 97, 287–296. doi: 10.1177/0037549720975075

Cockrell, R. C., and An, G. (2018). Examining the controllability of sepsis using genetic algorithms on an agent-based model of systemic inflammation. *PLoS Comput. Biol.* 14:e1005876. doi: 10.1371/journal.pcbi.1005876

D'Amour, A., Heller, K., Moldovan, D., Adlam, B., Alipanahi, B., Beutel, A., et al. (2020). Underspecification presents challenges for credibility in modern machine learning. *arXiv* [Preprint] https://arxiv.org/abs/2011.03395

Ferguson, N., Galley, H., and Webster, N. (1999). T helper cell subset ratios in patients with severe sepsis. *Intensive Care Med.* 25, 106–109. doi: 10.1007/s001340050795

Gibot, S., Béné, M. C., Noel, R., Massin, F., Guy, J., Cravoisy, A., et al. (2012). Combination biomarkers to diagnose sepsis in the critically ill patient. *Am. J. Respir. Crit. Care Med.* 186, 65–71. doi: 10.1164/rccm.201201-0037OC

Gulli, A., and Pal, S. (2017). *Deep Learning With Keras*. Birmingham: Packt Publishing Ltd.

Hochreiter, S., and Schmidhuber, J. (1997). Long short-term memory. *Neural Comput.* 9, 1735–1780. doi: 10.1162/neco.1997.9.8.1735

Hotchkiss, R. S., Monneret, G., and Payen, D. (2013a). Immunosuppression in sepsis: a novel understanding of the disorder and a new therapeutic approach. *Lancet Infect. Dis.* 13, 260–268. doi: 10.1016/S1473-3099(13)70001-X

Hotchkiss, R. S., Monneret, G., and Payen, D. (2013b). Sepsis-induced immunosuppression: from cellular dysfunctions to immunotherapy. *Nat. Rev. Immunol.* 13, 862–874. doi: 10.1038/nri3552

Kingma, D. P., and Ba, J. (2014). Adam: a method for stochastic optimization. *arXiv* [Preprint] https://arxiv.org/abs/1412.6980

Lagaris, I. E., Likas, A., and Fotiadis, D. I. (1998). Artificial neural networks for solving ordinary and partial differential equations. *IEEE Trans. Neural Netw.* 9, 987–1000. doi: 10.1109/72.712178

Metzcar, J., Wang, Y., Heiland, R., and Macklin, P. (2019). A review of cell-based computational modeling in cancer biology. *JCO Clin. Cancer Inform.* 3, 1–13. doi: 10.1200/CCI.18.00069

Nelson, D. M., Pereira, A. C., and de Oliveira, R. A. (2017). "Stock market's price movement prediction with LSTM neural networks," in *Proceedings of the 2017 International Joint Conference on Neural Networks (IJCNN)*, (Anchorage, AK: IEEE). doi: 10.1109/IJCNN.2017.7966019

Osuchowski, M. F., Welch, K., Siddiqui, J., and Remick, D. G. (2006). Circulating cytokine/inhibitor profiles reshape the understanding of the SIRS/CARS

continuum in sepsis and predict mortality. *J. Immunol.* 177, 1967–1974. doi: 10.4049/jimmunol.177.3.1967

Ozik, J., Collier, N., Heiland, R., An, G., and Macklin, P. (2019). Learning-accelerated discovery of immune-tumour interactions. *Mol. Syst. Des. Eng.* 4, 747–760. doi: 10.1039/C9ME00036D

Petersen, B. K., Yang, J., Grathwohl, W. S., Cockrell, C., Santiago, C., An, G., et al. (2019). Deep reinforcement learning and simulation as a path toward precision medicine. *J. Comput. Biol.* 26, 597–604. doi: 10.1089/cmb.2018.0168

Riedel, S. (2012). Procalcitonin and the role of biomarkers in the diagnosis and management of sepsis. *Diagn. Microbiol. Infect. Dis.* 73, 221–227. doi: 10.1016/j.diagmicrobio.2012.05.002

Ross, C., and Swetlitz, I. (2017). *IBM Pitched Its Watson Supercomputer As a Revolution in Cancer Care. It's Nowhere Close*. Boston, MA: Stat.

Samraj, R. S., Zingarelli, B., and Wong, H. R. (2013). Role of biomarkers in sepsis care. *Shock* 40:358. doi: 10.1097/SHK.0b013e3182a66bd6

Strickland, E. (2019). IBM Watson, heal thyself: how IBM overpromised and underdelivered on AI health care. *IEEE Spectr.* 56, 24–31. doi: 10.1109/MSPEC.2019.8678513

Tamayo, E., Fernández, A., Almansa, R., Carrasco, E., Heredia, M., Lajo, C., et al. (2011). Pro- and anti-inflammatory responses are regulated simultaneously from the first moments of septic shock. *Eur. Cytokine Netw.* 22, 82–87. doi: 10.1684/ecn.2011.0281

Tsymbalov, E., Panov, M., and Shapeev, A. (2018). "Dropout-based active learning for regression," in *Proceedings of the International Conference on Analysis of Images, Social Networks and Texts*, (New York, NY: Springer). doi: 10.1007/978-3-030-11027-7_24

Vincent, J.-L., Moreno, R., Takala, J., Willatts, S., Mendonça, A. De, Bruining, H., et al. (1996). The SOFA (Sepsis-related Organ Failure Assessment) score to describe organ dysfunction/failure. *Intensive Care Med.* 22, 707–710. doi: 10.1007/BF01709751

Wang, S., Fan, K., Luo, N., Cao, Y., Wu, F., Zhang, C., et al. (2019). Massive computational acceleration by using neural networks to emulate mechanism-based biological models. *Nat. Commun.* 10:4354. doi: 10.1038/s41467-019-12342-y

Watkins, C. J., and Dayan, P. (1992). Q-learning. *Mach. Learn.* 8, 279–292. doi: 10.1007/BF00992698

Wood, K. A., and Angus, D. C. (2004). Pharmacoeconomic implications of new therapies in sepsis. *Pharmacoeconomics* 22, 895–906. doi: 10.2165/00019053-200422140-00001

Author Disclaimer: The content of the information does not necessarily reflect the position or the policy of the Government, and no official endorsement should be inferred.

Conflict of Interest: The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

Publisher's Note: All claims expressed in this article are solely those of the authors and do not necessarily represent those of their affiliated organizations, or those of the publisher, the editors and the reviewers. Any product that may be evaluated in this article, or claim that may be made by its manufacturer, is not guaranteed or endorsed by the publisher.

frontiers in Physiology

ORIGINAL RESEARCH
published: 26 January 2022
doi: 10.3389/fphys.2021.718276

An Evolutionary Algorithm to Personalize Stool-Based Colorectal Cancer Screening

Luuk A. van Duuren[1], Jonathan Ozik[2], Remy Spliet[3], Nicholson T. Collier[2], Iris Lansdorp-Vogelaar[1] and Reinier G. S. Meester[1]*

[1] Department of Public Health, Erasmus University Medical Center, Rotterdam, Netherlands, [2] Decision and Infrastructure Sciences, Argonne National Laboratory, Lemont, IL, United States, [3] Econometric Institute, Erasmus University Rotterdam, Rotterdam, Netherlands

Background: Fecal immunochemical testing (FIT) is an established method for colorectal cancer (CRC) screening. Measured FIT-concentrations are associated with both present and future risk of CRC, and may be used for personalized screening. However, evaluation of personalized screening is computationally challenging. In this study, a broadly applicable algorithm is presented to efficiently optimize personalized screening policies that prescribe screening intervals and FIT-cutoffs, based on age and FIT-history.

Methods: We present a mathematical framework for personalized screening policies and a bi-objective evolutionary algorithm that identifies policies with minimal costs and maximal health benefits. The algorithm is combined with an established microsimulation model (MISCAN-Colon), to accurately estimate the costs and benefits of generated policies, without restrictive Markov assumptions. The performance of the algorithm is demonstrated in three experiments.

Results: In Experiment 1, a relatively small benchmark problem, the optimal policies were known. The algorithm approached the maximum feasible benefits with a relative difference of 0.007%. Experiment 2 optimized both intervals and cutoffs, Experiment 3 optimized cutoffs only. Optimal policies in both experiments are unknown. Compared to policies recently evaluated for the USPSTF, personalized screening increased health benefits up to 14 and 4.3%, for Experiments 2 and 3, respectively, without adding costs. Generated policies have several features concordant with current screening recommendations.

Discussion: The method presented in this paper is flexible and capable of optimizing personalized screening policies evaluated with computationally-intensive but established simulation models. It can be used to inform screening policies for CRC or other diseases. For CRC, more debate is needed on what features a policy needs to exhibit to make it suitable for implementation in practice.

Keywords: colorectal cancer, personalized screening, fecal immunochemical test, screening interval, cutoff, microsimulation models, evolutionary algorithm, FIT-history

OPEN ACCESS

Edited by:
Nicole Y. K. Li-Jessen, McGill University, Canada

Reviewed by:
Dominic G. Whittaker, University of Nottingham, United Kingdom
Tianhong Dai, Imperial College London, United Kingdom

****Correspondence:***
Luuk A. van Duuren
l.vanduuren@erasmusmc.nl

Specialty section:
This article was submitted to Computational Physiology and Medicine, a section of the journal Frontiers in Physiology

Received: *31 May 2021*
Accepted: *21 December 2021*
Published: *26 January 2022*

Citation:
van Duuren LA, Ozik J, Spliet R, Collier NT, Lansdorp-Vogelaar I and Meester RGS (2022) An Evolutionary Algorithm to Personalize Stool-Based Colorectal Cancer Screening. Front. Physiol. 12:718276. doi: 10.3389/fphys.2021.718276

1. INTRODUCTION

Colorectal cancer (CRC) is an important cause of cancer deaths. In 2020, it was the third most incident cancer type and the second leading cause of cancer deaths worldwide (Sung et al., 2021). CRC is preventable through screening, and screening programs for CRC have been implemented in many countries. A large proportion of these are based on the Fecal Immunochemical Test (FIT) (Schreuders et al., 2015). This test measures the concentration of hemoglobin (Hb) in an individual's stool sample. An increased concentration may be caused by a precancerous lesion or a cancer. Participants with a concentration above a prespecified threshold for a positive result, commonly referred to as the cutoff, are referred for a colonoscopy, an endoscopic test with which the colon and rectum are directly observed by a specialized practitioner. Participants with a concentration below the cutoff are invited for a new FIT after a fixed time interval.

However, the FIT provides opportunities which currently remain unexploited. Grobbee et al. (2017) showed that measured FIT-concentrations, also below the cutoff, are strongly associated with the future risk of developing CRC. While screening intervals and cutoffs are equal across the population in current FIT-based programs, Grobbee et al. (2017) conclude that FIT-programs may be improved by implementing a screening policy with personalized intervals and cutoffs based on an individual's history of measured fecal Hb-concentrations.

Screening policies come with benefits, as they are likely to prevent CRC cases, and with harms such as overtreatment, for example when participants are treated for screen-detected lesions that would not have progressed to a cancer during their lifetime. These harms and benefits are measured in Quality-Adjusted Life Years (QALYs): one QALY represents one life year in perfect health. Screening policies also come with costs. Given their financial budget, policy makers aim to maximize the number of QALYs gained, and screening policies need to be developed that achieve precisely this. Implementing personalized screening policies may help to achieve this.

The amount of feasible personalized screening policies is endless, making it infeasible to evaluate the costs and health benefits of all of them in practice in randomized controlled trials. Instead, advanced simulation models such as those by Loeve et al. (1999) and Rutter and Savarino (2010) have been developed to evaluate screening policies. Still, the sheer amount of possible personalized screening options based on FIT-concentrations is so large, that it prohibits evaluating all options even by simulation. This underlines the need for optimization algorithms to design effective personalized screening policies without the need to evaluate all options.

Though algorithms have been developed to optimize personalized policies, none of them have the flexibility to incorporate detailed and computationally heavy simulation models, which is required for accurate evaluation of costs and benefits. Instead, strong assumptions are typically imposed to ensure computational tractability. Maillart et al. (2008), Ayer et al. (2012), Erenay et al. (2014) and Otten et al. (2017) use the framework of (Partially Observable) Markov Decision Processes (POMDPs) to develop personalized screening policies for a variety of cancer types, modeling the progression of the cancer by a Markov process. However, these Markov models assume, for example, that the transition rates between the different cancer states are independent. In reality, these transition rates are highly correlated within an individual. Consequently, POMDPs optimize their policies to a simpler model of the disease progression. Ahuja et al. (2017) adapt a method for POMDPs to incorporate such correlations in the cancer progression. However, they impose strong restrictions to the costs associated with screening and treatment, and don't allow for optimizing the costs and benefits as a bi-objective problem.

In this study, we present an algorithm that optimizes screening policies while incorporating MISCAN-Colon (Loeve et al., 1999). This is a detailed simulation model for CRC screening which is able to realistically evaluate the costs of and QALYs gained by a screening policy and that is commonly used to inform e.g., the United States Preventive Services Task Force on their CRC screening policy (Knudsen et al., 2020). We present a bi-objective evolutionary algorithm (EA), a heuristic algorithm which is frequently applied to difficult optimization problems. An EA is an ideal tool to combine with a computationally heavy evaluation procedure, in this case required to evaluate the costs and QALYs of a screening policy with MISCAN-Colon. Moreover, the EA is very well-suited to generate a frontier of screening policies with varying preference weights for costs and benefits, allowing policy makers to make a well-informed choice for a particular screening policy within their given budget. Finally, the EA is a flexible tool that is to some extent modular for the evaluation procedure. This means that the algorithm can be applied to inform screening programs for any disease, as long as there is a simulation tool to evaluate the costs and benefits of a screening policy, and the program uses a test with a quantitative test result. Examples include prostate cancer screening based on Prostate Specific Antigen (PSA), lung cancer screening based on smoking behavior and breast cancer screening based on nodule size, for all of which model consortia exist within the Cancer Intervention and Surveillance Modeling Network (CISNET) (Gulati et al., 2011; Alagoz et al., 2018; Criss et al., 2019).

In this paper, we present a proof-of-concept of our computational approach by (1) showing how our evolutionary algorithm can be combined with an established simulation model to optimize personalized screening policies, and (2) showing the potential of personalized screening in the case of CRC.

The remainder of this paper is structured as follows. In section 2, we discuss all aspects of the algorithm and how personalized screening policies are evaluated. In section 3, we present the outcomes of our experiments and compare them with screening policies from practice. Finally, in section 4 we discuss the outcomes of the experiments and the advantages and limitations of our algorithm.

2. METHODS

In this section, we introduce all aspects related to our evolutionary algorithm and the experiments we performed. First,

we give background on the microsimulation model MISCAN-Colon that is used to evaluate the costs and benefits of personalized screening policies obtained by the algorithm. Next, we introduce our mathematical framework for personalized screening policies. After that, we formalize the bi-objective optimization problem that we aim to solve in this study using our algorithm. Then, we present all details on the evolutionary algorithm. Finally, we introduce the experiments that we used to illustrate the performance of the algorithm.

2.1. MISCAN-Colon

The microsimulation model MISCAN-Colon was developed by the Department of Public Health within Erasmus University Medical Center, Rotterdam, The Netherlands. It is an established model, and has been used to inform the American Cancer Society (ACS) and the United States Preventive Services Task Force (USPSTF) guidelines (Knudsen et al., 2020). It has been validated on the results of three clinical trials on the effects of screening for colorectal cancer: the United Kingdom Flexible Sigmoidoscopy Screening (UKFSS) trial (DeYoreo et al., 2020); the Norwegian Colorectal Cancer Prevention (NORCCAP) trial (Buskermolen et al., 2018); and the Screening for Colon and Rectum (SCORE) trial (Gini et al., 2021).

The structure of the model, the underlying assumptions, and the calibration and validation studies have been described in detail by Loeve et al. (1999) and van Hees et al. (2014). In brief, the model simulates individual life histories from birth to death. At birth, all individuals are free of disease, but they may develop CRC during their lives. MISCAN-Colon assumes that all cancers develop from precancerous lesions, called adenomas, via the conventional adenoma-carcinoma pathway. Individuals may develop one or more adenomas over time. These lesions grow and may progress to preclinical CRC. Preclinical cancers are asymptomatic but may become symptomatic, resulting in clinical detection. Once a cancer becomes clinical, the person is treated, and a time to death is determined, depending on the stage of the cancer. The parameters of the natural history of CRC were calibrated to high-quality data sources, such as autopsy studies on age-specific adenoma prevalence and multiplicity (Meester et al., 2018) and age-, stage-, and location-specific CRC incidence data from the Surveillance, Epidemiology and End Results (SEER) program from the period before screening was common practice (1975–1979) (SEER, 2021).

The model also has an optional screening component. When activated, the simulated individuals undergo screening according to a specified screening policy. Some lifetimes are altered because some cancers are prevented by removal of the precedent adenomas, or are detected at an early stage, leading to more favorable survival. The effect of screening depends on the implemented policy and the test characteristics such as the sensitivity and specificity and the reach of endoscopic tests. Endoscopic tests also have a risk of complications. The characteristics of the screening tests in MISCAN-Colon are based on various studies to assess the diagnostic performance of FIT and colonoscopy (Knudsen et al., 2016).

Screening policies are associated with monetary costs and benefits in terms of QALYs, related to the total number of screening tests and the life years spent on cancer treatment in a simulated population. After simulation, the model aggregates these quantities to calculate the policy's costs and benefits. The costs and benefits used in this study are listed in Gini et al. (2017).

Up to now, MISCAN-Colon has not been used before to evaluate personalized screening policies based on FIT-history. FITs were modeled as binary tests that return either a positive or negative test result based on sensitivity and specificity. For our study, the model was extended with a prototype module describing individuals' fecal occult blood loss over time, such that FIT-concentrations were returned. A model was developed with a linear mixed-effects model (GLMM) structure. Its parameter values were estimated using population-based data on measured FIT-concentrations and corresponding outcomes observed in the Dutch national colorectal cancer screening program (Toes-Zoutendijk et al., 2017). This module is a prototype and still needs further calibration before informing actual policy changes. However, the quality of this module is not relevant for the purpose of this study which is to provide a proof-of-concept of the presented computational technique.

An overview of the model assumptions for the natural history, test characteristics and the module for FIT-concentrations is presented in Supplementary Section 1 of the **Supplementary Material**.

2.2. Personalized Screening Policies

In this section, we provide the mathematical framework for personalized screening policies. In short, an individual is represented by a pair (r, τ) that contains its perceived risk of CRC based on its FIT-history r and its age at the most recent FIT τ. The two-dimensional space of all possible pairs is called the *belief space*. Each individual is represented by a point in this space. A screening policy prescribes an action for each point in the belief space. There are two types of actions: either a participant is referred to a hospital for a follow-up colonoscopy, denoted by *COL*, or an interval of I years until the next FIT is prescribed, denoted by FIT_I. In fact, a personalized policy is a mapping that partitions the belief space and relates each part to an action. An example is given in **Figure 1A**, in which screening intervals of 1, 2 and 3 years are prescribed. The remainder of this section provides a more extensive formalization of the framework of personalized screening policies.

First, the framework requires a discrete set of screen-eligible age groups $\mathcal{T}$. In this study, individuals were assumed eligible for screening between ages 40 and 85. This range was split in age groups of 5 years and we assumed that the policy is the same for each age group, i.e., two individuals aged 40 and 44 with equal perceived risk are prescribed the same action. Age groups are represented by their lowest age and in our study, the set of age groups was $\mathcal{T} := \{40, 45, \ldots, 80\}$.

Second, the framework requires a measure for perceived risk of CRC. In this study, perceived risk was estimated by the average of an individual's k most recently measured FIT-concentrations. We used $k = 1$ as a base case and $k = 2, 3$ for sensitivity analyses. The average was mapped linearly to a value in the range $[0, 1]$ where risk values of 0 corresponded to a negligible risk and 1 to a very high risk. This way, more advanced risk estimators

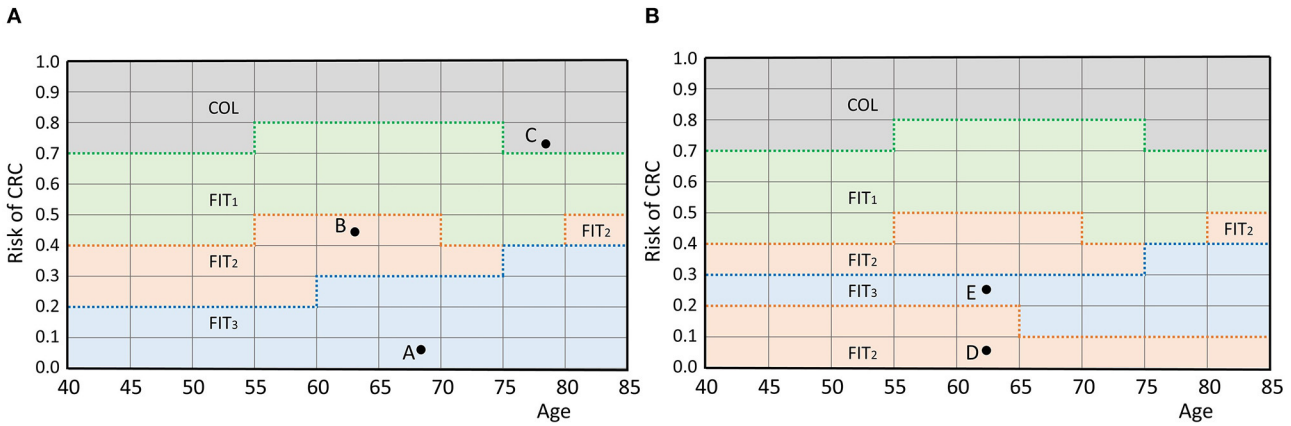


FIGURE 1 | (A) An example policy. Individuals are represented by a pair (age, perceived risk) which is a point in the belief space. For example, the participant represented by A is aged 68 and has a risk of 0.075. A policy defines an action for each individual by partitioning the belief space and relating each part to an action, as shown by the colors. The individuals represented by points A and B are prescribed actions FIT_3 and FIT_2 (a screening interval of 3 and 2 years), respectively. Individual C is referred to a hospital for a colonoscopy. The discretization of $\mathcal{R}$ and $\mathcal{T}$ causes the grid structure. **(B)** An infeasible policy. Participant E is assigned a longer interval than participant D while they are of the same age and the risk of E is higher. Therefore, this policy does not comply with the test order assumption.

can easily be incorporated in the future. Since most countries use a cutoff between 15 and 80 μg/g (Schreuders et al., 2015), we assumed that average FIT-concentrations above 100 μg/g correspond with a perceived risk of 1. In our method, individuals with a FIT-concentration above 100 μg/g were always referred for a colonoscopy. Formalizing the above, the perceived risk of CRC R^k after the participant's n^{th} FIT was calculated as

$$R^k := \frac{1}{100k} \sum_{i=0}^{k} C_{n-i},$$

with C_j the measured concentration at the participant's j^{th} FIT. Similar to the age groups, we discretized the interval [0,1] in parts of length 0.1 and assumed that the action is the same within each part for a given age group, i.e., two individuals with risk 0.11 and 0.19 of equal age were prescribed the same action. This restricted the number of feasible cutoffs. The set of feasible cutoffs $\mathcal{R}$ was $\{0, 0.1, \ldots, 1\}$. Note that the discrete nature of $\mathcal{T}$ and $\mathcal{R}$ causes the grid structure in **Figure 1A**. Finer discretization increases the number of potential personalized screening policies, but also increases the size of the search space of the algorithm.

Third, the framework needs a set of actions $\mathcal{A}$. In our study, we used two types of actions. The first, denoted by *COL*, was equivalent to a positive FIT and referred an individual for a colonoscopy in a hospital. After a positive colonoscopy result, the individual left the screening program and was referred to a surveillance program instead. After a negative colonoscopy result, the individual re-entered the screening program and obtained a new FIT after a fixed 5-year interval. The second action type corresponded with a negative FIT and prescribed a screening interval I. Such actions were denoted by FIT_I. The set of feasible intervals was denoted by $\mathcal{I}$. The resulting action set $\mathcal{A}$ was

$$\mathcal{A} := \{COL\} \cup \{FIT_I | I \in \mathcal{I}\}.$$

Considering larger action sets allows for more potential screening policies, but also increases the size of the algorithm's search space.

The space $\mathbb{B} := \mathcal{R} \times \mathcal{T}$ is called the *belief space*. The current status of a participant is represented by a point in this space. A screening policy $\pi : \mathbb{B} \to \mathcal{A}$ partitions the belief space and maps each part to an action in the action set (see **Figure 1A**), defining an action for each participant.

The framework assumes that the actions have a test burden and that the order of the actions in the belief space is fixed with respect to this test burden. In our case, colonoscopies, for which participants are referred to a hospital, have a relatively high test burden compared to FIT which is done at home. Short FIT-intervals were also assumed to have a higher burden than longer intervals. Only screening policies that adhere to this ordering by test burden are considered. That is, a participant is only assigned a test with a higher burden than another participant of the same age, if also the perceived risk is higher. **Figure 1B** shows an example of a screening policy that does not comply with the *test order assumption*. We consider such a policy infeasible in this framework.

As the ordering of the actions is fixed per age group, screening policies can also be characterized by the bounds of their partitions. The upper bound of the parts of the belief space that correspond to an action are considered a function in the belief space. In our study, this concerned the actions FIT_I with corresponding policy bounds $\beta_I : \mathcal{T} \to \mathcal{R}$. In **Figure 1A**, these functions are represented by the bold, dotted lines. A screening policy is characterized by the set of its policy bounds

$$\pi = \{\beta_I\}_{I \in \mathcal{I}}.$$

Note that the characterization only included the policy bounds of the screening intervals $I \in \mathcal{I}$, because the upper bound of the part corresponding with the action *COL* was not relevant. This

characterization of personalized screening policies is used in the remainder of this paper.

Policies that are obtained by a combination of two other policies are also considered. By prescribing policy π to a fraction $\lambda \in (0,1)$ of the population and prescribing policy σ to the remaining fraction $(1-\lambda)$, a new policy ρ is generated.

2.3. Optimization Problem

Next, we introduce the optimization problem solved in this paper. In particular we present a multi-objective (specifically bi-objective) optimization problem.

A policy π has associated costs and QALYs, denoted as $C(\pi)$ and $Q(\pi)$, respectively, and measured per 1,000 individuals, as is common. We define $\boldsymbol{o}(\pi) := \begin{bmatrix} C(\pi) \\ Q(\pi) \end{bmatrix}$ as the vector containing both objectives of π.

The bi-objective optimization problem is to find policies minimizing the costs and maximizing the QALYs gained. A single policy optimizing both objectives is unlikely to exist as screening policies with an increased number of QALYs gained generally come with higher costs. Therefore, we aim to find a set of policies that contains those with maximal benefits for given costs. Given this set, policy makers can choose policies based on their budget constraints or on what they find a suitable balance between the two criteria.

In a multi-objective setting, the concept of Pareto dominance is used to compare policies. A policy π *dominates* another policy σ if π is a better choice than σ, i.e., if (1) the costs and QALYs of π are at least as good as those of σ:

$$Q(\pi) \geq Q(\sigma) \text{ and } C(\pi) \leq C(\sigma),$$

and (2) at least one of the objectives is better:

$$Q(\pi) > Q(\sigma) \text{ or } C(\pi) < C(\sigma).$$

Figure 2 shows the costs and QALYs of several example policies. Here, policy B dominates E because its costs are lower and its QALYs are higher. B does not dominate D. A policy that is not dominated by any other policy is called *Pareto optimal*. The set of all Pareto optimal policies is referred to as the *Pareto frontier*. The multi-objective optimization problem is to find the Pareto frontier. The Pareto frontier potentially includes an infinite number of policies, and is computationally difficult to identify precisely. Therefore, the algorithm aims to find a set of policies that best approximates the Pareto frontier.

Next, we explain how an approximation of the Pareto frontier is represented using the approximation set as introduced in Zitzler et al. (2003). This set makes use of combinations of policies, i.e., prescribing policy π to a fraction $\lambda \in (0,1)$ of the population and prescribing policy σ to the remaining fraction $(1-\lambda)$ which results in a new policy ρ. Observe that the objective values of ρ are convex combinations of the objective values of π and σ in the conventional sense:

$$\boldsymbol{o}(\rho) = \lambda \boldsymbol{o}(\pi) + (1-\lambda)\boldsymbol{o}(\sigma).$$

By varying λ, an infinite number of new policies can be generated using only two policies.

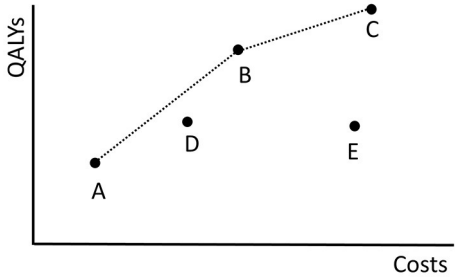


FIGURE 2 | The costs of and QALYs gained by the fictive screening policies in $\Pi = \{A, B, C, D, E\}$ are plotted. The dotted black line shows the objective values of the policies in the approximation set $\psi(\Pi)$. The minimal representation of $\psi(\Pi)$ is $\{A, B, C\}$.

We use the above observation to create an approximation set of the Pareto frontier of the following form. An approximation set is represented using a finite set of policies Π. This approximation set contains all non-dominated policies among Π and all their non-dominated combinations, and is denoted by $\psi(\Pi)$. This way, (if $|\Pi| \geq 2$) the approximation $\psi(\Pi)$ consists of an infinite set of policies, but can be represented using a, typically small, finite set of policies. In our computations, but also when presenting the results in this paper, it is beneficial to consider a minimal representation of $\psi(\Pi)$, which is a smallest subset Π' of Π such that $\psi(\Pi') = \psi(\Pi)$.

As an example, the dotted black line in **Figure 2** shows the approximation set $\psi(\Pi)$ represented by $\Pi = \{A, B, C, D, E\}$. The same approximation set can also be represented by policies $\Pi' = \{A, B, C\}$ because D is dominated by a combination of A and B and E is dominated by B.

2.4. Evolutionary Algorithm

In this section, we describe the evolutionary algorithm (EA) which we developed to identify approximation sets of the Pareto frontier. EAs are based on the principle of *survival of the fittest* (Holland, 1975).

In general, the algorithm keeps track of two sets of policies. Firstly, it maintains a *population* of screening policies. This set evolves over time, i.e., it changes at every iteration of the EA. Secondly, it maintains a *memory* which is a set of policies that is a minimal representation of the best approximation set found so far. This set is updated every time that a policy appears in the population which is non-dominated by any found policy. This new policy is then added to the memory, and others are removed if they are dominated. Therefore, the population can be thought of as the current generation, while the memory simply contains the best policies observed over all generations. Although we are interested in the approximation set represented by the final memory as the final solution to our optimization problem, the population does not necessarily have to be a non-dominated set of policies. In fact, for diversification purposes it can be beneficial to allow inferior policies in the population.

As an example, if policies $A, \ldots, E$ in **Figure 2** are the policies found by the algorithm, policies $A, \ldots, D$ form the memory as

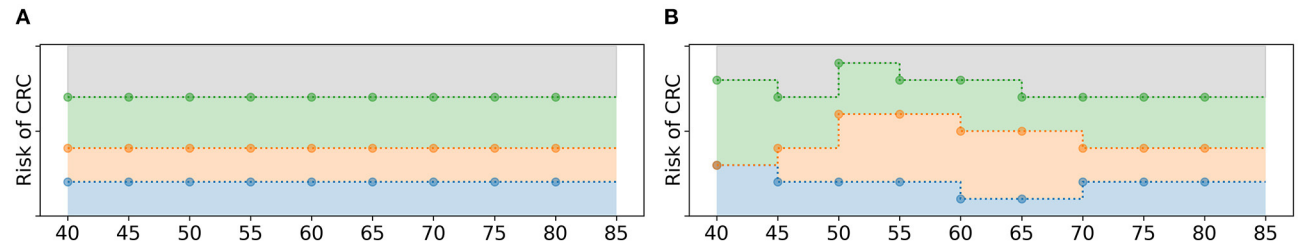


FIGURE 3 | Schematic overview of the initialization of a policy. **(A)** First, the policy bounds are assigned a constant value over all age groups, adhering to the test order assumption. **(B)** After that, random mutations are applied.

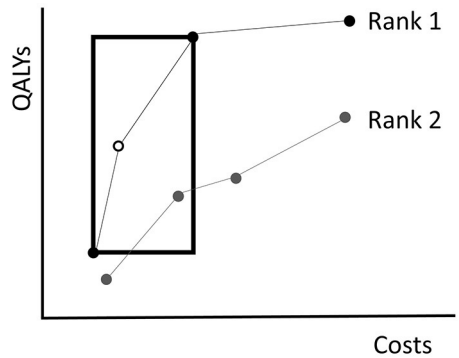


FIGURE 4 | An illustration of the NSGA-II algorithm. The dots represent the costs of and QALYs gained by several fictive policies. Policies with equal rank are connected by the lines. Non-dominated policies obtain rank 1. Those that are dominated by policies with rank 1 only, obtain rank 2, etc. The crowding distance of the white policy equals the circumference of the rectangle drawn around it. The rectangle touches the next less costly and next more expensive policies with equal rank.

policy E is dominated by B. Policy D is also part of the memory because it is not dominated by a found policy.

The EA starts with an *initial population* that consists of a predefined number of screening policies. It evaluates the *fitness*, or quality, of these policies in terms of the objectives. Then, it *selects* half of the policies to stay in the population and discards the other half. This is a semi-random selection procedure where solutions of higher fitness are more likely to be selected. The selected policies are paired up randomly to form pairs of parents. Together, these parents generate two child policies by exchanging some of their features, called *cross-over*. Some of the child policies undergo random *mutations* in which their features are changed randomly. Finally, the algorithm adds the children to the population, which results in a new population, and updates the memory such that it contains the best policies observed until then. It repeats the cycle of fitness evaluation, selection, cross-over and mutation until some stopping criterion is met.

In the remainder of this section we provide a more detailed description of the key elements of the EA and its interaction with MISCAN-Colon.

2.4.1. Initialization

A screening policy is initialized in two steps as illustrated in **Figure 3**. First, each policy bound $\beta_I(\tau)$ is assigned a constant value for all age groups $\tau \in \mathcal{T}$. For that, $|\mathcal{I}|$ random values are uniformly drawn from $\mathcal{R}$ and assigned to the policy bounds, adhering to the test order assumption. That is, the smallest value drawn from $\mathcal{R}$ is assigned to the policy bound that relates to the action with the lowest test burden, the second smallest value to the action with the second lowest test burden, etc. Then, the mutation operator as described in section 2.4.5 is applied such that the policy bounds are not necessarily constant over the age groups anymore. The algorithm repeats these two steps N_{pop} times to obtain an initial population of policies, where N_{pop} denotes the number of screening policies in the population.

2.4.2. Fitness Evaluation

The algorithm bases the fitness of a policy in the population on its costs and QALYs as simulated by MISCAN-Colon. In MISCAN-Colon, both objectives were discounted by 3% annually from the age of 40 and were calculated relative to a situation without screening. Simulations used one million individuals. Common seeds ensured that the results of different simulation runs were comparable.

The EA uses the Non-Dominated Sorting Genetic Algorithm-II (NSGA-II) introduced by Deb et al. (2002) to evaluate fitness. NSGA-II summarizes fitness of policies with two quantities: the *rank* and *crowding distance*. Given a population of policies P, the rank of a policy represents to what extent it is dominated by other policies in P (excluding combinations of policies). Non-dominated policies in P obtain rank 1. Then these policies are excluded and the non-dominated policies of the remainder are assigned rank 2. This is repeated until every policy is ranked (**Figure 4**). Consequently, the solution quality increases for decreasing rank.

It is likely that multiple policies in the population obtain an equal rank. To break a tie in the selection procedure, NSGA-II evaluates for each policy a crowding distance. The crowding distance is a statistic that reflects the level of isolation with respect to other policies with equal rank. For a policy π, the crowding distance is the circumference of the rectangle that touches the next less costly and next more expensive policies with the same rank as π, see **Figure 4** for an example. Note that the crowding distance of the cheapest and most expensive policies in a frontier are considered to be infinite. In case two policies have equal rank, the algorithm prefers the one with a higher crowding distance. The idea behind the crowding distance is to get a good spread of different screening policies, i.e., expensive as well as cheap

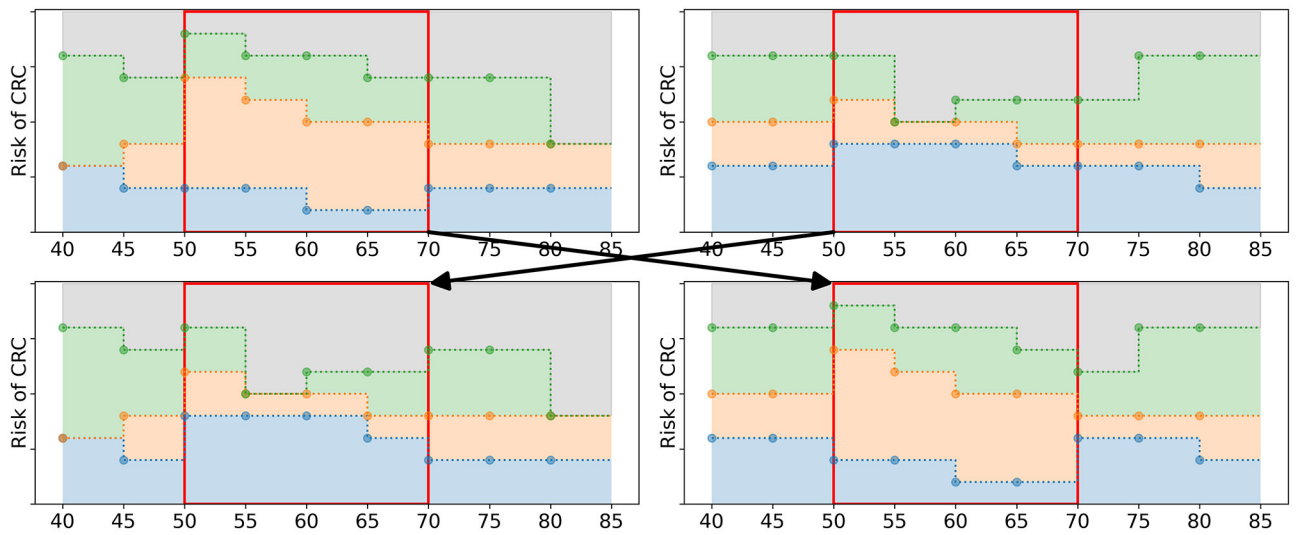


FIGURE 5 | Schematic overview of the cross-over operator. The two upper figures represent the parent policies, a random pair of policies in M_g. The lower figures represent two offspring policies that are added to O_g. The two age groups 50 and 65 were randomly selected. The policy bounds in between these age groups (marked by the red box) are exchanged in the offspring.

policies. This may help in achieving a high quality approximation of the complete Pareto frontier.

In our particular bi-objective case, the time complexity of the NSGA-II algorithm is $O(N_{pop}^2)$ (Deb et al., 2002).

2.4.3. Selection Operator

Once the rank and the crowding distance of the policies in the population are evaluated, the EA selects which policies are maintained in the population and which are discarded. The maintained policies form the *mating pool*. In iteration g, the pool is denoted by M_g. During the selection procedure, exactly $N_{sel} := N_{pop}/2$ policies are selected and added to M_g. The selection operator consists of two phases.

First, the mating pool is (partially) filled by an elitist selection procedure. Given the current population P_g and the approximation set $\psi(P_g)$, the EA adds the solutions in the minimal representation of $\psi(P_g)$ to the mating pool, i.e., it adds the policies in the population that are not dominated by (combinations of) other policies in the population. This ensures that the best policies are selected. Note that this is a subset of the policies with rank 1. Tests with our benchmark have shown that adding the complete set of policies with rank 1 in this phase leads to poorer algorithm performance. If more than N_{sel} policies are selected in this first, elitist phase, the algorithm randomly discards policies until N_{sel} policies remain.

In the second phase, the remainder of the mating pool is filled by tournament selection: two policies are randomly sampled from the population and the fittest of the two policies in terms of rank and crowding distance is added to the mating pool. This is repeated until the mating pool is filled with N_{sel} policies. Note that this procedure may lead to duplicates in the mating pool. Policies may be selected once in both phases and/or multiple times in the second phase.

2.4.4. Cross-Over Operator

Having filled the mating pool M_g, the algorithm applies 2-point cross-over (Whitley, 1994) to generate offspring. The policies in M_g are paired up randomly. For each of the pairs, two age groups $\tau_1, \tau_2 \in \mathcal{T}$ are randomly selected. The policy bounds in the interval $[\tau_1, \tau_2]$ are exchanged, see **Figure 5** for an example. This results in two new offspring policies which are added to O_g, the set of offspring obtained in iteration g. After all pairs of parents have generated offspring, O_g has a size of $N_{pop}/2$.

2.4.5. Mutation Operator

The offspring policies in O_g are subject to random mutations with probability p_M. If the EA selects a policy to undergo mutation, the following steps are taken. First, a fraction p_e of the age groups in $\mathcal{T}$ is randomly selected. For these age groups, the values of all policy bounds $\{\beta_I\}_{I \in \mathcal{I}}$ are mutated: they are replaced by random values from $\mathcal{R}$. However, these values are not sampled from $\mathcal{R}$, instead they are sampled from a subset of $\mathcal{R}$. For each selected age group, a value $\tilde{r} \in \mathcal{R}$ is sampled. This value is an upper or a lower bound with 50% probability. If it is an upper bound, $|\mathcal{I}|$ random values are drawn from the values in $\mathcal{R}$ smaller or equal to $\tilde{r}$. If it is a lower bound, they are drawn from the values in $\mathcal{R}$ larger or equal to $\tilde{r}$. These new values are assigned as policy bounds, adhering to the test order assumption, see **Figure 6** for an example.

The reason to sample the new values from a subset of $\mathcal{R}$ is that this is more likely to result in a larger variety of policies. For example, the policy bound related to the largest screening interval always obtains the smallest of the $|\mathcal{I}|$ new values. If these values are drawn from the complete set $\mathcal{R}$, it is unlikely that a value close to 1 is assigned to this bound. This is more likely to occur when sampling from a subinterval of $\mathcal{R}$.

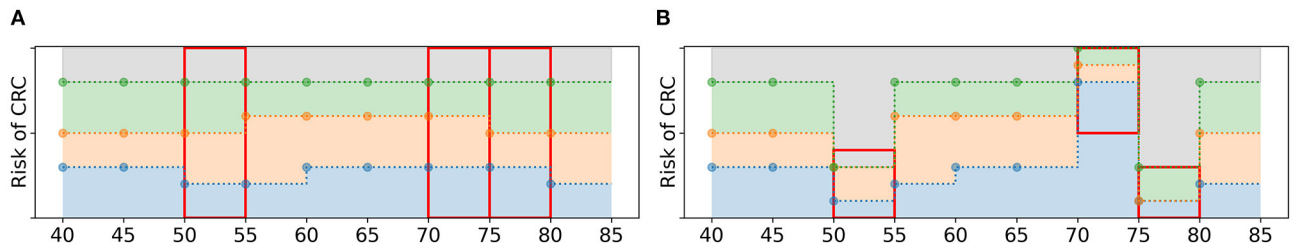


FIGURE 6 | Schematic overview of the mutation of a policy. **(A)** A policy before mutation. If $p_e = 3/9$, the algorithm is likely to select three age groups, marked by the three red boxes. **(B)** For each age group, new policy bounds are sampled from a subset of $\mathcal{R}$ (red boxes). For age groups 50 and 75, $\tilde{r}$ is 0.4 and 0.3, resp., and it is an upper bound. Therefore, the new values of the policy bounds must be smaller or equal to 0.4 and 0.3, resp. For age group 70, $\tilde{r} = 0.5$ is a lower bound and all new values are larger or equal to 0.5.

2.4.6. Updating Procedures and Stopping Condition

After applying all operators, the algorithm obtains (1) a mating pool M_g that contains the selected policies from the current population P_g, and (2) a set of newly generated offspring O_g. Both sets have size $N_{pop}/2$. The algorithm merges these sets to obtain the population for the next iteration, i.e., $P_{g+1} = M_g \cup O_g$.

Additionally, it updates its memory with the best found policies. It adds all newly found policies that are not dominated by the policies in the current memory, and removes all policies that are dominated by the newly added policies.

The algorithm repeats the procedures for selection, fitness, cross-over, mutation and updating until no new solutions are added to the memory for $N_{stop} = 30$ consecutive iterations. The approximation set represented by the memory at the final iteration is considered the best approximation of the Pareto frontier and is the final solution to our problem.

2.5. Experiments and Implementation

We demonstrate the performance of the algorithm with three different experiments. First, we evaluated how well the algorithm approximated a Pareto frontier, i.e., the optimal solution to the multi-objective optimization problem, using a benchmark problem. We considered an instance of the problem with a relatively small number of feasible policies, which enabled us to enumerate all feasible policies, evaluate their costs and QALYs and identify the Pareto frontier. All policies were simulated with 2 million individuals using common random numbers to ensure that each policy was evaluated for exactly the same population. Based on this benchmark, we also identified the best values for the parameters N_{pop}, p_M and p_e.

The benchmark problem size was reduced by restricting the assumed screen eligibility to ages 55 to 75, resulting in the age groups $\mathcal{T} = \{55, 60, 65, 70\}$, and restricting the set of feasible cutoffs to $\mathcal{R} = \{0, 0.125, 0.25, 0.375, 0.5\}$. We used R^1 to estimate perceived risk. As shown in Supplementary Section 2 of the **Supplementary Material**, this combination of parameters gave approximately 1.5 million feasible policies.

To quantify how well the Pareto frontier was approached by an approximation set, we used the relative difference between the hypervolume (HV) of both sets. The HV is a quality indicator introduced by Zitzler and Thiele (1998) and is very common in multi-objective optimization (Riquelme et al., 2015).

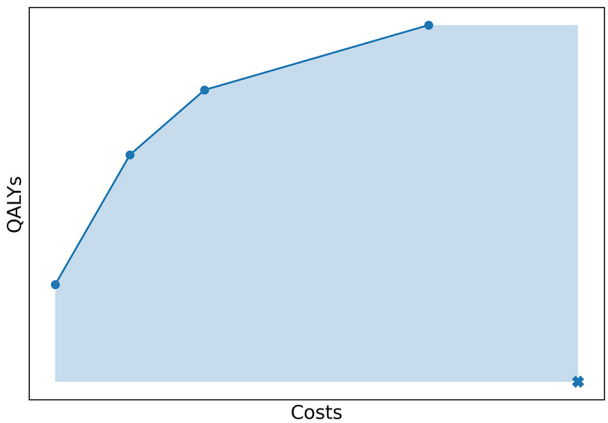


FIGURE 7 | The costs and QALYs of several policies that form a minimal representation of an approximation set. The hypervolume of this approximation set is equal to the area of the objective space dominated by the approximation set, bounded by a reference point. This point is marked with a cross in the figure.

In our study, the hypervolume of an approximation set was defined as the area of the objective space dominated by the approximation set, bounded in some sense by a reference point as illustrated in **Figure 7**. The reference point was chosen as (costs, QALYs) = (4,000,000; 0). In Experiment 1, we evaluated the HV of both the approximation set represented by the Pareto frontier and the approximation set obtained by the algorithm. The relative difference between the two quantified the optimality gap, i.e., how well the approximation set approaches the Pareto frontier.

Next, we used two larger problem instances to test the algorithm. Experiment 2 used the original settings for $\mathcal{T}$ and $\mathcal{R}$ and used the action set $\mathcal{A} = \{COL, FIT_1, FIT_2, FIT_3\}$ such that both the cutoff for FIT-positivity and screening intervals were optimized. In Experiment 3, we considered a simplified situation in which $\mathcal{A} = \{COL, FIT_2\}$. It effectively means that we used a fixed screening interval of 2 years and only optimized the cutoff per age group. This is an improvement already compared to current practice in which the cutoff is fixed for all ages. The size

of the search space was much smaller compared to Experiment 2 (see **Supplementary Material**).

For both experiments, it was computationally impossible to evaluate all feasible policies and to find the exact Pareto frontier. To evaluate the obtained approximation sets, we compared them in terms of costs and QALYs with policies recently evaluated for the United States Preventive Services Task Force (USPSTF) by Knudsen et al. (2020) that include FIT and/or colonoscopies. For a fair comparison, we only used reference policies that start screening no later than age 45, because the policies generated by the algorithm all start at age 40 due to our chosen parameter settings. An overview of the reference policies is shown in Supplementary Section 3 of the **Supplementary Material**. The reference policies and those in the memory of the algorithm were (re-)evaluated with MISCAN-Colon using 2.5 million individuals and with a different random number stream than used in the algorithm. This prevented a biased comparison, since the policies of the algorithm may have been optimized to the random number stream used for simulations in the EA.

As is common in health economics, we made use of a statistic, the *incremental cost-effectiveness ratio* (ICER) (Sanders et al., 2016), to identify a single policy in an approximation set which is cost-effective, for comparative purposes. We evaluated the ICER for the policies in the finite set that is a minimal representation of the approximation set. The ICER of policy π is defined as the extra costs per extra QALY gained when opting for policy π instead of the next less costly policy in the minimal representation, i.e., it is defined as the ratio between the difference in costs and the difference in QALYs gained between the two. Due to our definition of an approximation set, the ICER of a policy increases for increasing costs. The cost-effective policy is defined as the policy that has maximum benefits for which the ICER is still below a predetermined threshold, often called the willingness-to-pay threshold. In this study, we used a threshold of $100,000 per QALY gained to determine the cost-effective strategy.

The running time of the algorithm strongly depends on the implementation and computational resources. In our experiments, the algorithm was implemented using the Python DEAP evolutionary computation framework (Fortin et al., 2012) and implemented as a high-performance computing (HPC) workflow using the EMEWS framework (Ozik et al., 2016). The first, second and majority of the third experiment were run on Bebop, an HPC cluster managed by the Laboratory Computing Resource Center at Argonne National Laboratory. Bebop has 1,024 nodes comprised of 672 Intel Broadwell processors with 36 cores per node and 128 GB of RAM and 372 Intel Knights Landing processors with 64 cores per node and 96 GB of RAM.

3. RESULTS

In this section, we present the results of the three experiments introduced in section 2.5. All presented costs and QALYs are relative to a situation without screening for CRC. Also, they were discounted by 3% annually from age 40, as is common in cost-effectiveness analyses.

3.1. Experiment 1: Benchmark

Figure 8 shows the costs and QALYs of all feasible policies in the benchmark problem, evaluated in 10 phases on Bebop, 9 of which used 1,792 cores each and 1 which used 2,016 cores. It was completed in 97.07 h, resulting in 177,528.31 core hours in total.

Experiments were done with varying values for N_{pop}, p_M, and p_e. After convergence, the hypervolume (HV) of the obtained approximation set was highest for the values $(N_{pop}, p_M, p_e) = (400, 0.3, 0.6)$. This approximation set, obtained after 499 iterations of the algorithm, is included in **Figure 8**. The three selected parameters values are used in the remainder of our study.

We observe that nearly all feasible policies are dominated by the approximation set, suggesting it is a good approximation of the Pareto frontier. This is further confirmed by the hypervolume. The HV of the approximation set and Pareto frontier (PF) equal 108,116,896 and 108,124,226, respectively, effectively resulting in an optimality gap of 0.007%.

The PF contains 12 policies, the minimal representation of the approximation set contains 11. Further analysis showed that the 11 policies representing the approximation set are all part of the representation of the PF: the approximation set misses only one of the policies on the PF, which explains the optimality gap. The missing policy is marked in **Figure 8**.

3.2. Experiment 2: Optimizing Cutoffs and Screening Intervals

In the second experiment, using R^1 to estimate perceived risk, the algorithm took 1263 iterations until convergence. This was performed in 5 phases on Bebop. Each phase of the experiment was run on 432 cores, enabling 430 individual policies to be evaluated in parallel with the remaining two processors being used for workflow management. The total number of 1,263 iterations was completed in 101.7 h for a total compute time of 43,934.4 core hours, four times faster than the enumeration in Experiment 1 despite the factor 10^{16} increase in search space (see **Supplementary Material**). The evolutionary operators consumed 0.11% of the total computation time, the remainder was used by MISCAN-Colon.

Incorporating extra FIT-concentrations in the perceived risk value did not affect the performance and the outcomes of the algorithm. Experiments with perceived risk estimators R^2 and R^3 resulted in similar computation times and policies with similar costs, QALYs and patterns. In the remainder of this section, we only discuss the outcomes using R^1.

Figure 9 shows the total number of policies added to the memory in each iteration, and how many of these policies were added to the minimal representation of its approximation set, i.e., the number of new policies that were not dominated by any *combination* of other policies in the memory. We observe that the latter group is a minority. Especially in the final 600 iterations, only 9 of such policies were found.

Figure 10 shows the costs and QALYs of the best approximation set of the PF obtained by the algorithm and of all reference policies. The minimal representation of the approximation set contains twelve personalized policies, and dominates all reference policies. For similar costs, the QALYs of

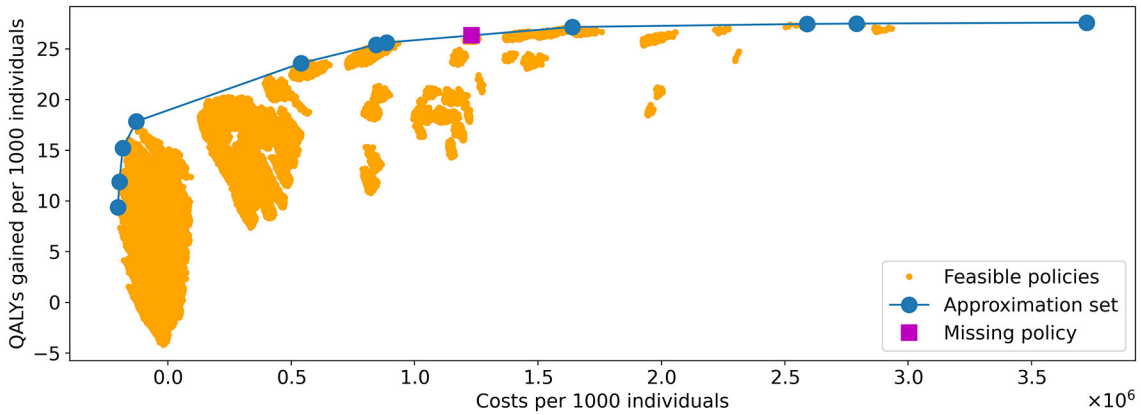


FIGURE 8 | Visualization of all feasible policies in the problem instance of Experiment 1. The yellow dots represent the costs and QALYs of the feasible policies. The blue line shows the approximation set obtained by the algorithm. It is represented by the 11 policies indicated by the blue dots. The policy indicated by the purple square is the only strategy on the PF that was not in the approximation set found by the algorithm.

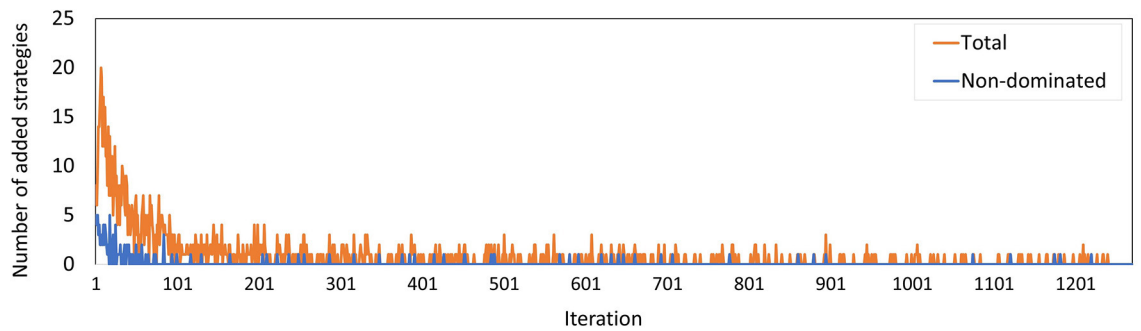


FIGURE 9 | The number of policies added to the memory in each iteration for Experiment 2. The blue line counts all added policies, the orange line only those that are not dominated by (combinations of) other policies in the current memory. The latter are part of the minimal representation of the memory's approximation set.

the obtained screening policies increased up to 14% compared to the reference policies. This shows that the algorithm succeeded in finding personalized screening policies that are more effective than the uniform reference policies as evaluated using MISCAN-Colon.

To characterize the obtained approximation set, **Figure 11** shows the cost-effective personalized policy in more detail (policy 6, marked blue-red in **Figure 10**), as well as two reference policies with comparable costs and QALYs (marked green-red in **Figure 10**). The reference policies initiate screening at age 45. Policy 6 prescribes screening before 45, but limits colonoscopy referrals by prescribing a high FIT-cutoff of 90 μg/g. The reference policies both stop screening at age 75. Policy 6 prescribes high cutoffs and long intervals from age 70. Since the algorithm is forced to design screening policies that start at age 40 and stop at age 85, we suspect that it tries to reduce the screening intensity by prescribing long intervals and high cutoffs for younger/older age ranges. Interestingly, the FIT-cutoffs at age ranges 55 and 65 in policy 6 are 0 μg/g, effectively resulting in a guaranteed referral for a colonoscopy regardless of the measured FIT-concentration. After such a colonoscopy, provided it was negative, screening is first halted for 5 years by design. We see that screening is then offered with higher cutoffs for another 5 years. Effectively, the colonoscopies are applied with a 10-year interval for most participants between these ages, in line with current USPSTF recommendations for colonoscopy-based screening and policy C3.

Figure 12 displays all policies that represent the blue approximation set in **Figure 10** to observe the effect of decreasing or increasing the costs compared to policy 6. All policies offer intermittent colonoscopy and FIT-screening by prescribing at least one guaranteed colonoscopy and prescribing FIT-screening with higher cutoffs after a guaranteed colonoscopy with a negative result. The cheaper policies focus on FIT-screening during the ages 50 through 65. They apply higher cutoffs and longer screening intervals for other ages, limiting the screening intensity for those age ranges. This is a consequence of the lower risk of CRC for younger age ranges in general and the shorter life expectancy for older age ranges, effectively resulting in less life years to gain from screening. More expensive policies focus relatively more on colonoscopy screening (FIT-cutoffs of 0 μg/g) and decrease the cutoffs and the intervals first for those

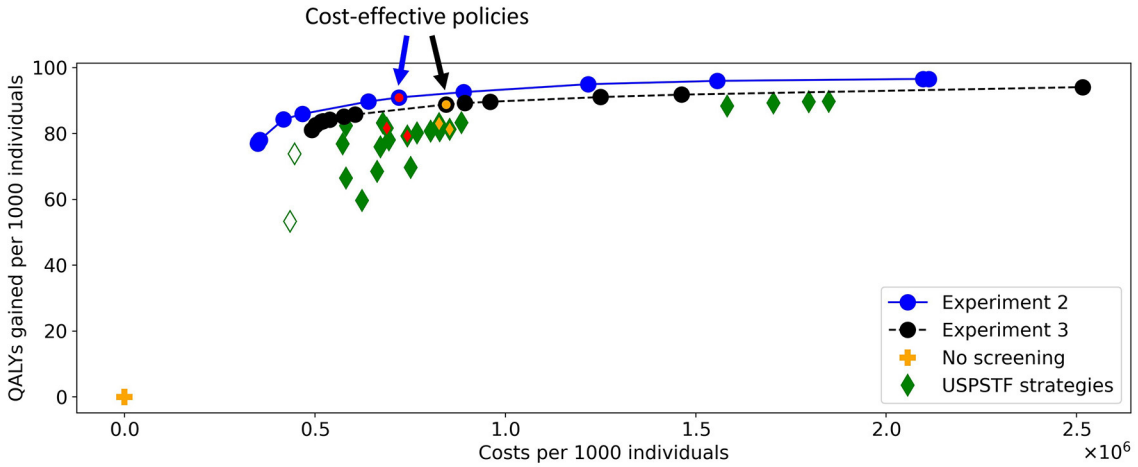


FIGURE 10 | The costs and QALYs of the reference policies and of the best approximation sets generated in Experiments 2 and 3. The blue line shows the best approximation set obtained in Experiment 2, the black line shows that of Experiment 3. The blue-red and black-orange policies in the approximation sets are cost-effective. The plus represents a situation without screening, the diamonds represent the policies evaluated for the USPSTF with FIT and/or colonoscopies that start at age 45. The two white diamonds are the two reference policies that are not dominated in Experiment 3. The four green-red/green-orange diamonds are referred to in **Figures 11**, **14**.

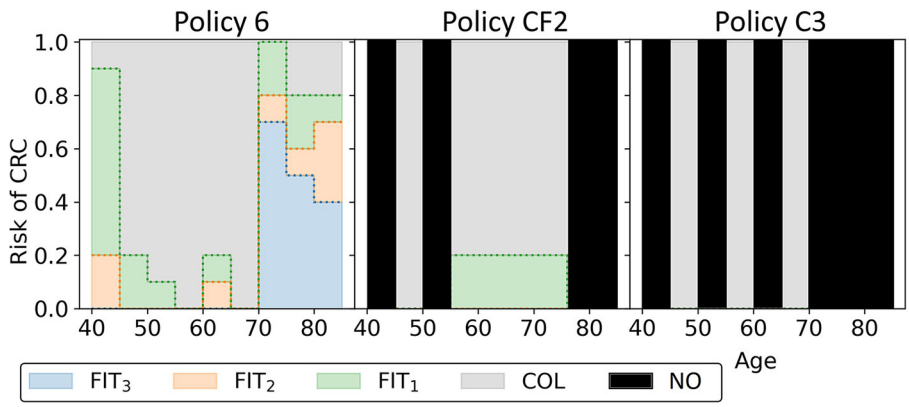


FIGURE 11 | The three blue-red/green-red policies in **Figure 10** are shown. Policy 6 is the cost-effective policy within the willingness-to-pay threshold in the approximation set for Experiment 2. Policies CF2 and C3 are the closest reference policies in terms of costs and QALYs (these policies are listed in **Supplementary Table 3**). For age groups with a black bar, reference policies do not offer screening.

aged 40 and then for the 70+ age ranges. The most expensive policies prescribe multiple guaranteed colonoscopies, similar to the colonoscopy-based policies evaluated for the USPSTF.

3.3. Experiment 3: Optimizing Cutoffs

Experiment 3 has a smaller number of feasible policies compared to Experiment 2 because the action space was smaller. Nonetheless, the algorithm converged after 2,111 iterations, more than in Experiment 2. The third experiment was run in two phases. The first 505 iterations were run on a virtual machine managed by Erasmus Medical Center, the remaining 1,606 iterations on Bebop. The part run on Bebop was performed on 288 cores, enabling 286 concurrent model runs, with a total walltime of 65.5 h, and a computation time of 18,864 core hours. The evolutionary operators used 0.07% of the computation time, MISCAN-Colon used the remainder. The running times of Experiments 2 and 3 are incomparable because MISCAN-Colon was accelerated in between the two runs.

The number of policies added to the memory per iteration (**Figure 13**) evolved along similar lines as in Experiment 2, where the minority of the policies added are not dominated by a combination of other policies, especially during the last few iterations. The peak at iteration 505 is caused by the changed random number stream for MISCAN-Colon when the runs were transferred from the virtual machine to the Bebop.

In Experiment 3, there were 13 policies to minimally represent the obtained approximation set (**Figure 10**). The figure shows that nearly all reference policies were dominated, except for

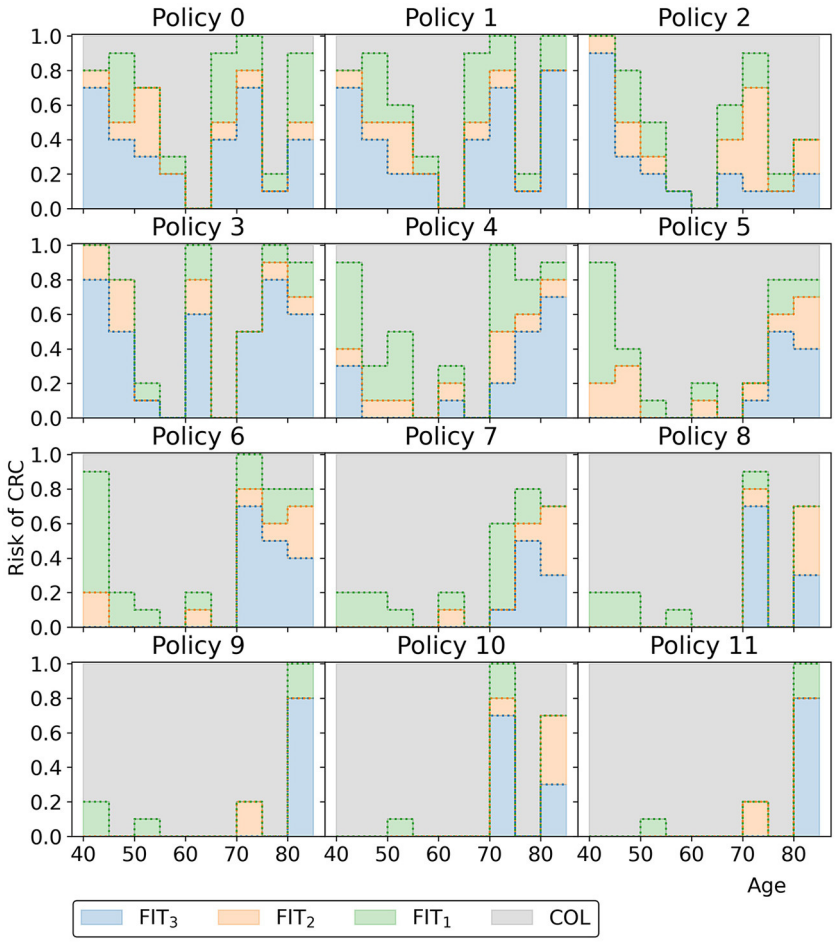


FIGURE 12 | The policies that form the minimal representation of the approximation set of Experiment 2 as shown in **Figure 10**.

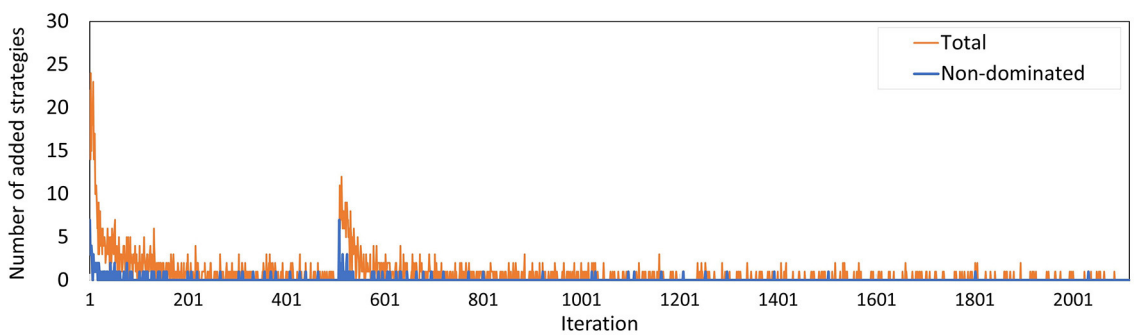


FIGURE 13 | The number of policies added to the memory in each iteration for Experiment 3. The blue line counts all added policies, the orange line only those that are not dominated by (combinations of) other policies in the current memory. The latter are part of the minimal representation of the memory's approximation set. The peak at iteration 506 is caused by the different seeds used on the virtual machine and the Bebop.

two. The two exceptions are marked by white-green diamonds: triennial FIT for ages 45 through 70, and colonoscopy for age ranges 45 and 60 (policies F1 and C1 in **Supplementary Table 3**, resp.). Both policies quit screening relatively early whereas the personalized policies have a fixed stopping age of 85 by design. Disregarding these two reference policies, the QALYs of the obtained screening policies were up to 4.3% higher than the QALYs of the reference policies for similar costs.

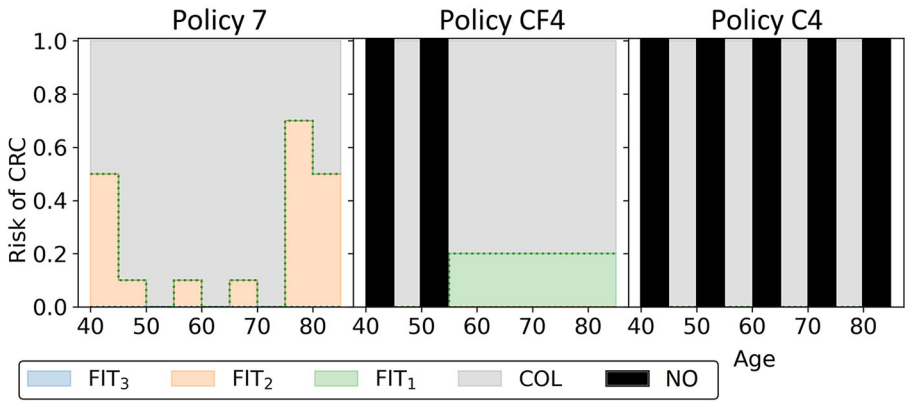


FIGURE 14 | The three black-orange/green-orange policies in **Figure 10** are shown. Policy 7 is the cost-effective policy within the willingness-to-pay threshold in the approximation set for Experiment 3. Policies CF4 and C4 are the closest reference policies in terms of costs and QALYs (see also **Supplementary Table 3**). For age groups with a black bar, no screening is offered.

In this experiment, the black-orange policy in **Figure 10** was the cost-effective policy within the willingness-to-pay threshold (policy 7 in **Figure 14**). The two most similar reference policies with respect to costs and QALYs (marked green-yellow in **Figure 10**) commence screening at 45. Also, the screening intensity of policy 7 is low until age 45 as a cutoff of 50 μg/g is prescribed. The policies stop screening at age 80 or 85, though policy 7 has high cutoffs for colonoscopy referral from age 75. In between, policy 7 effectively prescribes 10-yearly colonoscopy for most participants, in line with US colonoscopy-based screening recommendations and policy C4.

Overall, the other policies in the minimal representation of the obtained approximation set (**Figure 15**) have patterns similar to the policies found in Experiment 2. Screening is primarily focused on the ages 50/55 through 75 for policies cheaper than policy 7. More expensive policies allow more screening in other age ranges, and the most expensive policies are more colonoscopy-based.

3.4. Comparing Experiments 2 and 3

Screening policies in Experiment 2 are more flexible as they have a larger variety in screening intervals compared to Experiment 3. However, with this flexibility, the number of feasible policies increases by a factor 10^{13} (see **Supplementary Section 2**). This means that the algorithm has a larger search space.

Figure 10 shows that the approximation set of Experiment 3 is dominated by that of Experiment 2. **Figures 9**, **13** show that the set was found in fewer iterations in the second experiment compared to the third. This suggests that it may be beneficial to increase the flexibility of the problem by increasing the action space, despite the increased search space.

4. DISCUSSION

In this paper, we demonstrated the computational viability of designing and optimizing personalized FIT-based screening policies using an evolutionary algorithm. The algorithm combines with an advanced simulation model to evaluate the policies. The generated policies prescribed varying screening intervals or referral for a colonoscopy, based on a person's age and measured fecal haemoglobin concentrations. The evolutionary algorithm was used to generate a collection of personalized screening policies, also called an approximation set, that approximates the Pareto frontier, the set of policies with maximum benefits, measured in QALYs gained, for given costs. In our study, an established microsimulation model, MISCAN-Colon, was used to estimate the costs and QALYs of a screening policy.

We demonstrated the performance of the algorithm in three experiments. In the first, we used a relatively small problem instance with 1.5 million feasible policies. We calculated the exact optimal Pareto frontier and tested how well it was approximated by the algorithm. The algorithm could solve this instance to near-optimality, with an optimality gap of 0.007%.

The problem instances of the second and third experiments were too large to derive the exact Pareto frontier. We evaluated the performance of the evolutionary algorithm by (1) comparing the generated policies to a set of reference policies, previously evaluated with MISCAN-Colon in a decision analysis for the United States Preventive Services Task Force (USPSTF), in terms of costs and benefits and (2) assessing the face validity of the obtained policies. First, the generated personalized screening policies generally outperformed the reference policies in terms of costs and QALYs. For a given level of costs, the QALYs gained by the generated policies increased by 14% in Experiment 2 and 4.3% in Experiment 3. In Experiment 2, the computation time of the algorithm was four times shorter than the time of the enumeration process in Experiment 1, despite the 10^{16} times larger search space. This underscores the potential of personalized screening, and of the computational approach presented in this study.

Second, the obtained policies have several interesting features. The cost-effective policies allocated screening predominantly to the ages 50–70 or 45–70 through short intervals and low

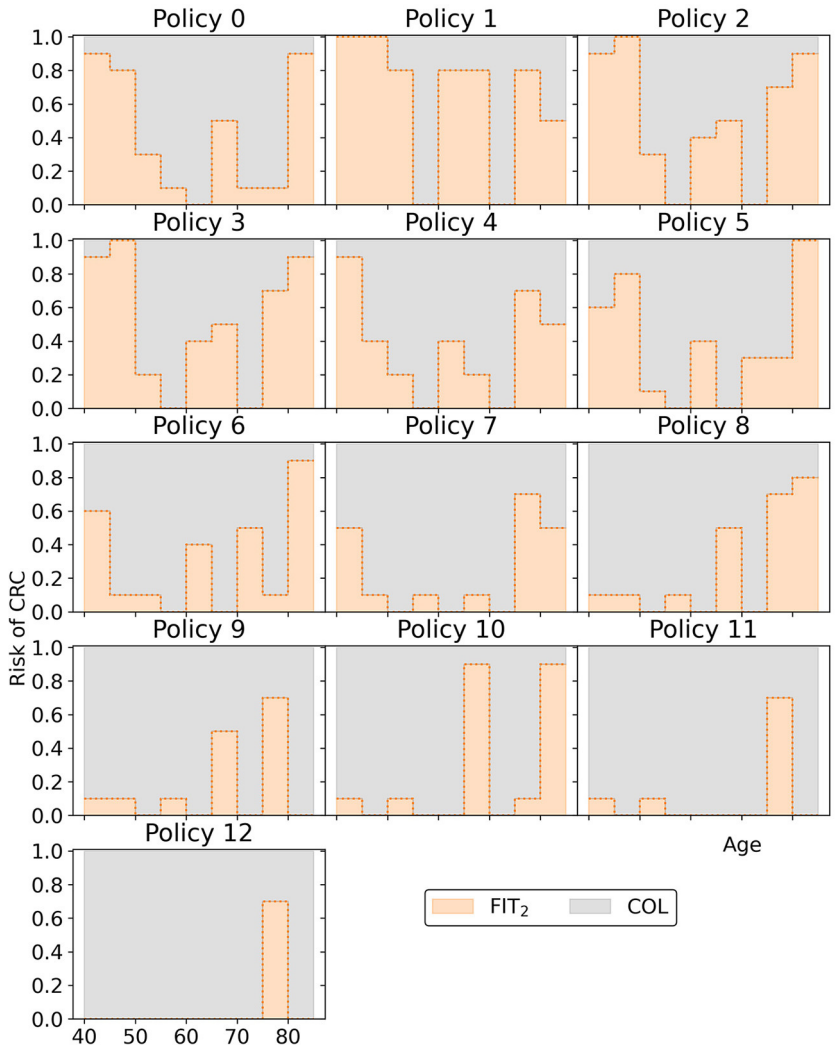


FIGURE 15 | The policies that form the minimal representation of the approximation set of Experiment 3 as shown in **Figure 10**. Note that a cutoff at a perceived risk of 1.0 implies that participants with a FIT-concentration above 100 μg/g are referred for a colonoscopy.

cutoffs for these ages. This is in line with currently implemented policies, which mostly prescribe screening to those aged 50–70 (Schreuders et al., 2015). Cheaper policies increased the intervals and cutoffs for the ages below 55 and above 65. This way, the algorithm narrows the focus of the policies to the ages 55–65 since policies are forced to apply screening from age 40 to 85 by design. More expensive policies expanded the age ranges with low cutoffs and short intervals. Remarkably, all policies guaranteed at least one colonoscopy to all participants by prescribing a FIT-cutoff of 0 μg/g for at least one age range. However, whenever a second guaranteed colonoscopy was offered, the interval from the previous colonoscopy was at least 10 years. This is in accordance with current US colonoscopy-based screening recommendations (Lin et al., 2021). The above observations support the algorithm's face validity, i.e., its ability to generate sensible policies.

In the second experiment, the policies prescribed a larger variety of screening intervals than in the third experiment, resulting in an increase of the search space by a factor 10^{13}. Still, the approximation set found in Experiment 2 dominates the set found in Experiment 3. This suggests that a larger set of screening intervals is beneficial, despite the increased search space.

To the best of our knowledge, this is the first algorithm that optimizes personalized FIT-screening policies evaluated by an advanced microsimulation model. Whereas current methods impose strong Markov assumptions to evaluate generated policies, we evaluated them without such assumptions. The described algorithm is flexible: an individual's risk can be estimated by a variety of estimators, a wide range of actions can be incorporated in the action set, and custom age ranges to which policies apply may be considered. It may also be applied to other diseases when combined with a suitable simulation model that evaluates the costs and benefits of policies, as long as their screening program is based on a test with a quantitative test

result. Examples include prostate specific antigen (PSA) based screening for prostate cancer or mammography screening for breast cancer. Such models are increasingly developed and our algorithm provides enough flexibility that it can be combined with many existing models.

The developed algorithm may be amenable for further improvement. First, it may be possible to enhance the evolutionary operators to search the space of screening policies more efficiently, for example by applying semi-random mutations directed by other simulation outcomes. Second, more fine-grained variations of the belief and action space may be considered, for example including information on prior colonoscopy results in addition to FIT-history, and the option to "stop screening". Furthermore, additional user constraints may be applied to the policies generated by our algorithm, to facilitate easier implementation in practice. For example, it may not be desirable to prescribe guaranteed colonoscopies, or policy makers may want age-independent cutoffs for FIT-positivity for practical reasons. Decision scientists and policy makers should come up with a guideline of what features a policy requires for real-world implementation. We believe the computational framework presented in this paper is sufficiently flexible to incorporate such additional features.

As with any model, results from a microsimulation model are subject to uncertainty, and should be interpreted with caution. MISCAN-Colon was extensively validated in the past on randomized clinical trial data for screening, including fecal-based screening. However, the module for FIT-concentrations was a prototype model for which direct clinical validation was not possible in the scope of this study. It needs further development and validation when more data on the relation between FIT-concentrations and presence of lesions become available. On the other hand, the study shows that using a simpler but faster model could decrease the algorithm's computation time. In Experiments 2 and 3, 99.9% of the algorithm's running time was spent on simulation by MISCAN-Colon, despite parallel computations. However, this may be at the cost of decreased accuracy in the evaluation of the policies.

To conclude, we demonstrated a potential method for identifying optimized personalized screening policies while evaluating them with established simulation models from practice. This moves the field a step closer to implementing personalized screening in practice.

DATA AVAILABILITY STATEMENT

An implementation of the algorithm in Python is available at https://gitlab.com/luukvandEMC/ea_personalized_screening. It includes a fictive dataset similar to the benchmark set of Experiment 1.

AUTHOR CONTRIBUTIONS

LD developed the algorithm, performed the analysis, and wrote the manuscript. RM, RS, and IL-V contributed as supervisors. RM also helped develop the module to simulate fecal occult blood loss. NC and JO conducted the experiments on the High Performance Computer. All authors contributed to the manuscript and approved the submitted version.

FUNDING

The microsimulation analysis was supported by Grant U01-CA199335 and Grant U01-CA253913 from the National Cancer Institute (NCI) as part of the Cancer Intervention and Surveillance Modeling Network (CISNET). The content is solely the responsibility of the authors and does not necessarily represent the official views of the National Institutes of Health. The work was supported in part by the U.S. Department of Energy, Office of Science, under contract (No. DE- AC02-06CH11357).

ACKNOWLEDGMENTS

This research was completed with resources provided by the Laboratory Computing Resource Center at Argonne National Laboratory (Bebop cluster).

SUPPLEMENTARY MATERIAL

The Supplementary Material for this article can be found online at: https://www.frontiersin.org/articles/10.3389/fphys.2021.718276/full#supplementary-material

REFERENCES

Ahuja, K., Zame, W., and van der Schaar, M. (2017). "DPScreen: dynamic personalized screening," in *Advances in Neural Information Processing Systems*, eds I. Guyon, U. V. Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, and R. Garnett (Long Beach, CA: Curran Associates, Inc.), 1321–1332.

Alagoz, O., Berry, D. A., de Koning, H. J., Feuer, E. J., Lee, S. J., Plevritis, S. K., et al. (2018). Introduction to the cancer intervention and surveillance modeling network (CISNET) breast cancer models. *Med. Decis. Making* 38(1_Suppl):3S–8S. doi: 10.1177/0272989X17737507

Ayer, T., Alagoz, O., and Stout, N. K. (2012). Or forum–a pomdp approach to personalize mammography screening decisions. *Operat. Res.* 60, 1019–1034. doi: 10.1287/opre.1110.1019

Buskermolen, M., Gini, A., Naber, S. K., Toes-Zoutendijk, E., de Koning, H. J., and Lansdorp-Vogelaar, I. (2018). Modeling in colorectal cancer screening: assessing external and predictive validity of miscan-colon microsimulation model using norccap trial results. *Med. Decis. Making* 38, 917–929. doi: 10.1177/0272989X18806497

Criss, S. D., Cao, P., Bastani, M., Ten Haaf, K., Chen, Y., Sheehan, D. F., et al. (2019). Cost-effectiveness analysis of lung cancer screening in the United States: a comparative modeling study. *Ann. Internal Med.* 171, 796–804. doi: 10.7326/M19-0322

Deb, K., Pratap, A., Agarwal, S., and Meyarivan, T. (2002). A fast and elitist multiobjective genetic algorithm: NSGA-II. *IEEE Trans. Evol. Comput.* 6, 182–197. doi: 10.1109/4235.996017

DeYoreo, M., Lansdorp-Vogelaar, I., Knudsen, A. B., Kuntz, K. M., Zauber, A. G., and Rutter, C. M. (2020). Validation of colorectal

cancer models on long-term outcomes from a randomized controlled trial. *Med. Decis. Making* 40, 1034–1040. doi: 10.1177/0272989X20961095

Erenay, F. S., Alagoz, O., and Said, A. (2014). Optimizing colonoscopy screening for colorectal cancer prevention and surveillance. *Manufactur. Serv. Operat. Manage.* 16, 381–400. doi: 10.1287/msom.2014.0484

Fortin, F.-A., Rainville, F.-M. D., Gardner, M.-A., Parizeau, M., and Gagne, C. (2012). DEAP: evolutionary algorithms made easy. *J. Mach. Learn. Res.* 13, 2171–2175. Available online at: http://jmlr.org/papers/v13/fortin12a.html

Gini, A., Buskermolen, M., Senore, C., Anttila, A., Novak Mlakar, D., Veerus, P., et al. (2021). Development and validation of three regional microsimulation models for predicting colorectal cancer screening benefits in europe. *MDM Policy Pract.* 6:2381468320984974. doi: 10.1177/2381468320984974

Gini, A., Zauber, A. G., Cenin, D. R., Omidvari, A.-H., Hempstead, S. E., Fink, A. K., et al. (2017). Cost-effectiveness of screening individuals with cystic fibrosis for colorectal cancer. *Gastroenterology*. doi: 10.1053/j.gastro.2017.12.011. [Epub ahead of print].

Grobbee, E. J., Schreuders, E. H., Hansen, B. E., Bruno, M. J., Lansdorp-Vogelaar, I., Spaander, M. C., et al. (2017). Association between concentrations of hemoglobin determined by fecal immunochemical tests and long-term development of advanced colorectal neoplasia. *Gastroenterology* 153, 1251–1259. doi: 10.1053/j.gastro.2017.07.034

Gulati, R., Wever, E. M., Tsodikov, A., Penson, D. F., Inoue, L. Y., Katcher, J., et al. (2011). What if i don't treat my psa-detected prostate cancer? Answers from three natural history models. *Cancer Epidemiol. Prev. Biomark.* 20, 740–750. doi: 10.1158/1055-9965.EPI-10-0718

Holland, J. (1975). *Adaptation in Natural and Artificial Systems.* Ann Arbor: University of Michigan Press.

Knudsen, A. B., Rutter, C. M., Peterse, E. F. P., Lietz, A. P., Seguin, C. L., Meester, R. G. S., et al. (2020). Colorectal cancer screening: a decision analysis for the U.S. preventive services task force. *JAMA* 325, 1998–2011. doi: 10.1001/jama.2021.5746

Knudsen, A. B., Zauber, A. G., Rutter, C. M., Naber, S. K., Doria-Rose, V. P., Pabiniak, C., et al. (2016). Estimation of benefits, burden, and harms of colorectal cancer screening strategies: modeling study for the us preventive services task force. *JAMA* 315, 2595–2609. doi: 10.1001/jama.2016.6828

Lin, J. S., Perdue, L. A., Henrikson, N. B., Bean, S. I., and Blasi, P. R. (2021). Screening for colorectal cancer: updated evidence report and systematic review for the US preventive services task force. *JAMA* 325, 1978–1997. doi: 10.1001/jama.2021.4417

Loeve, F., Boer, R., van Oortmarssen, G. J., van Ballegooijen, M., and Habbema, J. D. F. (1999). The miscan-colon simulation model for the evaluation of colorectal cancer screening. *Comput. Biomed. Res.* 32, 13–33. doi: 10.1006/cbmr.1998.1498

Maillart, L. M., Ivy, J. S., Ransom, S., and Diehl, K. (2008). Assessing dynamic breast cancer screening policies. *Operat. Res.* 56, 1411–1427. doi: 10.1287/opre.1080.0614

Meester, R. G., Peterse, E. F., Knudsen, A. B., de Weerdt, A. C., Chen, J. C., Lietz, A. P., et al. (2018). Optimizing colorectal cancer screening by race and sex: microsimulation analysis II to inform the american cancer society colorectal cancer screening guideline. *Cancer* 124, 2974–2985. doi: 10.1002/cncr.31542

Otten, J., Witteveen, A., Vliegen, I., Siesling, S., Timmer, J. B., and IJzerman, M. J. (2017). "Stratified breast cancer follow-up using a partially observable MDP," in *Markov Decision Processes in Practice*, eds R. J. Boucherie and N. M. van Dijk (Cham: Springer), 223–244. doi: 10.1007/978-3-319-47766-4_7

Ozik, J., Collier, N. T., Wozniak, J. M., and Spagnuolo, C. (2016). "From desktop to large-scale model exploration with Swift/T," in *2016 Winter Simulation Conference (WSC)* (Arlington), 206–220. doi: 10.1109/WSC.2016.7822090

Riquelme, N., Von Lücken, C., and Baran, B. (2015). "Performance metrics in multi-objective optimization," in *2015 Latin American Computing Conference (CLEI)* (Arequipa: IEEE), 1–11. doi: 10.1109/CLEI.2015.7360024

Rutter, C. M., and Savarino, J. E. (2010). An evidence-based microsimulation model for colorectal cancer: validation and application. *Cancer Epidemiol. Prev. Biomarkers* 19, 1992–2002. doi: 10.1158/1055-9965.EPI-09-0954

Sanders, G. D., Neumann, P. J., Basu, A., Brock, D. W., Feeny, D., Krahn, M., et al. (2016). Recommendations for conduct, methodological practices, and reporting of cost-effectiveness analyses: second panel on cost-effectiveness in health and medicine. *JAMA* 316, 1093–1103. doi: 10.1001/jama.2016.12195

Schreuders, E. H., Ruco, A., Rabeneck, L., Schoen, R. E., Sung, J. J., Young, G. P., et al. (2015). Colorectal cancer screening: a global overview of existing programmes. *Gut* 64, 1637–1649. doi: 10.1136/gutjnl-2014-309086

SEER (2021). *Surveillance, Epidemiology, and End Results (SEER) Program* (www.seer.cancer.gov). seer*stat database: Incidence - seer research data, 9 registries.

Sung, H., Ferlay, J., Siegel, R. L., Laversanne, M., Soerjomataram, I., Jemal, A., et al. (2021). Global cancer statistics 2020: GLOBOCAN estimates of incidence and mortality worldwide for 36 cancers in 185 countries. *CA Cancer J. Clin.* 71, 209–249. doi: 10.3322/caac.21660

Toes-Zoutendijk, E., van Leerdam, M. E., Dekker, E., Van Hees, F., Penning, C., Nagtegaal, I., et al. (2017). Real-time monitoring of results during first year of dutch colorectal cancer screening program and optimization by altering fecal immunochemical test cut-off levels. *Gastroenterology* 152, 767–775. doi: 10.1053/j.gastro.2016.11.022

van Hees, F., Habbema, J. D. F., Meester, R. G., Lansdorp-Vogelaar, I., van Ballegooijen, M., and Zauber, A. G. (2014). Should colorectal cancer screening be considered in elderly persons without previous screening? A cost-effectiveness analysis. *Ann. Internal Med.* 160, 750–759. doi: 10.7326/M13-2263

Whitley, D. (1994). A genetic algorithm tutorial. *Stati. Comput.* 4, 65–85. doi: 10.1007/BF00175354

Zitzler, E., and Thiele, L. (1998). "Multiobjective optimization using evolutionary algorithms–A comparative case study," in *Parallel Problem Solving from Nature–PPSN V. PPSN 1998*, eds A. E. Eiben, T. Bäck, M. Schoenauer, H. P. Schwefel (Berlin: Springer), 292–301. doi: 10.1007/BFb0056872

Zitzler, E., Thiele, L., Laumanns, M., Fonseca, C. M., and Da Fonseca, V. G. (2003). Performance assessment of multiobjective optimizers: an analysis and review. *IEEE Trans. Evol. Comput.* 7, 117–132. doi: 10.1109/TEVC.2003.810758

Conflict of Interest: The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

Publisher's Note: All claims expressed in this article are solely those of the authors and do not necessarily represent those of their affiliated organizations, or those of the publisher, the editors and the reviewers. Any product that may be evaluated in this article, or claim that may be made by its manufacturer, is not guaranteed or endorsed by the publisher.

frontiers | Frontiers in Physiology

ORIGINAL RESEARCH
published: 09 May 2022
doi: 10.3389/fphys.2022.780917

Characterization and Valuation of the Uncertainty of Calibrated Parameters in Microsimulation Decision Models

Fernando Alarid-Escudero[1]*, *Amy B. Knudsen*[2], *Jonathan Ozik*[3,4], *Nicholson Collier*[3,4] and *Karen M. Kuntz*[5]

[1]Division of Public Administration, Center for Research and Teaching in Economics (CIDE), Aguascalientes, Mexico, [2]Institute for Technology Assessment, Massachusetts General Hospital, Boston, MA, United States, [3]Decision and Infrastructure Sciences Division, Argonne National Laboratory, Argonne, IL, United States, [4]Consortium for Advanced Science and Engineering, University of Chicago, Chicago, IL, United States, [5]Division of Health Policy and Management, University of Minnesota School of Public Health, Minneapolis, MN, United States

Background: We evaluated the implications of different approaches to characterize the uncertainty of calibrated parameters of microsimulation decision models (DMs) and quantified the value of such uncertainty in decision making.

Methods: We calibrated the natural history model of CRC to simulated epidemiological data with different degrees of uncertainty and obtained the joint posterior distribution of the parameters using a Bayesian approach. We conducted a probabilistic sensitivity analysis (PSA) on all the model parameters with different characterizations of the uncertainty of the calibrated parameters. We estimated the value of uncertainty of the various characterizations with a value of information analysis. We conducted all analyses using high-performance computing resources running the Extreme-scale Model Exploration with Swift (EMEWS) framework.

Results: The posterior distribution had a high correlation among some parameters. The parameters of the Weibull hazard function for the age of onset of adenomas had the highest posterior correlation of −0.958. When comparing full posterior distributions and the maximum-a-posteriori estimate of the calibrated parameters, there is little difference in the spread of the distribution of the CEA outcomes with a similar expected value of perfect information (EVPI) of $653 and $685, respectively, at a willingness-to-pay (WTP) threshold of $66,000 per quality-adjusted life year (QALY). Ignoring correlation on the calibrated parameters' posterior distribution produced the broadest distribution of CEA outcomes and the highest EVPI of $809 at the same WTP threshold.

Conclusion: Different characterizations of the uncertainty of calibrated parameters affect the expected value of eliminating parametric uncertainty on the CEA. Ignoring inherent correlation among calibrated parameters on a PSA overestimates the value of uncertainty.

Keywords: microsimulation models, uncertainty quantification, calibration, Bayesian, value of information analysis, decision-analytic models, high-performance computing, EMEWS

OPEN ACCESS

Edited by:
Nicole Y. K. Li-Jessen, McGill University, Canada

Reviewed by:
Daniele E. Schiavazzi, University of Notre Dame, United States
Rowan Iskandar, sitem-insel, Switzerland

***Correspondence:**
Fernando Alarid-Escudero
fernando.alarid@cide.edu

Specialty section:
This article was submitted to Computational Physiology and Medicine, a section of the journal Frontiers in Physiology

Received: *22 September 2021*
Accepted: *04 April 2022*
Published: *09 May 2022*

Citation:
Alarid-Escudero F, Knudsen AB, Ozik J, Collier N and Kuntz KM (2022) Characterization and Valuation of the Uncertainty of Calibrated Parameters in Microsimulation Decision Models. Front. Physiol. 13:780917. doi: 10.3389/fphys.2022.780917

BACKGROUND

Decision models (DMs) are commonly used in cost-effectiveness analysis where uncertainty in the parameters is inherent (Kuntz et al., 2017). The impact of parameter uncertainty can be assessed with a probabilistic sensitivity analysis (PSA) to characterize decision uncertainty (i.e., the probability of a strategy being cost-effective) (Briggs et al., 2012; Sculpher et al., 2017) and to quantify the value of potential future research by determining the potential consequences of a decision with value of information (VOI) analysis (Schlaifer, 1959; Raiffa and Schlaifer, 1961).

The parameters of DMs can be split into two categories, those obtained from the literature or estimated from available data (i.e., external parameters) and those that need to be estimated through calibration (i.e., calibrated parameters). External parameters are estimated either from individual-level or aggregated data that directly inform the parameters of interest. There are recommendations on the type of distributions that characterize their uncertainty based on the characteristics of the parameters or the statistical model used to estimate them (Briggs et al., 2012). For example, a probability could be modeled with a beta distribution and a relative risk with a lognormal distribution (Briggs et al., 2002). For calibrated parameters, no such data exist that can directly inform their uncertainty because a research study hasn't been conducted or is unfeasible to conduct, or because the parameters reflect unobservable phenomena, as is often the case in natural history models of chronic diseases (Welton and Ades, 2005; Karnon et al., 2007; Rutter et al., 2009; Rutter et al., 2011) or in infectious disease dynamic models (Enns et al., 2017). The choice of distribution for these parameters is often less clear. One option is to define uniform distributions with wide bounds or generate informed distributions based on moments of the calibrated parameters, such as the mean and standard error. However, the impact of these approaches to characterize the uncertainty of calibrated parameters on decision uncertainty and the VOI on reducing that uncertainty has not been studied.

Model calibration is the process of estimating unobserved or unobservable parameters by matching model outputs to observed clinical or epidemiological data (known as calibration targets) (Kennedy and O'Hagan, 2001; Stout et al., 2009; Kuntz et al., 2017). While there are several approaches for searching the parameter space in the calibration process, most approaches are insufficient to characterize the uncertainty in the calibrated model parameters because they do not provide interval estimates. For example, direct-search optimization algorithms like Newton-Raphson Nelder-Mead (Nelder and Mead 1965) simulated annealing or genetic algorithms (Kong et al., 2009) treat the calibration targets as if they were known with certainty, so are primarily useful when identifying a single or a set of parameters that yield good fit to the targets (Kennedy and O'Hagan, 2001).

A sample of calibrated parameter sets that correctly characterizes the uncertainty of the calibration target data is obtained from their joint distribution, conditional on the calibrated targets. To obtain the joint distribution, calibration could be specified as a statistical estimation problem under at least two different frameworks, through maximum likelihood (ML) or Bayesian methods. ML can fail in obtaining interval estimates by not being able to estimate the Hessian matrix when the likelihood is intractable or computationally intensive to simulate and when the calibration problem is non-identifiable (Gustafson, 2005; Alarid-Escudero et al., 2018); thus, we focus on Bayesian methods (Romanowicz et al., 1994; Kennedy and O'Hagan, 2001; Oakley and O'Hagan, 2004; Gustafson, 2005; Kaipio and Somersalo, 2005; Oden et al., 2010; Gustafson, 2015; Alarid-Escudero et al., 2018).

Despite their suitability to correctly characterize the uncertainty of calibrated model parameters, Bayesian methods are generally computationally expensive because they require evaluating the model thousands and sometimes millions of times. The computational burden of Bayesian methods does not seem to be an impediment when calibrating non-computationally intensive DMs (e.g., Markov cohort models, difference equations, relatively small systems of differential equations, etc.) (Whyte et al., 2011; Hawkins-Daarud et al., 2013; Jackson et al., 2016; Menzies et al., 2017). Still, they become more challenging to apply to DMs that could be computationally intensive to solve, such as models that simulate underlying stochastic processes (Iskandar, 2018) (e.g., microsimulation, discrete-event simulation, and agent-based models), limiting their use to only a few of such models (Rutter et al., 2009).

However, the increasing availability of high-performance computing (HPC) systems in an academic, national laboratory and commercial settings enables such systems for model calibration and model exploration of microsimulation DMs at a large scale to a broader audience. HPC resources allow running large numbers of DMs concurrently, allowing calibration algorithms to generate large batches of parameters simultaneously, such as the incremental mixture importance sampling (IMIS) described below, to be run efficiently. In many cases, particularly in the academic and national laboratory settings, computing allocations can be obtained through proposals with no cost to researchers (e.g., the Advanced Scientific Computing Research (ASCR) Leadership Computing Challenge (ALCC), https://science.osti.gov/ascr/Facilities/Accessing-ASCR-Facilities/ALCC). However, implementing dynamic calibration algorithms for HPC resources has generally proved difficult, requiring specialized knowledge across various disciplines. The Extreme-scale Model Exploration with Swift (EMEWS) framework was designed to facilitate large-scale model calibration and exploration on HPC resources (Ozik et al., 2016a) to a broad community. EMEWS can run very large, highly concurrent ensembles of microsimulation DMs of varying types with a broad class of calibration algorithms, including those increasingly available to the community via. Python and R libraries, using HPC workflows. EMEWS workflows provide interfaces for plugging in DMs (and any other simulation or black box model) and algorithms, through an inversion of control scheme (Ozik et al., 2018), to control the dynamic execution of those DMs for calibration and other heuristics for "model exploration" purposes. These interfaces help reduce the need for an in-depth understanding of how task coordination and inter-task dependencies are

implemented for HPC resources. The general use of EMEWS can be seen on the EMEWS website (https://emews.github.io), which includes links to tutorials.

The purpose of our study is threefold. First, to use recently developed HPC capabilities to characterize the uncertainty of calibrated parameters of a microsimulation model of the natural history of colorectal cancer (CRC). Second, to explore the impact of different approaches to characterize the uncertainty of calibrated parameters on decision uncertainty, and third, to use VOI analysis to quantify the value of eliminating parameter uncertainty when assessing the cost-effectiveness of CRC screening.

METHODS

We developed a microsimulation model of the natural history of CRC and calibrated it using a Bayesian approach. We then overlaid a simple CRC screening strategy onto the natural history model and conducted a cost-effectiveness analysis (CEA) of screening, including a PSA. Instead of using the posterior means to represent the best estimates of each calibrated parameter, we obtained the posterior distribution using a Bayesian approach that represents the joint uncertainty of all the calibrated parameters that can then be used in a PSA. We then evaluated the impact of different approaches to characterize the uncertainty of calibrated parameters on the joint distribution of incremental costs and incremental effects of the screening strategy compared with no screening through a PSA while also accounting for the uncertainty of the external parameters (e.g., test characteristics, costs, etc.). Finally, we quantified the amount of money that a decision maker should be willing to spend to eliminate all parameter uncertainty (i.e., the expected value of perfect information (EVPI)).

Microsimulation Model of the Natural History of CRC

We developed a state-transition microsimulation model of the natural history of CRC implemented in R (Krijkamp et al., 2018) based on a previously developed model (Alarid-Escudero et al., 2018). The progression between health states follows a continuous-time age-dependent Markov process. There are two age-dependent transition intensities (i.e., transition rates), $\lambda_1(a)$ and $\mu(a)$, that govern the age of onset of adenomas and non-cancer-specific mortality, respectively. Following Wu et al. (2006) we specify $\lambda_1(a)$ as a Weibull hazard with the following specification

$$\lambda_1(a) = l\gamma a^{\gamma-1},$$

where a is the age of the simulated individuals, and l and γ are the scale and shape parameters of the Weibull hazard function, respectively. The model simulates two adenoma categories: small (adenoma smaller than 1 cm in size) and large (adenoma larger than or equal to 1 cm in size). All adenomas start small and can transition to the large size category at a constant annual rate λ_2. Large adenomas may become preclinical CRC at a constant annual rate λ_3. Both small and large adenomas may progress to preclinical CRC, although most will not in a simulated individual's lifetime. Early preclinical cancers (preclinical stages I and II) progress to late stages (preclinical stages III and IV) at a constant annual rate λ_4 and could become symptomatic at a constant annual rate λ_5. Late preclinical cancer could become symptomatic at a constant annual rate λ_6. After clinical detection, the model simulates the survival time to early and late CRC death using cancer-specific constant mortality rates, λ_7 and λ_8, respectively. The model has nine health states: normal, small adenoma, large adenoma, early preclinical CRC, late preclinical CRC, early clinical CRC, late clinical CRC, CRC death, and death from other causes. The state-transition diagram of the continuous-time model is shown in **Figure 1**.

The continuous-time age-dependent Markov process of this natural history model of CRC can be represented by an age-dependent 9×9 transition intensity matrix, $Q(a)$. To translate $Q(a)$ to discrete-time, we compute the annual-cycle age-dependent transition probability matrix, $P(a,t)$, using the Kolmogorov differential equations (Kolmogorov, 1963; Cox and Miller, 1965; Welton and Ades, 2005)

$$P(a,t) = Exp(tQ(a)),$$

where $t = 1$ and $Exp()$ is the matrix exponential. In discrete time, the natural history model of CRC allows individual transitions across multiple health states in a single year. Small and large adenomas may progress to preclinical or clinical CRC, and preclinical cancers may progress through early and late stages.

We simulated a hypothetical cohort of 50-year-old women in the United States over a lifetime. The cohort starts the simulation with a prevalence of adenoma of p_{adeno}, from which a proportion, p_{small}, corresponds to small adenomas, and a prevalence of preclinical early and late CRC of 0.12% (Rutter et al., 2007) and 0.08% (Wu et al., 2006), respectively. The parameters p_{adeno} and p_{small} are calibrated parameters. The simulated cohort is at risk of all-cause mortality, $\mu(a)$, from all health states obtained from 2014 United States life tables (Arias et al., 2017).

Calibration Targets

We used the microsimulation model of the natural history of CRC to generate synthetic calibration targets by selecting a set of parameter values based on plausible estimates from the literature (**Table 1**) (Wu et al., 2006; Rutter et al., 2007). We simulated four different age-specific synthetic targets, including adenoma prevalence, the proportion of small adenomas, and CRC incidence for early and late stages, which resemble commonly used calibration targets for this type of model (Rutter et al., 2009; Whyte et al., 2011; Frazier et al., 2000; Kuntz et al., 2011). To simulate the calibration targets, we ran the microsimulation model 100 times to get a stable estimate of the standard errors (SEs) using the fixed values in **Table 1**. We then aggregated each outcome across all 100 model replications to compute their mean and SE. To

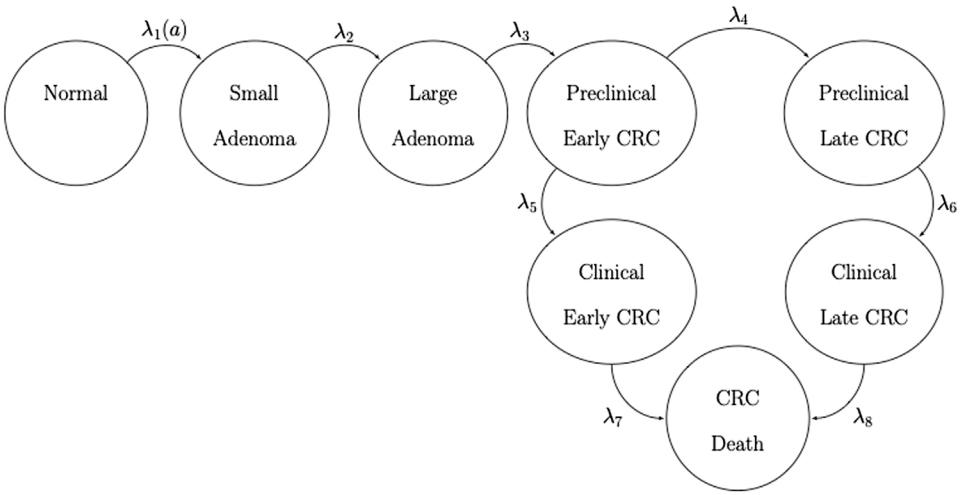


FIGURE 1 | State-transition diagram of the nine-state microsimulation model of the natural history of colorectal cancer. Individuals in all health states face an age-specific mortality of dying from other causes (state not shown) (Jalal et al., 2021).

TABLE 1 | Description of parameters of the natural history model.

Symbol	Description	Value	Source	Prior distribution	Calibrated
Initial state of 50-year-old cohort					
		Proportions			
p_{adeno}	Prevalence of adenoma at age 50	0.25	Rutter et al. (2007)	Beta(3, 8)	Yes
p_{small}	Proportion adenomas that are small at age 50	0.71	Wu et al. (2006)	Beta(6, 3)	Yes
—	Prevalence of preclinical early CRC at age 50	0.12	Wu et al. (2006)	Fixed	No
—	Prevalence of preclinical late CRC at age 50	0.08	Wu et al. (2006)	Fixed	No
Disease dynamics					
		Transition rates (annual)			
l	Scale parameter of Weibull hazard	2.86e-06	Wu et al. (2006)	Log-normal(m = –11.97, s = 0.59)	Yes
γ	Shape parameter of Weibull hazard	2.78	Wu et al. (2006)	Log-normal(m = 1.04, s = 0.18)	Yes
λ_2	Small adenoma to large adenoma	0.0346	Wu et al. (2006)	Log-normal(m = –3.45, s = 0.59)	Yes
λ_3	Large adenoma to preclinical early CRC	0.0215	Wu et al. (2006)	Log-normal(m = –3.91, s = 0.35)	Yes
λ_4	Preclinical early CRC to preclinical late CRC	0.3697	Wu et al. (2006)	Log-normal(m = –1.15, s = 0.23)	Yes
λ_5	Preclinical early CRC to clinical early CRC	0.2382	Wu et al. (2006)	Log-normal(m = –1.41, s = 0.10)	Yes
λ_6	Preclinical late CRC to clinical late CRC	0.4582	Wu et al. (2006)	Log-normal(m = –0.78, s = 0.22)	Yes
λ_7	CRC mortality in early stage	0.0302	Wu et al. (2006)	Fixed	No
λ_8	CRC mortality in late stage	0.2099	Wu et al. (2006)	Fixed	No
$\mu(a)$	Age-specific mortality	Age-specific	Arias, (2017)	Fixed	No

account for different levels of uncertainty across targets given the amount of data to estimate their summary measures, we simulated various targets based on cohorts of different sizes (Rutter et al., 2009). Adenoma-related targets were based on a cohort of 500 individuals, and cancer incidence targets were based on 100,000 individuals.

Calibration of the Microsimulation Model of the Natural History

To state the calibration of the microsimulation model as an estimation problem (Alarid-Escudero et al., 2018), we define M as the microsimulation model of the natural history of CRC with 11 input parameters. Cancer-specific mortality rates from early and late stages of CRC could be obtained from cancer population registries (e.g., the Surveillance, Epidemiology and End Results (SEER) registry in the United States), so calibration of these rates was unnecessary. That is, $\theta_k = [\lambda_7, \lambda_8]$ is a set of 2 parameters that are either known or could be obtained from external data (i.e., are external parameters). The model has a set of 9 parameters $\theta_u = [p_{adeno}, p_{small}, l, \gamma, \lambda_2, \lambda_3, \lambda_4, \lambda_5, \lambda_6]$ that cannot be directly estimated from sample data and need to be calibrated. M's full set of parameters is $\theta = [\theta_u, \theta_k]$.

To calibrate M, we adopted a Bayesian approach that allowed us to obtain a joint posterior distribution that characterizes the uncertainty of both the calibration targets and previous knowledge of the parameters of interest in the form of prior distributions. Prior distributions can reflect experts' opinions, or when little knowledge is available, these could be specified as uniform distributions. We constructed the likelihood function by assuming that each type of target t, including adenoma prevalence, proportion of small adenomas, early clinical CRC incidence, and late clinical CRC incidence for each age group a, y_{t_a}, are normally distributed with mean ϕ_{t_a} and standard deviation σ_{t_a} (Alarid-Escudero et al., 2018). That is,

$$y_{t_a} \sim Normal\left(\phi_{t_a}, \sigma_{t_a}\right),$$

where $\phi_{t_a} = \mathrm{E}[\mathrm{M}(\theta)]$ is the expected value of the model-predicted output from parameter set θ. We added the log-likelihoods across all targets to compute an aggregated likelihood measure. We defined prior distributions for all θ_u based on previous knowledge or the nature of the parameters (**Table 1**). We defined beta distributions for the prevalence of adenomas and the proportion of small adenomas at age 50, bounded between 0 and 1. We assumed that the annual transition rates follow a log-normal distribution for their priors, defined over positive numbers. The ranges given in **Table 1** are assumed to represent the 95% equal-tailed interval for the beta and log-normal distributions.

To conduct the Bayesian calibration, we used the incremental mixture importance sampling (IMIS) algorithm (Steele et al., 2006; Raftery and Bao, 2009), which has been previously used to calibrate health policy models (Menzies et al., 2017; Ryckman et al., 2020). We ran the IMIS algorithm on the Midway2 cluster at the University of Chicago Research Computing Center (https://rcc.uchicago.edu/resources/high-performance-computing). Midway2 is a hybrid cluster, including both central processing unit (CPU) and graphics processing unit (GPU) resources. For this work, we used the CPU resources. Midway2 consists of 370 nodes of Intel E5-2680v4 processors, each with 28 cores and 64 GB of RAM. Using EMEWS, we developed a workflow that parallelized the likelihood evaluations over 1,008 processes using 36 compute nodes. In other words, we reduced the computation time approximately by 250 had the analysis been conducted in a laptop with four processing cores.

Consistent with previous analyses, we deemed that convergence had occurred when the target effective sample size (ESS) got as close as 5,000 (Rutter et al., 2019; DeYoreo et al., 2022). An advantage of IMIS over other Monte Carlo methods, such as Markov chain Monte Carlo, is that with IMIS, we parallelize the evaluation of the likelihood for different sampled parameter sets, making its implementation perfectly suitable for an HPC environment using EMEWS. IMIS requires defining and computing the likelihood, which we could do with our model. However, when computing the likelihood is intractable, modelers could use the incremental mixture approximate Bayesian computation (IMABC) algorithm (Rutter et al., 2019), which an approximate Bayesian version of IMIS.

Propagation of Uncertainty

We sampled 5,000 parameter sets from the IMIS joint posterior distribution for the nine calibrated model parameters. To compare the outputs of the calibrated model against the calibration targets, we propagated the uncertainty of the calibrated parameters through the microsimulation model of the natural history of CRC. We simulated a cohort of 100,000 (i.e., the largest cohort size used to generate the targets). We generated the model-predicted adenoma and cancer outcomes for each of the 5,000 calibrated parameter sets drawn from their joint posterior distribution. We computed the 95% posterior predicted interval (PI), defined as the estimated range between the 2.5th and 97.5th percentiles of the model-predicted posterior outputs to quantify the uncertainty limit model outputs.

Cost-Effectiveness Analysis of Screening for CRC

With the calibrated microsimulation model of the natural history of CRC, we assessed the cost-effectiveness of 10-yearly colonoscopy screening starting at age 50 years compared to no screening. For adenomas detected with colonoscopy, a polypectomy was performed during the procedure. Individuals diagnosed with a small or large adenoma underwent surveillance with colonoscopy every 5 or 3 years, respectively. We assumed screening or surveillance continued until 85 years of age. Individuals with a history of polyp diagnosis had higher recurrence rates after polypectomy, that is, a higher transition rate from normal to small adenoma (i.e., $\lambda_1(a)$). We assumed a hazard ratio of 2 for small adenomas and 3 for the large adenomas. The costs and utilities of CRC care varied by stage, and individuals without clinical CRC had a utility of 1. **Table 2** shows the parameters used in the CEA with their corresponding distributions.

Uncertainty Quantification

We performed four different approaches to quantify the uncertainty of the two types of parameters—calibrated parameters and external (i.e., CEA) parameters. The first approach for uncertainty quantification considers uncertainty in both types of parameters, with uncertainty of the calibrated parameters characterized by their joint posterior distribution obtained from the IMIS algorithm. The second approach only considers uncertainty in the external parameters while fixing the calibrated parameters at the *maximum-a-posteriori* (MAP) estimate, defined as the parameter with the highest posterior density. The third approach considers uncertainty only in the calibrated parameters characterized by their joint posterior distribution and no uncertainty in the external parameters, fixed at their mean values. The fourth approach considers uncertainty in both types of parameters, but instead of using the IMIS posterior distribution of the calibrated parameters, we constructed distributions based solely on the IMIS posterior moments (i.e., means and standard deviations) and the type of calibrated parameters ignoring correlations.

We conducted a PSA to evaluate the impact of uncertainty in model parameters on the cost-effectiveness of 10-years

TABLE 2 | Description of cost-effectiveness analysis parameters.

Parameter	Value (range)	Distribution	Source
Screening test characteristics (location-specific)			
Small adenomas			
Sensitivity	0.773 (0.734–0.808)	Beta	Van Rijn et al. (2006)
Specificity	0.868 (0.855–0.880)	Beta	Schroy et al. (2013)
Large adenomas and CRC			
Sensitivity	0.950 (0.920–0.990)	Beta	Van Rijn et al. (2006)
Specificity	0.868 (0.855–0.880)	Beta	Schroy et al. (2013)
Increased rates after polypectomy (hazard ratio)			
Low risk	2 (1–3)	Log-normal	Assumed
High risk	3 (2–4)	Log-normal	Assumed
Costs ($)			
Colonoscopy	10,000 (9,000–11,000)	Log-normal	Assumed
Early clinical CRC, annual costs	21,524 (20,000–23,000)	Log-normal	Assumed
Late clinical CRC, annual costs	37,000 (35,000–39,000)	Log-normal	Assumed
Utilities			
Preclinical CRC	1.000 (0.980–1.000)	Log-normal	Assumed
Early clinical CRC	0.855 (0.700–0.900)	Log-normal	Ness et al. (1999)
Late clinical CRC	0.300 (0.200–0.400)	Log-normal	Ness et al. (1999)

TABLE 3 | Posterior means, standard deviations, maximum-a-posteriori (MAP) estimate and 95% credible interval (CrI) of calibrated parameters of the microsimulation model of the natural history of CRC.

Parameter	Mean	SD	MAP	95% CrI	
				LB	UB
p_{adeno}	0.264	0.008	0.264	0.248	0.281
p_{small}	0.706	0.019	0.711	0.667	0.741
l	6.24E–06	3.16E–06	4.52E–06	1.92E–06	1.41E–05
γ	2.639	0.112	2.635	2.432	2.877
λ_2	0.035	0.002	0.035	0.031	0.039
λ_3	0.021	0.001	0.021	0.020	0.023
λ_4	0.374	0.036	0.368	0.310	0.448
λ_5	0.247	0.021	0.251	0.209	0.288
λ_6	0.457	0.076	0.435	0.345	0.664

colonoscopy screening vs. no screening for CRC. A separate PSA was performed for the four different approaches to quantify the uncertainty of the two types of parameters. We used EMEWS to distribute the samples of each PSA across HPC resources.

Value of Information Analysis

We quantified the theoretical value of eliminating uncertainty in the external and calibrated model parameters using VOI analysis. VOI measures the losses (i.e., foregone benefits) from choosing a strategy given imperfect information (Raiffa and Schlaifer, 1961), providing the amount of resources a decision maker should be willing to spend to obtain information that would reduce the uncertainty. Specifically, we estimated the value of eliminating parametric uncertainty (i.e., the EVPI) in the cost-effectiveness of a 10-years colonoscopy screening strategy. This entailed computing the difference in net benefit between perfect information and current information (Oostenbrink et al., 2008). The EVPI was calculated across a wide range of willingness-to-pay (WTP) thresholds (Eckermann et al., 2010). We repeated this VOI analysis for the different approaches to characterize the uncertainty of the calibrated and external parameters.

RESULTS

We sampled 5,000 parameter sets from the posterior distribution using IMIS, including 3,241 unique parameter sets with an expected sample size (ESS) of 2,098. With the sample from the posterior distribution, we estimated posterior means and standard deviations, MAP estimates, and 95% credible intervals (CrI) for all calibrated parameters (**Table 3**). The posterior means of the calibrated parameter were similar to the prior means (**Table 3**). Still, the major contrast is that the width of the posterior distributions shrunk, meaning that the calibration targets informed the calibrated parameters through a Bayesian updating (**Figure 2**).

The Bayesian calibration also correlated the parameters, showing the dependency among some of them (**Figure 3**). There are pairs of parameters with high correlation. The scale and shape parameters of the Weibull hazard function for the age of onset of adenomas, l, and γ, respectively, have the highest negative correlation of −0.958. The high correlation results from the calibration of the microsimulation model of the natural history of CRC being non-identifiable when calibrating all 9 parameters to all the targets. The transition rates from early preclinical CRC to late preclinical and early clinical have a correlation of 0.784. The prevalence of adenomas and the proportion of small adenomas at age 50, which inform the initial distribution of

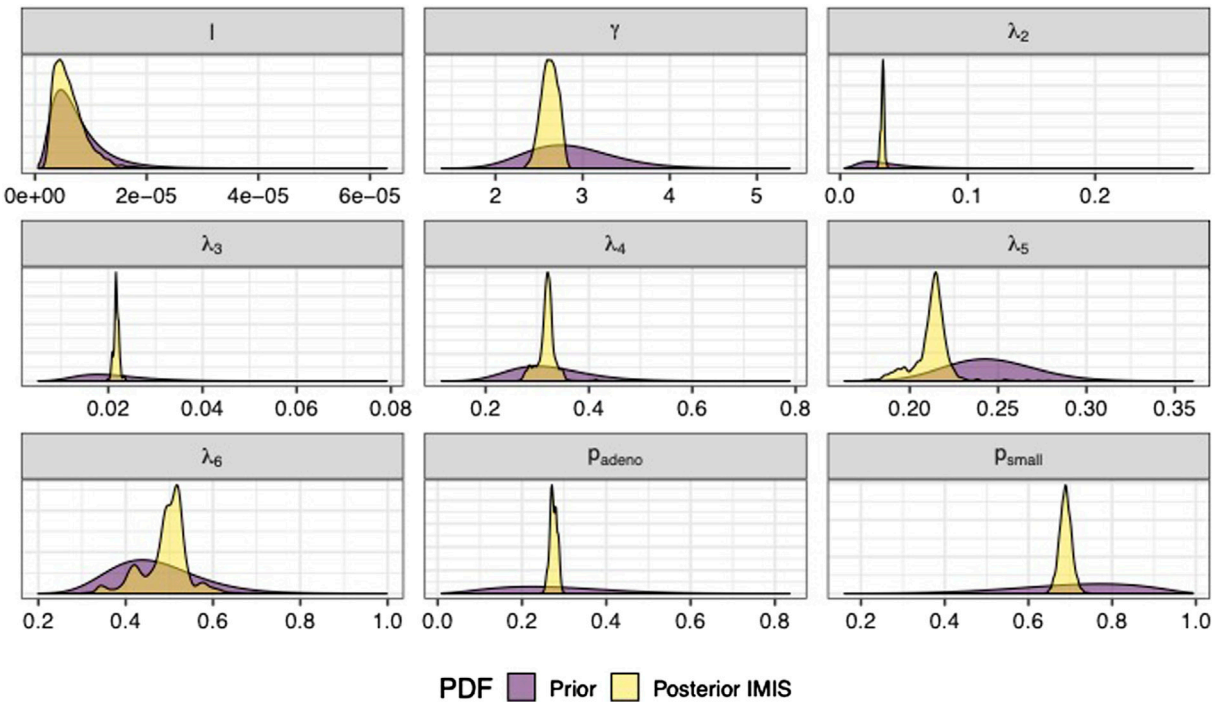


FIGURE 2 | Prior and posterior marginal distributions of calibrated parameters of the microsimulation model of the natural history of CRC.

the cohort across the adenoma health states, also have a high correlation of 0.482. These high correlations result from the model calibration being non-identifiable. In a previous study, we found that the estimation of the 9 parameters of this model structure is non-identifiable via. calibration because the relationship between the parameters is highly colinear when using the current four calibration targets (Alarid-Escudero et al., 2018).

The calibrated model accurately predicted the calibration targets for both the means and the uncertainty intervals. **Figure 4** shows the internal validation of the calibrated model by comparing calibration targets with their 95% confidence interval (CI) and the model-predicted posterior means together with their 95% posterior PI.

The joint distribution of the incremental quality-adjusted life years (QALYs) and incremental costs of the 10 years colonoscopy screening strategy vs. the no-screening strategy resulting from the PSA for the four uncertainty quantification approaches of the calibrated parameters are shown in **Figure 5**. When accounting for the uncertainty on the external parameters, there is little difference in the spread of the CEA outcomes when considering the joint distribution of the calibrated parameters vs. using only the MAP estimates (approaches 1 and 2 on the top row of **Figure 5**, respectively). The joint distribution of the outcomes is slightly wider when considering uncertainty on all parameters compared to when fixing the calibrated parameters at their MAP estimate. The third approach reflects the impact of only varying the calibrated parameters on the joint distribution of incremental QALYS and incremental costs, which is much narrower than approaches 1 and 2. The fourth approach, which characterizes uncertainty of the calibrated parameters using the method of moments without accounting for correlation, has the widest spread on the distribution of the outcomes.

For the VOI analysis, we found value in eliminating uncertainty by having a positive EVPI in the parameters of the CEA of the 10-years colonoscopy screening strategy (**Figure 6**). However, the value varies by uncertainty quantification and WTP threshold. The first and second approaches to uncertainty quantification had similar EVPI, reaching their maximum of $653 and $685, respectively, at a $66,000/QALY WTP threshold. For WTP thresholds greater than $66,000/QALY, the first approach had a higher EVPI than the second approach. When we consider only the uncertainty for the calibrated parameters (approach 3), the EVPI is the lowest across all WTP thresholds with an EVPI of $0.1 at a WTP threshold of $66,000/QALY and reaching its highest of $212 at a WTP threshold of $71,000/QALY. The fourth approach reaches a maximum of $809 at a WTP threshold of $66,000/QALY and is the highest compared to the other approaches up to a WTP threshold of $81,000/QALY, at which the first approach has the highest EVPI.

DISCUSSION

In this study, we characterized the uncertainty of a realistic microsimulation model of the natural history of CRC by calibrating its parameters to different targets with varying degrees of uncertainty using a Bayesian approach on an HPC environment using EMEWS. We also quantified the value of the uncertainty of the

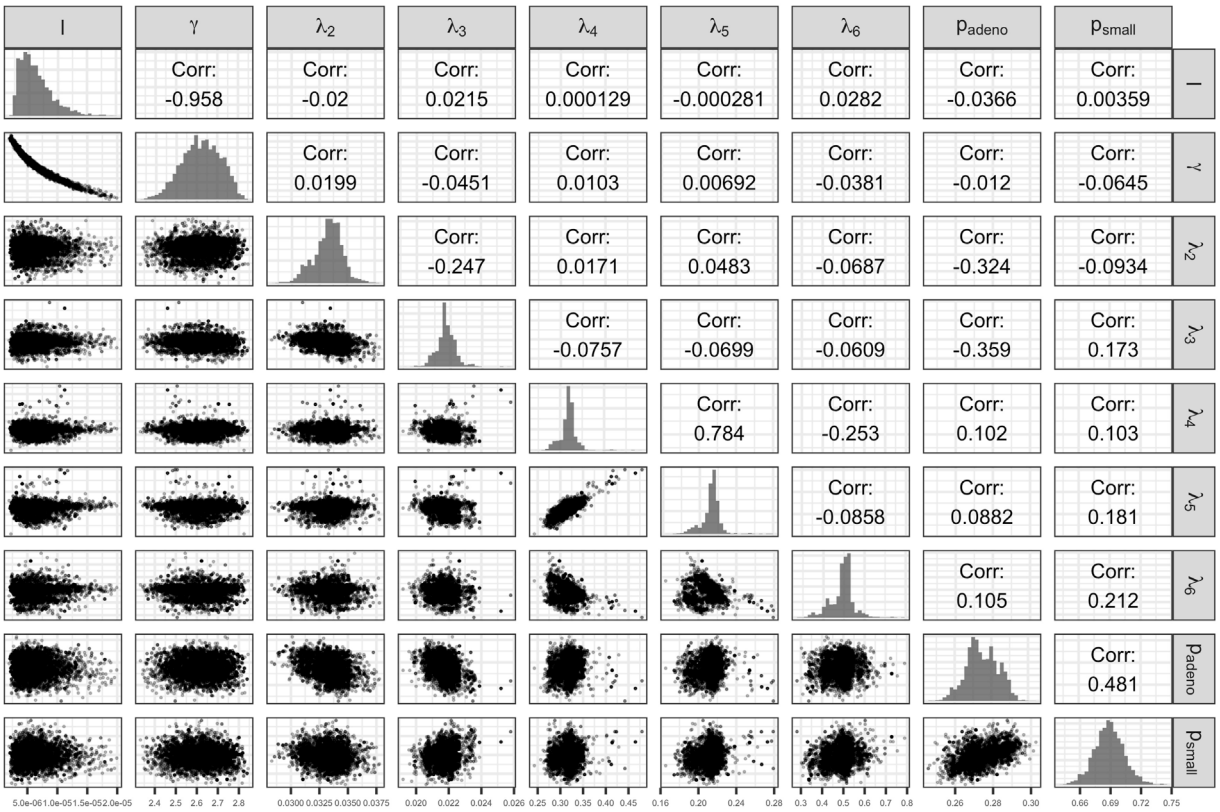


FIGURE 3 | Scatter plot of pairs of deep model parameters with correlation coefficient and posterior marginal distributions.

calibrated parameters on the cost-effectiveness of a 10-year colonoscopy screening strategy with a VOI analysis. EMEWS has been previously used to calibrate other microsimulation DMs (Rutter et al., 2019; Rutter et al., 2019) but has not been previously used to conduct a PSA with the calibrated parameters and calculate the VOI. Although Bayesian calibration can be a computationally intensive task, we reduce the computation time by evaluating the likelihood of different parameter sets in multiple cores simultaneously on an HPC setup, which IMIS allows.

We found that different characterizations of the uncertainty of calibrated parameters affect the expected value of reducing uncertainty on the CEA. Ignoring inherent correlation among calibrated parameters on a PSA overestimates the value of uncertainty. When the full posterior distribution of the calibrated parameters is not readily available, the MAP could be considered the best parameter set. In our example, not considering the uncertainty of calibrated parameters on the PSA did not seem to have a meaningful impact on the uncertainty of the CEA outcomes and the EVPI of the screening strategy. The uncertainty associated with the natural history was less valuable than the uncertainty of the external parameters. However, these results should be taken with caution because this analysis is conducted on a fictitious model with simulated calibrated targets. Modelers should analyze the impact of a well-conducted characterization of the uncertainty of calibrated parameters on CEA outcomes and VOI measures on a case-by-case basis.

There are examples of calibrated parameters being included in a PSA. For instance, by taking a certain number of good-fitting parameter sets (Kim et al., 2007; Kim et al., 2009), bootstrapping with equal probability good-fitting parameter sets obtained through directed search algorithms (e.g., Nelder-Mead) (Taylor et al., 2012), or conducting a Bayesian calibration, which produces the joint posterior distribution of the calibrated parameters (Menzies et al., 2017). However, this is the first manuscript to conduct a PSA and VOI analysis using distributions of calibrated microsimulation DM parameters that accurately characterize their uncertainty.

Currently, Bayesian calibration of microsimulation DMs might not be feasible on regular desktops or laptops. To circumvent current computational limitations from using Bayesian methods in calibrating microsimulation models, surrogate models -often called metamodels or emulators-have been proposed (O'Hagan et al., 1999; O'Hagan, 2006; Oakley and Youngman, 2017). Surrogate models are statistical models like Gaussian processes (Sacks et al., 1989a; Sacks et al., 1989b; Oakley and O'Hagan, 2002) or neural networks (Hauser et al., 2012; Jalal et al., 2021) that aim to replace the relationship between inputs and outputs of the original microsimulation DM (Barton et al., 1992; Kleijnen, 2015), which, once fitted, are computationally more efficient to run than the microsimulation DM. Constructing an emulator might not be a straightforward task because the microsimulation DM still needs to be evaluated at different

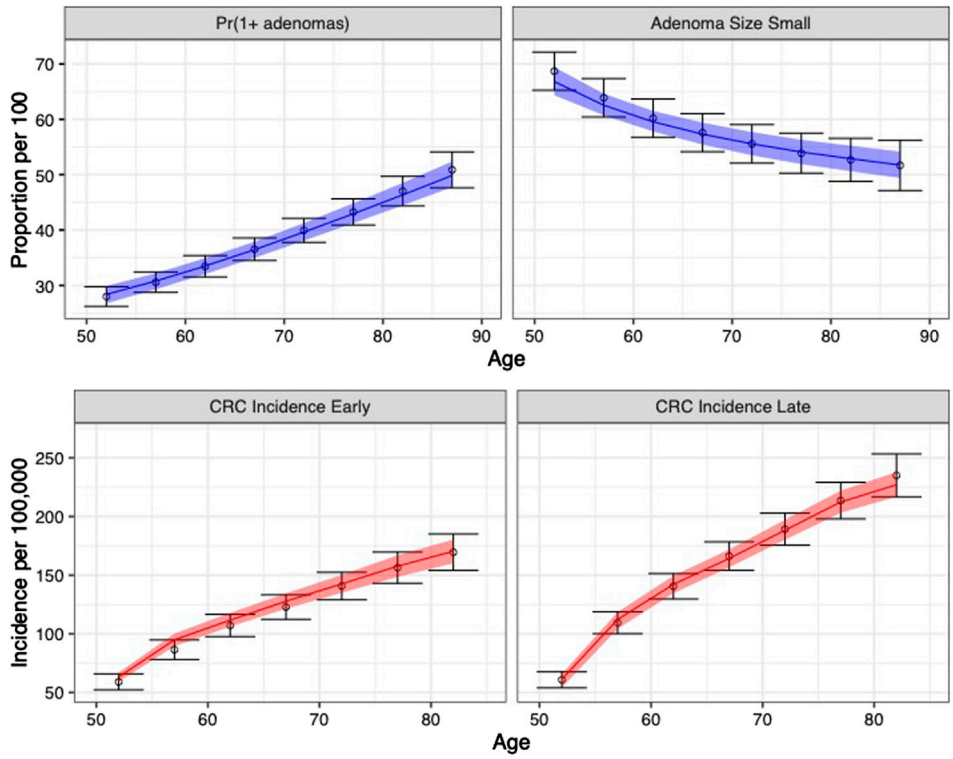


FIGURE 4 | Comparison between posterior model-predicted outputs and calibration targets. Calibration targets with their 95% CI are shown in black. The shaded area shows the 95% posterior model-predictive interval of the outcomes and colored lines shows the posterior model-predicted mean based on 5,000 simulations using samples from the posterior distribution. Upper panel refers to adenoma-related targets and lower panel refers to CRC incidence targets by stage.

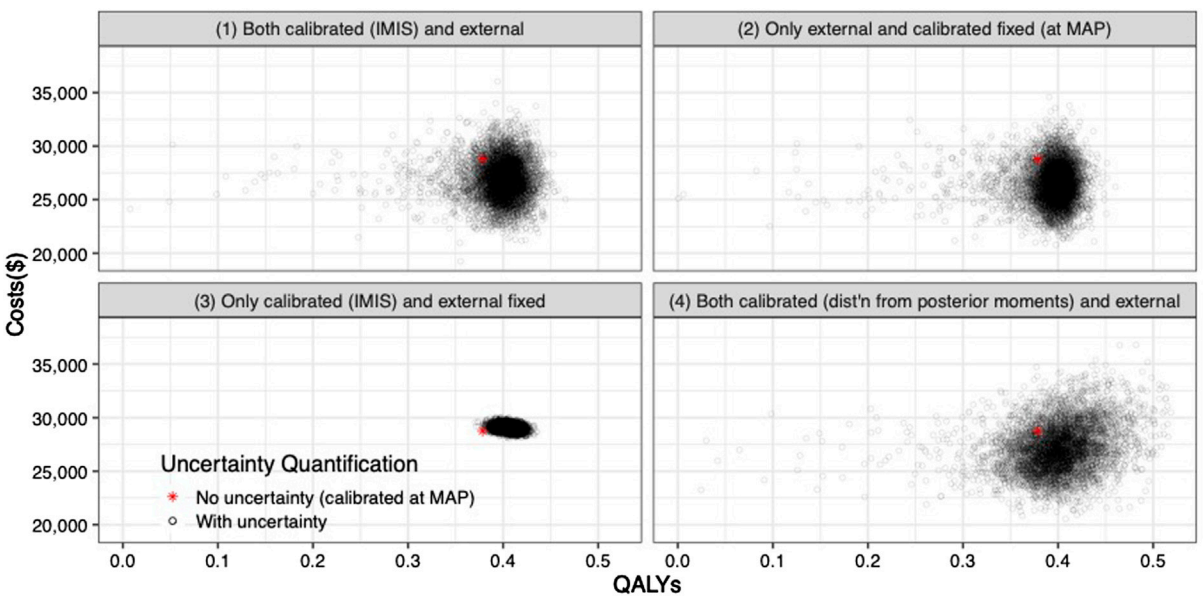


FIGURE 5 | Incremental costs and incremental QALYs of 10-years colonoscopy screening vs. no screening under different assumptions of characterization of the uncertainty of both calibrated and external parameters. The red star corresponds to the incremental costs and incremental QALYs evaluated at the maximum-a-posteriori estimate of the calibrated parameters and the mean values of the external parameters.

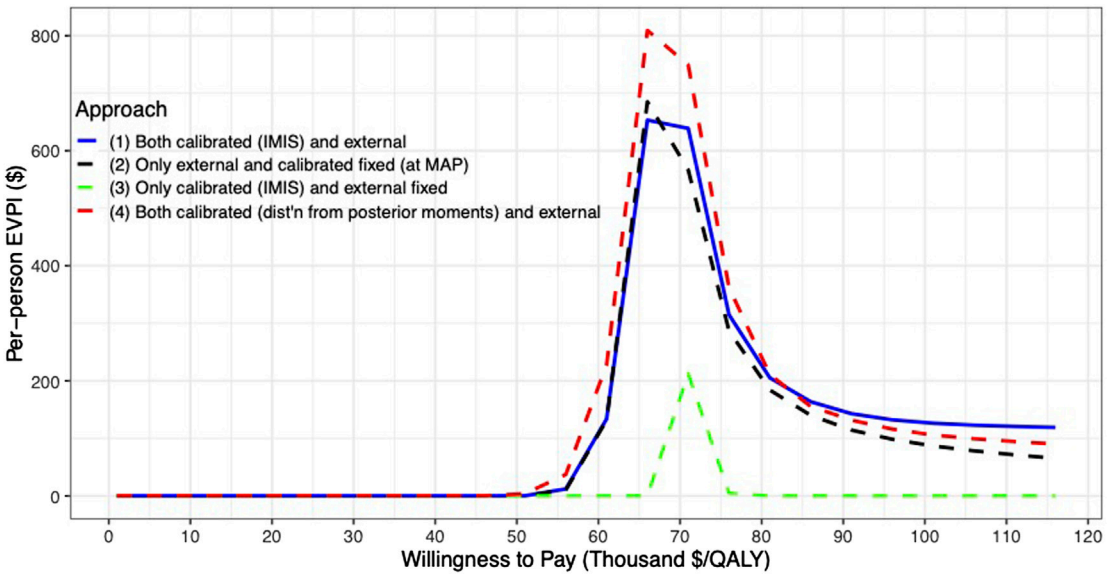


FIGURE 6 | Per-patient EVPI of 10-year colonoscopy screening vs no screening under different approaches to characterize the uncertainty of both the calibrated and external parameters.

parameter sets, which could also be computationally expensive. Furthermore, the statistical routines to build the emulator may not be readily available in the programming language in which the microsimulation DM is coded. These are situations where EMEWS can be used to construct metamodels efficiently; however, this is a topic for further research.

Researchers might actively avoid questions that would require HPC due to the perceived difficulties involved or make do with less-than-ideal smaller-scale analyses (e.g., choosing the maximum likelihood estimate or a small set of parameters instead of the posterior distribution for uncertainty quantification) and the robustness of the conclusions can suffer as a result.

In this article, we showed that EMEWS could facilitate the use of HPC to implement computationally demanding Bayesian calibration routines to correctly characterize the uncertainty of the calibrated parameters of microsimulation DMs and propagate it in the evaluation of CEA of screening strategies and quantify their value of information. This study's methodology and results could guide a similar VOI analysis on CEAs using microsimulation DMs to determine where more research is needed and guide research prioritization.

DATA AVAILABILITY STATEMENT

The raw data supporting the conclusion of this article will be made available by the authors, without undue reservation.

AUTHOR CONTRIBUTIONS

FA-E contributed to the conceptualization, data curation, formal analysis, funding acquisition, investigation, methodology, project administration, resources, software, supervision, validation, visualization, and writing of the original draft and of reviewing and editing; AK contributed to the conceptualization, formal analysis, funding acquisition, investigation, methodology, resources, validation, and writing of the original draft and of reviewing and editing; JO contributed to the conceptualization, formal analysis, funding acquisition, investigation, methodology, resources, validation, and writing of the original draft and of reviewing and editing; NC contributed to the conceptualization, formal analysis, funding acquisition, investigation, methodology, resources, validation, and writing of the original draft and of reviewing and editing; KK contributed to the conceptualization, formal analysis, funding acquisition, investigation, methodology, resources, validation, and writing of the original draft and of reviewing and editing.

FUNDING

Financial support for this study was provided in part by a grant from the National Council of Science and Technology of Mexico (CONACYT) and a Doctoral Dissertation Fellowship from the Graduate School of the University of Minnesota as part of Dr. Alarid-Escudero's doctoral program. All authors were supported by grants from the National Cancer Institute (U01- CA-199335 and U01-CA-253913) as part of the Cancer Intervention and Surveillance Modeling Network (CISNET). The work was supported in part by the U.S. Department of Energy, Office of Science, under contract (No. DE- AC0206CH11357). The funding agencies had no role in the study's design, interpretation of results, or

writing of the manuscript. The content is solely the authors' responsibility and does not necessarily represent the official views of the National Institutes of Health. The funding agreement ensured the authors' independence in designing the study, interpreting the data, writing, and publishing the report. This research was completed with resources provided by the Research Computing Center at the University of Chicago (Midway2 cluster).

REFERENCES

Alarid-Escudero, F., MacLehose, R. F., Peralta, Y., Kuntz, K. M., and Enns, E. A. (2018). Nonidentifiability in Model Calibration and Implications for Medical Decision Making. *Med. Decis. Mak [Internet]* 38 (7), 810–821. doi:10.1177/0272989x18792283

Arias, E., Heron, M., and Xu, J. (2017). United States Life Tables. *Natl. Vital Stat. Rep.* 66 (4), 63.

Barton, R. R. (1992). "Metamodels for Simulation Input-Output Relations," in Winter Simulation Conference. Editors J. J. Swain, D. Goldsman, R. C. Crain, and J. R. Wilson, 289–299. doi:10.1145/167293.167352

Briggs, A. H., Goeree, R., Blackhouse, G., and O'Brien, B. J. (2002). Probabilistic Analysis of Cost-Effectiveness Models: Choosing between Treatment Strategies for Gastroesophageal Reflux Disease. *Med. Decis. Mak* 22, 290–308. doi:10.1177/027298902400448867

Briggs, A. H., Weinstein, M. C., Fenwick, E. A. L., Karnon, J., Sculpher, M. J., and Paltiel, A. D. (2012). Model Parameter Estimation and Uncertainty Analysis. *Med. Decis. Making* 32 (5), 722–732. doi:10.1177/0272989x12458348

Cox, D. R., and Miller, H. D. (1965). *The Theory of Stochastic Processes*. London, UK: Chapman & Hall.

DeYoreo, M., Rutter, C. M., Ozik, J., and Collier, N. (2022). Sequentially Calibrating a Bayesian Microsimulation Model to Incorporate New Information and Assumptions. *BMC Med. Inform. Decis. Mak [Internet]. Biomed. Cent.* 22 (1), 12. doi:10.1186/s12911-021-01726-0

Eckermann, S., Karnon, J., and Willan, A. R. (2010). The Value of Value of Information: Best Informing Research Design and Prioritization Using Current Methods. *Pharmacoeconomics* 28 (9), 699–709. doi:10.2165/11537370-000000000-00000

Enns, E. A., Kao, S. Y., Kozhimannil, K. B., Kahn, J., Farris, J., and Kulasingam, S. L. (2017). Using Multiple Outcomes of Sexual Behavior to Provide Insights into Chlamydia Transmission and the Effectiveness of Prevention Interventions in Adolescents. *Sex. Transm. Dis.* 44 (10), 619–626. doi:10.1097/olq.0000000000000653

Frazier, A. L., Colditz, G. A., Fuchs, C. S., and Kuntz, K. M. (2000). Cost-effectiveness of Screening for Colorectal Cancer in the General Population. *JAMA* 284 (15), 1954–1961. doi:10.1001/jama.284.15.1954

Gustafson, P. (2015). *Bayeisan Inference for Partially Identified Models: Exploring the Limits of Limited Data*. Boca Raton, FL: CRC Press.

Gustafson, P. (2005). On Model Expansion, Model Contraction, Identifiability and Prior Information: Two Illustrative Scenarios Involving Mismeasured Variables. *Stat. Sci. [Internet]* 20 (2), 111–129. doi:10.1214/088342305000000098

Hauser, T., Keats, A., and Tarasov, L. (2012). Artificial Neural Network Assisted Bayesian Calibration of Climate Models. *Clim. Dyn.* 39 (1–2), 137–154. doi:10.1007/s00382-011-1168-0

Hawkins-Daarud, A., Prudhomme, S., van der Zee, K. G., and Oden, J. T. (2013). Bayesian Calibration, Validation, and Uncertainty Quantification of Diffuse Interface Models of Tumor Growth. *J. Math. Biol.* 67 (6–7), 1457–1485. doi:10.1007/s00285-012-0595-9

Iskandar, R. (2018). A Theoretical Foundation of State-Transition Cohort Models in Health Decision Analysis. *PLoS One* 13 (12), e0205543. doi:10.1371/journal.pone.0205543

Jackson, C. S., Oman, M., Patel, A. M., and Vega, K. J. (2016). Health Disparities in Colorectal Cancer Among Racial and Ethnic Minorities in the United States. *J. Gastrointest. Oncol.* 7 (Suppl. 1), 32–43. doi:10.3978/j.issn.2078-6891.2015.039

Jalal, H., Trikalinos, T. A., and Alarid-Escudero, F. (2021). BayCANN: Streamlining Bayesian Calibration with Artificial Neural Network Metamodeling. *Front. Physiol.* 12, 1–13. doi:10.3389/fphys.2021.662314

Kaipio, J., and Somersalo, E. (2005). *Statistical and Computational Inverse Problems*. New York, NY: Springer.

Karnon, J., Goyder, E., Tappenden, P., McPhie, S., Towers, I., Brazier, J., et al. (2007). A Review and Critique of Modelling in Prioritising and Designing Screening Programmes. *Health Technol. Assess. (Rockv).* 11 (52), 1–145. doi:10.3310/hta11520

Kennedy, M. C., and O'Hagan, A. (2001). Bayesian Calibration of Computer Models. *J. R. Stat Soc Ser B Stat. Methodol.* 63 (3), 425–464. doi:10.1111/1467-9868.00294

Kim, J. J., Kuntz, K. M., Stout, N. K., Mahmud, S., Villa, L. L., Franco, E. L., et al. (2007). Multiparameter Calibration of a Natural History Model of Cervical Cancer. *Am. J. Epidemiol.* 166 (2), 137–150. doi:10.1093/aje/kwm086

Kim, J. J. P., Ortendahl, B. S. J., and Goldie, S. J. M. (2009). Cost-Effectiveness of Human Papillomavirus Vaccination and Cervical Cancer Screening in Women Older Than 30 Years in the United States. *Ann. Intern. Med.* 151 (8), 538–545. doi:10.7326/0003-4819-151-8-200910200-00007

Kleijnen, J. P. C. (2015). *Design and Analysis of Simulation Experiments*. 2nd ed. New York, NY: Springer International Publishing.

Kolmogorov, A. N. (1963). On the Representation of Continuous Functions of Several Variables by Superposition of Continuous Functions of One Variable and Addition. *Am. Math. Soc. Transl Ser.* 28, 55–59. doi:10.1090/trans2/028/04

Kong, C. Y., McMahon, P. M., and Gazelle, G. S. (2009). Calibration of Disease Simulation Model Using an Engineering Approach. *Value Heal* 12 (4), 521–529. doi:10.1111/j.1524-4733.2008.00484.x

Krijkamp, E. M., Alarid-Escudero, F., Enns, E. A., Jalal, H. J., Hunink, M. G. M., and Pechlivanoglou, P. (2018). Microsimulation Modeling for Health Decision Sciences Using R: A Tutorial. *Med. Decis. Mak [Internet]* 38 (3), 400–422. doi:10.1177/0272989x18754513

Kuntz, K. M., Lansdorp-Vogelaar, I., Rutter, C. M., Knudsen, A. B., van Ballegooijen, M., Savarino, J. E., et al. (2011). A Systematic Comparison of Microsimulation Models of Colorectal Cancer: the Role of Assumptions about Adenoma Progression. *Med. Decis. Mak* 31, 530–539. doi:10.1177/0272989x11408730

Kuntz, K. M., Russell, L. B., Owens, D. K., Sanders, G. D., Trikalinos, T. A., and Salomon, J. A. (2017). "Decision Models in Cost-Effectiveness Analysis," in *Cost-Effectiveness in Health and Medicine*. Editors P. J. Neumann, G. D. Sanders, L. B. Russell, J. E. Siegel, and T. G. Ganiats (New York, NY: Oxford University Press), 105–136.

Menzies, N. A., Soeteman, D. I., Pandya, A., and Kim, J. J. (2017). Bayesian Methods for Calibrating Health Policy Models: A Tutorial. *Pharmacoeconomics* 25 (6), 613–624. doi:10.1007/s40273-017-0494-4

Nelder, J. A., and Mead, R. (1965). A Simplex Method for Function Minimization. *Comput. J.* 7 (4), 308–313. doi:10.1093/comjnl/7.4.308

Ness, R. M., Holmes, A. M., Klein, R., and Dittus, R. (1999). Utility Valuations for Outcome States of Colorectal Cancer. *Am. J. Gastroenterol.* 94 (6), 1650–1657. doi:10.1111/j.1572-0241.1999.01157.x

Oakley, J., and O'Hagan, A. (2002). Bayesian Inference for the Uncertainty Distribution of Computer Model Outputs. *Biometrika* 89 (4), 769–784. doi:10.1093/biomet/89.4.769

Oakley, J., and O'Hagan, A. (2004). Probabilistic Sensitivity Analysis of Complex Models: a Bayesian Approach. *J. R. Stat. Soc. Ser. B (Statistical Methodol. [Internet]* 66 (3), 751–769. doi:10.1111/j.1467-9868.2004.05304.x

Oakley, J. E., and Youngman, B. D. (2017). Calibration of Stochastic Computer Simulators Using Likelihood Emulation. *Technometrics* 59 (1), 80–92. doi:10.1080/00401706.2015.1125391

Oden, T., Moser, R., and Ghattas, O. (2010). *Computer Predictions with Quantified Uncertainty*. Austin. TX: The Institute for Computational Engineering and Sciences.

O'Hagan, A. (2006). Bayesian Analysis of Computer Code Outputs: A Tutorial. *Reliab Eng. Syst. Saf. [Internet]* 91 (10–11), 1290–1300. doi:10.1016/j.ress.2005.11.025

O'Hagan, A., Kennedy, M. C., and Oakley, J. E. (1999). Uncertainty Analysis and Other Inference Tools for Complex Computer Codes. *Bayesian Staistics 6:*

Proceedings of the Sixth Valencia International Meeting. Editors J. M. Bernardo, J. O. Berger, A. P. Dawid, and A. F. M. Smith. Oxford, United Kingdom: Clarendon Press, 503–24.

Oostenbrink, J. B., Al, M. J., Oppe, M., and Rutten-van Mölken, M. P. M. H. (2008). Expected Value of Perfect Information: An Empirical Example of Reducing Decision Uncertainty by Conducting Additional Research. Value Heal [Internet]. *Int. Soc. Pharmacoeconomics Outcomes Res. (Ispor)* 11 (7), 1070–1080. doi:10.1111/j.1524-4733.2008.00389.x

Ozik, J., Collier, N. T., Wozniak, J. M., Macal, C. M., and An, G. (2018). Extreme-scale Dynamic Exploration of a Distributed Agent-Based Model with the EMEWS Framework. *IEEE Trans. Comput. Soc. Syst. IEEE* 5 (3), 884–895. doi:10.1109/tcss.2018.2859189

Ozik, J., Collier, N. T., Wozniak, J. M., and Spagnuolo, C. (2016a). "From Desktop to Large-Scale Model Exploration with Swift/T," in Proceedings of the 2016 Winter Simulation Conference. Editors T. M. K. Roeder, P. I. Frazier, R. Szechtman, E. Zhou, T. Huschka, and S. E. Chick (IEEE Press), 206–220. doi:10.1109/wsc.2016.7822090

Raftery, A. E., and Bao, L. (2009). Estimating and Projecting Trends in HIV/AIDS Generalized Epidemics Using Incremental Mixture Importance Sampling. *Biometrics* 66, 1162–1173. doi:10.1111/j.1541-0420.2010.01399.x

Raiffa, H., and Schlaifer, R. O. (1961). *Applied Statistical Decision Theory*. Cambridge, MA: Harvard Business School.

Romanowicz, R. J., Beven, K. J., and Tawn, J. A. (1994). Evaluation of Predictive Uncertainty in Nonlinear Hydrological Models Using a Bayesian Approach. *Stat. Environ. 2 Water Relat. Issues*. Editors V. Barnett and K. F. Turkman. Hoboken, NJ: John Wiley & Sons (February), 297–317.

Rutter, C. M., Ozik, J., DeYoreo, M., and Collier, N. (2019). Microsimulation Model Calibration Using Incremental Mixture Approximate Bayesian Computation. *Ann. Appl. Stat. [Internet* 13 (4), 2189–2212. doi:10.1214/19-aoas1279

Rutter, C. M., Yu, O., and Miglioretti, D. L. (2007). A Hierarchical Non-homogenous Poisson Model for Meta-Analysis of Adenoma Counts. *Stat. Med.* 26 (1), 98–109. doi:10.1002/sim.2460

Rutter, C. M., Miglioretti, D. L., and Savarino, J. E. (2009). Bayesian Calibration of Microsimulation Models. *J. Am. Stat. Assoc.* 104 (488), 1338–1350. doi:10.1198/jasa.2009.ap07466

Rutter, C. M., Zaslavsky, A. M., and Feuer, E. J. (2011). Dynamic Microsimulation Models for Health Outcomes. *Med. Decis. Making* 31 (1), 10–18. doi:10.1177/0272989x10369005

Ryckman, T., Luby, S., Owens, D. K., Bendavid, E., and Goldhaber-Fiebert, J. D. (2020). Methods for Model Calibration under High Uncertainty: Modeling Cholera in Bangladesh. *Med. Decis. Mak* 40 (5), 693–709. doi:10.1177/0272989x20938683

Sacks, J., Schiller, S. B., and Welch, W. J. (1989). Designs for Computer Experiments. *Technometrics* 31 (1), 41–47. doi:10.1080/00401706.1989.10488474

Sacks, J., Welch, W. J., Mitchell, T. J., and Wynn, H. P. (1989). Design and Analysis of Computer Experiments. *Stat. Sci.* 4 (4), 409–423. doi:10.1214/ss/1177012413

Schlaifer, R. O. (1959). *Probability and Statistics for Business Decisions*. New York, NY: McGraw-Hill.

Schroy, P. C., III, Coe, A., Chen, C. A., O'Brien, M. J., and Heeren, T. C. (2013). Prevalence of Advanced Colorectal Neoplasia in White and Black Patients Undergoing Screening Colonoscopy in a Safety-Net Hospital. *Ann. Intern. Med.* 159 (1), 13. doi:10.7326/0003-4819-159-1-201307020-00004

Sculpher, M. J., Basu, A., Kuntz, K. M., and Meltzer, D. O. (2017). "Reflecting Uncertainty in Cost-Effectiveness Analysis," in *Cost-Effectiveness in Health and Medicine*. Editors PJ Neumann, GD. Sanders, L. B. Russell, J. E. Siegel, and T. G. Ganiats (New York, NY: Oxford University Press), 289–318.

Steele, R. J., Raftery, A. E., and Emond, M. J. (2006). Computing Normalizing Constants for Finite Mixture Models via Incremental Mixture Importance Sampling (IMIS). *J. Comput. Graph Stat.* 15 (3), 712–734. doi:10.1198/106186006x132358

Stout, N. K., Knudsen, A. B., Kong, C. Y., Mcmahon, P. M., and Gazelle, G. S. (2009). Calibration Methods Used in Cancer Simulation Models and Suggested Reporting Guidelines. *Pharmacoeconomics* 27 (7), 533–545. doi:10.2165/11314830-000000000-00000

Taylor, D. C., Pawar, V., Kruzikas, D. T., Gilmore, K. E., Sanon, M., and Weinstein, M. C. (2012). Incorporating Calibrated Model Parameters into Sensitivity Analyses: Deterministic and Probabilistic Approaches. *Pharmacoeconomics* 30 (2), 119–126. doi:10.2165/11593360-000000000-00000

Van Rijn, J. C., Reitsma, J. B., Stoker, J., Bossuyt, P. M., Van Deventer, S. J., and Dekker, E. (2006). Polyp Miss Rate Determined by Tandem Colonoscopy: A Systematic Review. *Am. J. Gastroenterol.* 101 (2), 343–350. doi:10.1111/j.1572-0241.2006.00390.x

Welton, N. J., and Ades, A. E. (2005). Estimation of Markov Chain Transition Probabilities and Rates from Fully and Partially Observed Data: Uncertainty Propagation, Evidence Synthesis, and Model Calibration. *Med. Decis. Making* 25, 633–645. doi:10.1177/0272989x05282637

Whyte, S., Walsh, C., and Chilcott, J. (2011). Bayesian Calibration of a Natural History Model with Application to a Population Model for Colorectal Cancer. *Med. Decis. Mak* 31 (4), 625–641. doi:10.1177/0272989x10384738

Wu, G. H-M., Wang, Y-M., Yen, A. M-F., Wong, J.-M., Lai, H.-C., Warwick, J., et al. (2006). Cost-effectiveness Analysis of Colorectal Cancer Screening with Stool DNA Testing in Intermediate-Incidence Countries. *BMC Cancer* 6, 136. doi:10.1186/1471-2407-6-136

Conflict of Interest: The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

Publisher's Note: All claims expressed in this article are solely those of the authors and do not necessarily represent those of their affiliated organizations, or those of the publisher, the editors and the reviewers. Any product that may be evaluated in this article, or claim that may be made by its manufacturer, is not guaranteed or endorsed by the publisher.

Advantages of publishing in Frontiers

OPEN ACCESS
Articles are free to read for greatest visibility and readership

FAST PUBLICATION
Around 90 days from submission to decision

HIGH QUALITY PEER-REVIEW
Rigorous, collaborative, and constructive peer-review

TRANSPARENT PEER-REVIEW
Editors and reviewers acknowledged by name on published articles

Frontiers
Avenue du Tribunal-Fédéral 34
1005 Lausanne | Switzerland

Visit us: www.frontiersin.org
Contact us: frontiersin.org/about/contact

REPRODUCIBILITY OF RESEARCH
Support open data and methods to enhance research reproducibility

DIGITAL PUBLISHING
Articles designed for optimal readership across devices

FOLLOW US
@frontiersin

IMPACT METRICS
Advanced article metrics track visibility across digital media

EXTENSIVE PROMOTION
Marketing and promotion of impactful research

LOOP RESEARCH NETWORK
Our network increases your article's readership