frontiers | Research Topics

Data science and digital service delivery in healthcare

Edited by
Koichi Fujiwara, Tetsuharu Nagamoto and Priit Kruus

Published in
Frontiers in Public Health
Frontiers in Computer Science
Frontiers in Neuroinformatics
Frontiers in Big Data

ISSN 1664-8714
ISBN 978-2-83251-374-3
DOI 10.3389/978-2-83251-374-3

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

Data science and digital service delivery in healthcare

Topic editors
Koichi Fujiwara — Nagoya University, Japan
Tetsuharu Nagamoto — Kyoto University, Japan
Priit Kruus — Tallinn University of Technology, Estonia

Citation
Fujiwara, K., Nagamoto, T., Kruus, P., eds. (2023). *Data science and digital service delivery in healthcare*. Lausanne: Frontiers Media SA.
doi: 10.3389/978-2-83251-374-3

Table of contents

frontiers | Frontiers in Computer Science

TYPE Editorial
PUBLISHED 05 January 2023
DOI 10.3389/fcomp.2022.1079583

OPEN ACCESS

EDITED AND REVIEWED BY
Patty Kostkova,
University College London,
United Kingdom

*CORRESPONDENCE
Koichi Fujiwara
✉ fujiwara.koichi@hps.material.nagoya-u.ac.jp

SPECIALTY SECTION
This article was submitted to
Digital Public Health,
a section of the journal
Frontiers in Computer Science

RECEIVED 25 October 2022
ACCEPTED 21 December 2022
PUBLISHED 05 January 2023

CITATION
Fujiwara K (2023) Editorial: Data science and digital service delivery in healthcare.
Front. Comput. Sci. 4:1079583.
doi: 10.3389/fcomp.2022.1079583

Editorial: Data science and digital service delivery in healthcare

Koichi Fujiwara*

Department of Material Process Engineering, Nagoya University, Nagoya, Japan

KEYWORDS

biological signal, medical AI, machine learning, data science, wearable sensors

Editorial on the Research Topic
Data science and digital service delivery in healthcare

The applications of machine learning and AI technologies in the medical and healthcare fields have advanced with deep learning technologies like other AI fields. Most real-world applications of medical AI technologies have remained diagnostic systems based on medical images obtained from computer technology (CT) and magnetic resonance imaging (MRI). On the other hand, the use of biological signals currently has been left behind.

Various biological signals have been used for testing in hospitals, for example, Electrocardiograms (ECG) for cardiovascular tastings, Electroencephalograms (EEG) for epilepsy diagnosis and sleep testing, Electromyography (EMG), Electrooculography (EOG), a saturation of percutaneous oxygen (SpO_2) and respiratory signals for sleep testing. In addition, acceleration signals have been used for monitoring daily activities. However, these biological signals have not been fully utilized for medical AI development.

This may be because: (1) the Spatio-temporal patterns of the biological signals are usually complex non-stationary, and high-dimensional. In EEG, around 20 electrodes are used for measurement according to the 10–20 International system. Thus, it is difficult to specify the expression of the target physiological phenomena in the biological signals. (2) The amount of obtained biological signals in hospitals is limited because the biological signal measurement is usually burdensome. In many studies, we need to perform additional tests to collect enough biological signals for AI development. (3) Even when a large number of biological signals is obtained, the occurrence frequency of the target phenomenon is still low, which results in highly imbalanced data. For example, when we focus on epileptic seizures and collect EEG data from epileptic patients, obtaining sufficient EEG data around seizure occurrences is difficult because most patients have seizures once or twice a day or week. (4) Various artifacts are easily contaminated in biological signals, such as motion artifacts or electrical noise from power supplies. We have to remove such artifacts before analysis appropriately. (5) Even when such artifacts can be removed appropriately, the signal-to-noise ratio (SNR) of the biological signal is not always high, and (6) Individuality among people is considerable, and generalization among them is difficult. Therefore, medical AI based on biological signals contains all of the difficulties in machine learning.

In our Research Topic, the published articles presented various AI and machine learning applications for healthcare services, which may suggest ideas on solutions to these problems. We hope that many researchers in machine learning will enter this area soon.

Author contributions

The author confirms being the sole contributor of this work and has approved it for publication.

Conflict of interest

The author declares 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 Computer Science

ORIGINAL RESEARCH
published: 04 February 2020
doi: 10.3389/fcomp.2020.00003

Data-Efficient Framework for Personalized Physiotherapy Feedback

Bryan Lao[1], *Tomoya Tamei*[2] *and Kazushi Ikeda*[1]*

[1] *Mathematical Informatics Laboratory, Graduate School of Information Science, Nara Institute of Science and Technology, Ikoma, Japan,* [2] *Center for Mathematical and Data Sciences, Kobe University, Kobe, Japan*

Physiotherapy is a labor-intensive process that has become increasingly inaccessible. Existing telehealth solutions overcome many of the logistical problems, but they are cumbersome to re-calibrate for the various exercises involved. To facilitate self-exercise efficiently, we developed a framework for personalized physiotherapy exercises. Our approach eliminates the need to re-calibrate for different exercises, using only few user-specific demonstrations available during collocated therapy. Two types of augmented feedback are available to the user for self-correction. The framework's utility was demonstrated for the sit-to-stand task, an important activity of daily living. Although further testing is necessary, our results suggest that the framework can be generalized to the learning of arbitrary motor behaviors.

Keywords: telehealth, physiotherapy, Gaussian process, latent variable model, sit-to-stand

OPEN ACCESS

Edited by:
Koichi Fujiwara,
Nagoya University, Japan

Reviewed by:
Jun-Wei Wang,
University of Science and Technology Beijing, China
Emi A. Yuda,
Tohoku University, Japan
Yukio Horiguchi,
Kyoto University, Japan

****Correspondence:***
Kazushi Ikeda
kazushi@is.naist.jp

Specialty section:
This article was submitted to Digital Public Health, a section of the journal Frontiers in Computer Science

Received: *08 November 2019*
Accepted: *14 January 2020*
Published: *04 February 2020*

Citation:
Lao B, Tamei T and Ikeda K (2020) Data-Efficient Framework for Personalized Physiotherapy Feedback. Front. Comput. Sci. 2:3. doi: 10.3389/fcomp.2020.00003

1. INTRODUCTION

Physiotherapy is a rehabilitation activity that improves and restores physical function. The process can be labor-intensive, involving multiple face-to-face sessions with a physical therapist (PT) (**Figure 1**). In each therapy session, the PT and patient practice a large variety of movements with limited amount of time. An average post-stroke therapy session was found to be 36 min long, requiring patients to perform up to 17 types of movements (Lang et al., 2007). This amount of practice is an order of magnitude lower than what is expected to induce neural reorganization (Lang et al., 2009). Thus, the patient is required to continue the exercises themselves, without the corrective guidance from the PT (Tang et al., 2015).

While conventional face-to-face therapy is effective in treating many common injuries, access has become increasingly difficult for many individuals. Among many factors, the shortage of practitioners and physical distance were identified as major contributors (Schopp et al., 2000). This shortage is expected to worsen as the world population is aging at an unprecedented rate. At present, so-called developed countries are already considered aged, led by Japan (28%) and Italy (23%); developing countries are following this trend at an even faster pace (United Nations, 2019). As a result, countries are shifting long-term elderly care from institutions to home- and community-based services, and remote therapy has emerged as an accessible alternative to conventional therapy (Higo and Khan, 2015).

Remote therapy, or telehealth, is defined as the delivery of health-related services and information via telecommunications technologies (ICT) (Lee et al., 2018). Recent developments use immersive technologies, like augmented reality (AR) and virtual reality (VR) to simulate environments close to conventional therapy. An AR example is the popular augmented mirror setup that guides users through pre-recorded exercises. For example, Physio@Home demonstrates

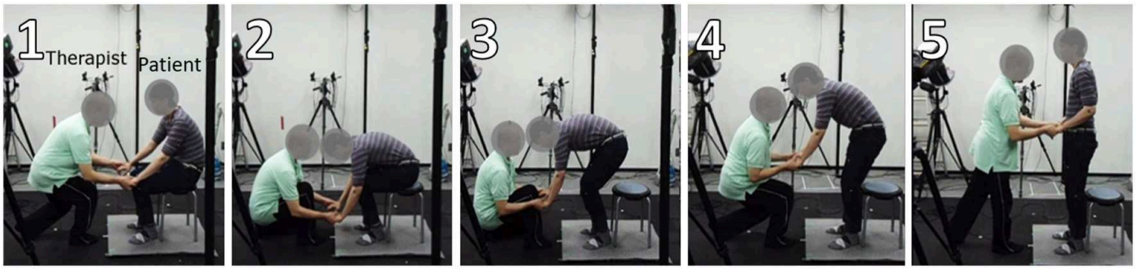


FIGURE 1 | Therapist (left) induces proper form of exercise on the patient (right).

four shoulder exercises. The system tracks a user's joints and overlays them on the user's body, where a target shoulder angle is presented (Tang et al., 2015). A VR example is the simulation of a 3D environment to facilitate arm-reaching exercises. This system also tracks users' joint positions, while vibratory feedback is given when task performance is successful (Kato et al., 2015). These systems use simple single-limb models to reduce the cognitive load on the patient and the computational load on the system.

The existing systems use single-purpose frameworks, which can be cumbersome to calibrate for the variety of different exercises and users. We highlight three issues in particular. First, complex multi-joint movements are largely unexplored. Focus has been largely on isolated individual movements, such as finger motion, knee, and shoulder movements. Important whole-body movements, such as the sit-to-stand exercise, cannot be addressed. Second, patient-specific calibration is difficult to achieve. On the one hand, many systems rely on generic expert templates. The template may differ greatly from the target user's body type and physical condition, potentially suggesting painful postures. On the other hand, automatic calibration would require large amounts of personalized training data, which is impractical with the limited time available during a therapy session. Finally, the physical therapist's motor skills are completely ignored. Many systems are designed according to the information provided by the expert PT. Expert *knowledge*, such as what types of exercise and target angles are useful, but expert *motor skills* play an equally important role during therapy (Tang and Dillman, 2013).

The present study aims to develop a data-efficient framework for personalized physiotherapy exercises. This framework solves the identified problems by allowing arbitrary whole-body motions, using only few user-specific demonstrations available during collocated therapy. We approach this problem by utilizing a model called Gaussian Process Dynamical Model (GPDM) (Wang et al., 2008). GPDM is part of a family of latent variable models which can represent high-dimensional observation data in a low-dimensional latent space (Lawrence, 2005). GPDM, a dynamical variant, has been demonstrated to work well with human motion data. The key idea is to embed and organize meaningful task demonstrations in the same latent space. However, there is currently no principled way to compare characteristics between multiple demonstrations.

We propose simple modifications to the GPDM to extract meaningful feedback mechanisms for self-correction. In the context of motor skill learning, *feedback* refers to performance-related information that a learner receives for performing a task. Two types of augmented feedback are typically presented to a learner. One conveys knowledge of results (KR) while the other conveys knowledge of performance (KP) (Sunaryadi, 2016). From the modified latent space, we extract features that convey both types of feedback.

This paper describes proposed modifications to the GPDM for robust self-correction of whole-body physiotherapy exercises. We then demonstrate our model's utility on the sit-to-stand task, an important activity of daily living. Our work eliminates the need for cumbersome calibration or large amounts of user-specific data for a personalized self-exercise system. Using only a limited amount of personalized data, expert-level feedback can be easily obtained. We will start with mathematical formulations of the GPDM and proposed modifications, followed by a description of the data collection and processing procedures. Subsequently, we present the results of the modifications then discuss them in relation to similar studies. Finally, we conclude with a brief summary and outlook.

2. MATERIALS AND METHODS

We first discuss the fundamental concepts of GPDM before introducing our proposed adjustments to the model. This is followed by details of the experiments and data-processing procedures performed for sit-to-stand motion data.

2.1. Gaussian Process Dynamical Model

The GPDM is a dynamical extension of the Gaussian Process Latent Variable Model (GPLVM), a class of latent variable models that allows non-linear generative mapping from latent space to observation space (Lawrence, 2005). The GPDM extends this model by introducing a dynamical prior in the latent space (Wang et al., 2008). For human motion, data is a sequence of poses indexed by discrete time t. The observation space is defined by a sequence of vector-valued poses $\mathbf{y}_t \in \mathbb{R}^D$, while the latent space is defined by a corresponding lower-dimensional sequence $\mathbf{x}_t \in \mathbb{R}^d$. Either can be written in the form

$$\mathbf{x}_t = \sum_i \mathbf{a}_i \phi_i(\mathbf{x}_{t-1}) + \mathbf{n}_{x,t}, \tag{1}$$

$$\mathbf{y}_t = \sum_j \mathbf{b}_j \psi_j(\mathbf{x}_t) + \mathbf{n}_{y,t}, \tag{2}$$

for weights $\mathbf{A} = [\mathbf{a}_1, \mathbf{a}_2, \ldots]$ and $\mathbf{B} = [\mathbf{b}_1, \mathbf{b}_2, \ldots]$, basis functions ϕ_i and ψ_j, and zero-mean white Gaussian noise $\mathbf{n}_{x,t}$ and $\mathbf{n}_{y,t}$.

The GPDM is calculated by marginalizing over parameters of the mappings (i.e., $\mathbf{A}$ and $\mathbf{B}$) and optimizing the latent coordinates of the training data. To obtain the data likelihood over the observations $\mathbf{Y}$, we assume an isotropic Gaussian prior on each $\mathbf{b}_j$ and marginalize over $\mathbf{B}$ to obtain

$$p(\mathbf{Y}|\mathbf{X}, \overline{\beta}) = \frac{|\mathbf{W}|^N}{\sqrt{(2\pi)^{ND}|\mathbf{K}_Y|^D}} \exp\left(-\frac{1}{2}\mathrm{tr}(\mathbf{K}_Y^{-1}\mathbf{Y}\mathbf{W}^2\mathbf{Y}^T)\right), \tag{3}$$

where $\mathbf{Y} = [\mathbf{y}_1, \ldots, \mathbf{y}_N]^T$ is a design matrix of poses, $\mathbf{X} = [\mathbf{x}_1, \ldots, \mathbf{x}_N]^T$ contains the corresponding latent coordinates, $\mathbf{W} \equiv diag(w_1, \ldots, w_D)$ is a scaling matrix, and $\mathbf{K}_Y$ is a kernel matrix. The elements of the kernel matrix are defined by a kernel function, $(\mathbf{K}_Y)_{i,j} = k_Y(\mathbf{x}_i, \mathbf{x}_j)$, chosen to be the default "RBF + bias + white" (Lawrence, 2005). The density over the latent coordinates can be obtained in a similar manner. We assume an isotropic Gaussian prior on each $\mathbf{a}_i$ and marginalize over $\mathbf{A}$ to obtain

$$p(\mathbf{X}|\overline{\alpha}) = \frac{p(\mathbf{x}_1)}{\sqrt{(2\pi)^{(N-1)d}|\mathbf{K}_X|^d}} \exp\left(-\frac{1}{2}\mathrm{tr}(\mathbf{K}_X^{-1}\mathbf{X}_{2:N}\mathbf{X}_{2:N}^T)\right), \tag{4}$$

where $\mathbf{X}_{2:N} = [x_2, \ldots, x_N]^T$, $\mathbf{K}_X$ is the $(N-1) \times (N-1)$ kernel matrix constructed from $\mathbf{X}_{1:N-1} = [\mathbf{x}_1, \ldots, \mathbf{x}_{N-1}]^T$, and $\mathbf{x}_1$ is given an isotropic Gaussian prior. The dynamics are chosen to be the default "RBF + linear + white" (Wang et al., 2008). The latent mapping, priors, and dynamics define a generative model for time series of the form

$$p(\mathbf{X}, \overline{\alpha}, \overline{\beta}|\mathbf{Y}) \propto p(\mathbf{Y}|\mathbf{X}, \overline{\beta})p(\mathbf{X}|\overline{\alpha})p(\overline{\alpha})p(\overline{\beta}), \tag{5}$$

where simple uninformative priors $p(\overline{\alpha}) \propto \prod_i \alpha_i^{-1}$ and $p(\overline{\beta}) \propto \prod_i \beta_i^{-1}$ are assumed. The GPDM is learned by minimizing the joint negative log-posterior of the unknowns $-\ln p(\mathbf{X}, \overline{\alpha}, \overline{\beta}|\mathbf{Y})$.

The definition for GPDM in Equation (5) is trained using pose sequence $\mathbf{Y}$, implying a single instance of motor behavior. However, Wang et al. also describe how the model can be extended to multiple sequences, explicitly modeling multiple instances simultaneously. To do so, the associated latent trajectories need to be embedded in a *shared* latent space. Now the observation space sequences $\{\mathbf{Y}^{(1)}, \ldots, \mathbf{Y}^{(P)}\}$ are still trained as a single data matrix, but each sequence is made independent by ignoring the temporal transitions between the last pose of sequence $i-1$ and the first pose of sequence i. Consequently, the associated latent trajectories $\{\mathbf{X}^{(1)}, \ldots, \mathbf{X}^{(P)}\}$ become disconnected (Wang et al., 2008).

2.2. Organizing Latent Trajectories

The original formulation of GPDM is insufficient for comparing multiple demonstrations. As is, we see a disorganized latent space where extraction of meaningful features becomes difficult (**Figure 2A**). We reorganize the space by introducing common reference points.

2.2.1. Common Pose

We hypothesize that the latent trajectories can be organized naturally by appending exact copies of common reference points to each motion sequence. We introduce the concept of a *common pose* (*CP*), which is appended to either end of each latent trajectory. The *CP* is calculated twice, based on the mean pose of trajectory end points. The first is done for the start of the motion sequence, and a second time for the end of the motion sequence. Formally, the common start pose is defined as

$$CP_{start} = \frac{1}{P}\sum_{p=1}^{P} \mathbf{y}_1^{(p)}, \tag{6}$$

where P is the number of sequences. Similarly, the common end pose is defined as

$$CP_{end} = \frac{1}{P}\sum_{p=1}^{P} \mathbf{y}_N^{(p)}, \tag{7}$$

where N is the number of poses in a sequence. The CP_{start} is appended to the *start* of each motion sequence $\mathbf{Y}^{(p)}$, while the CP_{end} is appended to the *end*. In our tests, we found that appending fifteen instances to each trajectory end works well.

2.2.2. Zone of Intermediate Poses

By connecting all latent trajectories through the common poses, any pair of trajectories creates an enclosed zone bounding all intermediate poses between them (**Figure 2B**). This Zone of Intermediate Poses (ZIP) has two useful properties for simple user feedback.

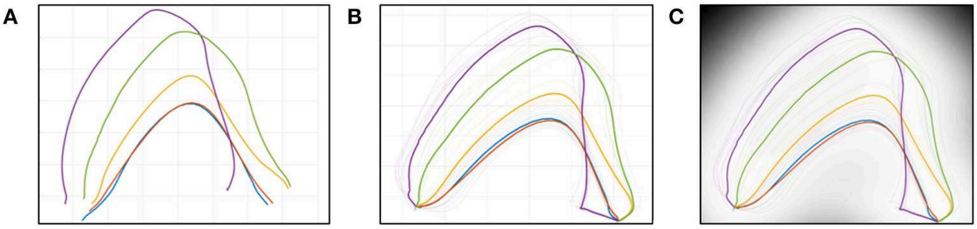


FIGURE 2 | Latent space representation, **(A)** GPDM, **(B)** GPDM with *CP*, **(C)** with precision.

The geometric area of the ZIP is a measure of performance similarity in the latent space. A large area indicates dissimilar movement patterns; a small area indicates similar movement patterns; and zero area (coincident lines) indicates exact-matching movement patterns. The area given by the coordinates (x_i, y_i) of two connected trajectories is defined by Gauss's Area Formula:

$$A = \frac{1}{2}\sum_{i=1}^{n}[x_i \cdot (y_{i-1} - y_{i+1})], \tag{8}$$

where A is the area of the polygon and n is the total number of vertices. The first and last points that create the polygon connect to each other, defined as $y_0 = y_n$ and $y_{n+1} = y_1$. Since our definition of similarity explicitly uses an area formula, the latent space is necessarily two-dimensional.

Poses sampled within the ZIP yield a smooth pose sequence, due to the proximity of the trajectories. Each latent point has an associated level of uncertainty in the pose space, with higher precision yielding better pose estimates. Precision is highest on the training points, but decreases rapidly as points are sampled farther away. In our implementation, the uncertainty of each point in the pose space is visualized by gray-scale coloring in the latent space. High-precision poses are indicated by a light color, while low-precision poses are indicated by darker colors (**Figure 2C**).

2.3. User Performance Feedback

We propose two ways to present both types of augmented feedback. Knowledge of Results is presented through a performance score, while Knowledge of Performance is presented by visualization of corrective poses. In other words, users can confirm if their performance is improving, and if not, how to correct their mistakes.

2.3.1. Performance Score

The user is given a *Performance Score* to indicate the quality of performance, starting from base performance to the desired expert-induced motion. The Performance Score is defined as

$$PS = 1 - \frac{A_{curr}}{A_{ref}}, \tag{9}$$

where A_{ref} is the area between the baseline and desired trajectories; A_{curr} is the area between the current and desired trajectories; and both A_{ref} and A_{curr} are calculated using (8). This convention implies that the baseline condition is assigned a 0% score while a goal condition is assigned a 100% score. Generally, a user's progress starts from 0% and improves all the way to 100%. However, it is possible to perform worse than the baseline by misinterpreting the expert's instructions. In this case, the score can go below 0%. The possible scoring outcomes are visualized with their corresponding ZIP in **Figure 3**.

2.3.2. Corrective Action

Poses can be visualized by sampling points from the latent space. Self-correction is facilitated by sampling a trajectory from the ZIP. By tracing a line from a *current* trajectory to a *desired* trajectory, a smooth corrective pose sequence can be inferred.

2.4. Data Collection

2.4.1. Participants

Nine healthy adult males (age: 27.2 ± 1.5 years, weight: 62.7 ± 10.7 kg) were recruited for the role of *subject* (person who stands up), while one PT (30 years experience) was recruited for the role of *expert* (person who induces change). All participants gave informed consent to participate in the experiment.

2.4.2. Experiment Protocol

Each subject was asked to perform a number of conditions during their respective session. At the start of every condition, a subject is instructed to sit comfortably on an armless, backless chair of fixed height (0.45 m), while the knee is flexed to 90°. Before recording, a subject is given a few minutes to familiarize themselves with the movement of the current sit-to-stand condition. There are three types of conditions, performed in the following order:

1. NATURAL: A subject is asked to stand up naturally, i.e., a self-selected pace and strategy. This condition is further subdivided into three typical sit-to-stand conditions. These conditions are distinguished by the arm position: (1) *N.folded*: folded across the chest, (2) *N.front*: to the front on the knees, and (3) *N.sides*: to the sides.
2. INDUCED: The PT is asked to induce the desired sit-to-stand motion as usually performed for his patients. This form of guidance is characterized by a light touch on the arms, requiring the subject to use his own strength to stand up.
3. LEARNED: A subject is asked to recall the new strategy that was learned from the INDUCED condition. The subject is then asked to replicate the motion as close as possible to the

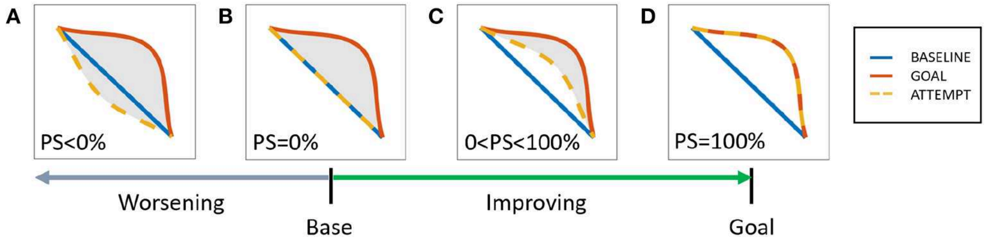


FIGURE 3 | Area-based performance score system, possible scores **(A)** worsening: *PS*<0%, **(B)** base: *PS* = 0%, **(C)** improving: 0<*PS*<100%, **(D)** mastery: *PS* = 100%.

taught movement. No additional instructions on timing or strategy was given.

Each subject performed all five conditions during their respective session, where each condition was repeated for six successful trials. A trial was considered a failure if data capture was affected in any way, e.g., occlusion of markers. A total of 270 successful trials (5 conditions × 6 trials × 9 subjects) were collected for analysis. The experiment protocol is summarized in **Figure 4**.

2.4.3. Motion Capture Recording

While the sit-to-stand tasks were being performed, the subject's whole-body motion was being recorded. The setup was an indoor MAC3D motion capture system (Motion Analysis Corp.), with 16 cameras mounted around the capture space (**Figure 5A**). The Cortex software from the same company provides the control panel for all devices and the tools for processing raw motion capture data.

Before data recording, 29 passive retroreflective markers were fitted to a subject's whole body, followed by a standard calibration procedure. The Helen Hayes marker set (**Figure 5B**) was used as a reference (Motion Analysis Corporation, 2006). During recording, the marker trajectories were sampled at 200 Hz with measurement units in millimeters. The x, y, z positions of each marker were continuously recorded, for a total of 87 channels (29 markers × 3 dimensions). An audible beep signals the subject when a trial starts and ends. The trial is ended a few seconds after the subject is fully standing.

2.4.4. Data Pre-processing

The marker data were first pre-processed before analysis, using built-in tools in Cortex (Motion Analysis Corp.) and custom code in Matlab (The MathWorks, Inc.). The procedures were performed in the following order:

1. **Noise removal**: Each trial was visually examined and corrected for occlusions and noise. The markers were then labeled and smoothed using a fourth-order Butterworth (6 Hz low-pass) filter.
2. **Data translation**: The coordinate system was standardized across trials. A common origin point was obtained using the static point between the R.Heel and L.Heel markers. The average point between the two markers were calculated, and the coordinates of all other markers were subtracted by this value. This procedure was performed for all trials individually.
3. **Data normalization**: To reduce inter-individual differences, the length units are normalized to a unitless value based on

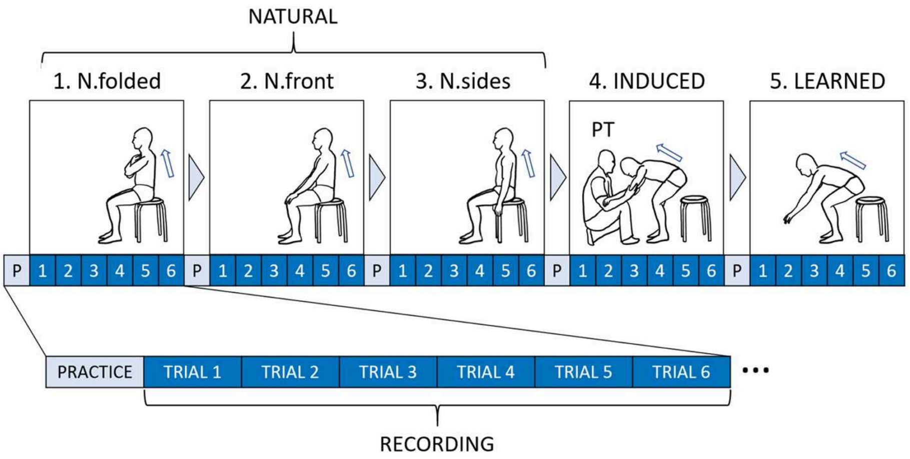


FIGURE 4 | Experimental flow diagram for each subject.

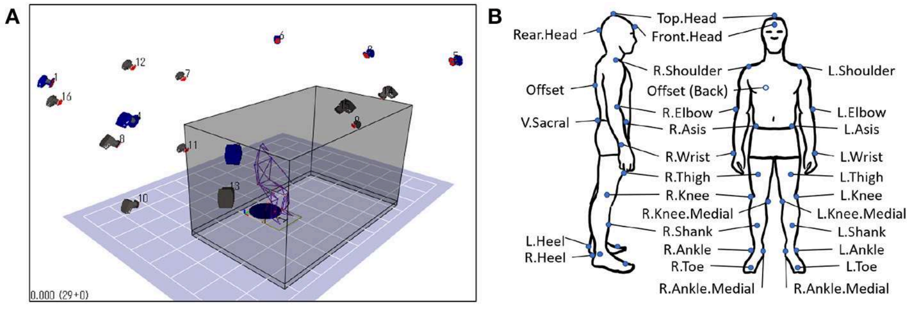


FIGURE 5 | Motion capture setup, **(A)** capture space, **(B)** Helen Hayes markerset.

height (Hof, 1996; Bahrami et al., 2000). For each trial, all coordinate values are divided by the vertical component of the Top.Head marker.

4. **Event standardization**: Each trial was truncated to retain only the relevant portion of the sit-to-stand motion. A *start* and an *end* event were defined based on a stable reference marker (Tully et al., 2005). The start event is defined as the moment when the speed of the R.Shoulder marker is >0 in the sagittal plane, while the end event is defined as the moment when the R.Shoulder reaches its highest vertical position.

3. RESULTS

A reorganized latent space was successfully extracted from the experiment data. Relevant properties of the new latent space are discussed in this section.

3.1. Latent Space Behavior

In conventional GPDM, multiple motor behaviors have no apparent relation in a shared latent space. This is true even when the *same* motor task is performed repeatedly. To organize the latent trajectories, we proposed to connect them according to known matching poses. Specifically, we appended reference common poses to both ends of each latent trajectory.

Results indicate that the latent trajectories have successfully connected at the common poses, CP_{start} and CP_{end}, found at either end. The trajectories of similar conditions stay close together, forming two subgroups. The NATURAL conditions stay close together, while the INDUCED and LEARNED conditions also stay close. However, the order within the subgroups vary among subjects. These results suggest that even small differences in the pose space can cause latent points to stay far apart. By connecting the trajectories through *exactly matching poses*, an organized latent space can be achieved. The extracted latent trajectories are shown in **Figure 6**.

3.2. Performance Score

Knowledge of results (KR) is one of two types of augmented feedback shown to be positively linked to motor skill learning (Sunaryadi, 2016). We proposed the Performance Score (PS) as a measure of performance success. The score uses the normalized geometric area between two connected latent trajectories as a measure of similarity between a baseline and a desired behavior. We assigned N.folded as the baseline condition and INDUCED as the desired condition.

Results indicate that the INDUCED scores are always higher than the LEARNED scores, while the LEARNED scores are always higher than the NATURAL scores. This consistent ordering suggests that all subjects were able to remember and perform part of the expert-induced movement. While all LEARNED scores are positive, the *N.front* and *N.sides* conditions report some negative values. The inconsistent negative scores of the other NATURAL conditions suggest that the expert-advised movement is not naturally achieved. These results indicate that the Performance Score can capture the expected performance improvements. A summary of all scores is shown in **Table 1**.

3.3. Corrective Pose Sequence

Knowledge of performance (KP) is the other type of augmented feedback for motor skill learning (Sunaryadi, 2016). We proposed sampling from the Zone of Intermediate Poses (ZIP) as a simple yet robust solution for movement self-correction, since a smooth pose sequence can be visualized by simply sampling adjacent points from the latent space.

Results show that by sampling along a trajectory, known pose sequences can be reconstructed. For example, tracing

TABLE 1 | Performance score summary.

Subjects	N.folded	N.front	N.sides	INDUCED	LEARNED
A	0	2.91	−14.69	100	75.52
B	0	23.06	27.44	100	36.95
C	0	−44.47	−56.51	100	52.14
D	0	−36.70	−38.41	100	65.77
E	0	15.34	11.46	100	38.97
F	0	−2.44	10.43	100	36.34
G	0	47.74	49.69	100	66.00
H	0	43.20	30.52	100	73.90
I	0	−19.33	4.08	100	62.48

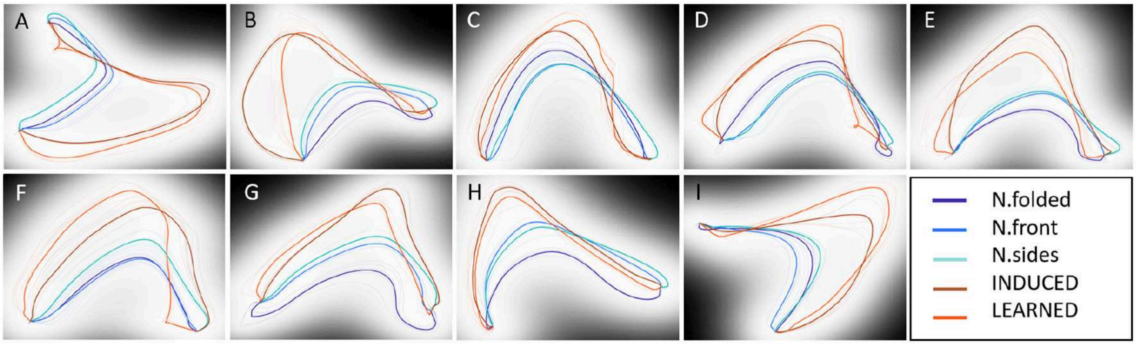


FIGURE 6 | Latent trajectory representation for all subjects **(A–I)**.

a NATURAL trajectory and the INDUCED trajectory shows the prototypical pose sequences. We can see some distinction between the two conditions as the INDUCED poses are lower and more forward-leaning (**Figure 7**). Further, by sampling the ZIP between the LEARNED trajectory and the INDUCED trajectory, one can visualize the corrective pose sequence (**Figure 8**). These results indicate that the generative portion of GPDM is unaffected by our modification, while the formed ZIPs can be used to identify the erratic portions of movement.

4. DISCUSSION

Our goal is to develop a data-efficient framework for personalized physiotherapy exercises. Our modified GPDM approach solves the common problems of existing telehealth applications, by providing personalized feedback for whole-body exercises, based on few expert-induced demonstrations. Specifically, two types of augmented feedback were extracted from the reorganized latent embedding, conveying both performance quality and the corrective action. By analyzing the feedback outcomes from sit-to-stand experiments, we confirmed the utility of our proposed method for an important physiotherapy exercise. In the discussion that follows, we compare our findings with related literature.

4.1. Latent Space Behavior

The introduction of the *common pose* solved the problem of relating multiple latent trajectories. The desired connecting effect is achieved because the common pose points "gather" nearby similar points. Effectively, the appended common poses reintroduce the *still* poses to both ends of the movement, i.e., subject is sitting still and standing still. The key differences with the original *still* poses are that they now match exactly and are shared by all sequences.

Multiple GPLVM studies demonstrate that multiple trajectories lie separately in the shared latent space, despite sharing common poses. One study modeled four golf swings from the same golfer, using conventional GPDM (Wang et al., 2008). Another study modeled sitting motion on surfaces of different heights, using Observation Driven GPLVM (Gupta et al., 2008). In both types of motion, the start poses are known to be the same pose, yet the latent starting points are represented by different points. One key difference with our target movement, i.e., sit-to-stand, is that the end pose is also known to be the same. This condition appears to be unique to our study.

4.2. Augmented Feedback

Our approach provides Knowledge of Results through the Performance Score, calculated as a function of the geometric area

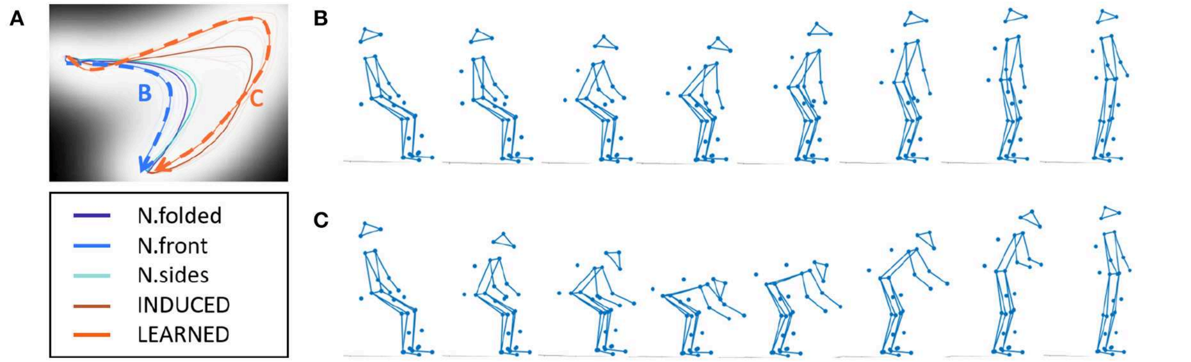


FIGURE 7 | Reconstructing demonstrated poses, **(A)** latent space, **(B)** N.folded sequence, **(C)** INDUCED sequence.

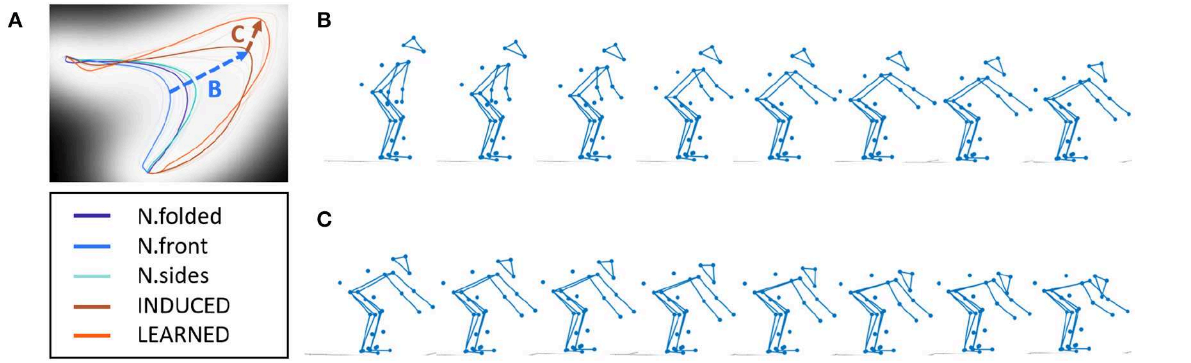


FIGURE 8 | Inferring corrective poses, **(A)** latent space, **(B)** N.folded to LEARNED sequence, **(C)** LEARNED to INDUCED sequence.

between two connected latent trajectories. Since the mapping from latent space to pose space is non-linear, the score does not necessarily convey the *scale* of progress, i.e., flexing a joint 50% more does not necessarily mean a subject improved by 50%. Other single-valued quantities for progress have been previously proposed, yet the scale of progress remains difficult to quantify exactly. Some examples include time to accomplish the task (Yang and Kim, 2002; Kato et al., 2015), joint flexion/extension angle (Piqueras et al., 2013), and root mean square error between poses (Anderson et al., 2013). Our proposed Performance Score has two advantages over the aforementioned metrics. First, it incorporates both spatial and sequential information. Second, calibration is not necessary when generalizing to other motions.

One interesting note regarding the Performance Score is that the area term in Equation (9) is a valid metric, i.e., distance function. Given any pair of connected latent trajectories, the four conditions of non-negativity, identity of indiscernibles, symmetry, and triangle inequality are all satisfied.

Our approach provides Knowledge of Performance through estimating the intermediate poses between two motor behaviors, allowing the user to visualize a corrective pose sequence. Due to the complexity of human motion, presenting the optimal amount and type of information is challenging. To reduce cognitive load, popular approaches use immersive technologies or only limb-specific movements. For example, some existing systems ask the user to move a target bone vertically and horizontally (Velloso et al., 2013) or move the shoulder laterally to a target number of degrees (Tang et al., 2015). Virtual reality applications, in particular, tend to be limited to upper-body movement for safety reasons. Our approach balances the amount of information captured and presented, by focusing only on the corrective poses of a whole-body model. Thus, removing the need to arbitrarily isolate body parts.

An important note regarding the corrective pose sequence is that it involves the *visualization* of the latent space, limiting the dimensionality to a maximum of three. Nonetheless, it would be of interest to discuss how different choices of dimensionality may affect our application. In particular, we highlight a three-dimensional example by Wang et al. (2008) and a nine-dimensional example by Damianou et al. (2016). In the first example, the original GPDM model was used to describe golf swing motion, and the latent space was set to three dimensions. The latent trajectories resulted in U-like shapes. In the second example, an extension of GPDM (dynamical variational GPLVM) was used to describe walking and running motions. In their model, the latent space was initially set to nine dimensions, but the model selects three "true" dimensions. Within the visualized three-dimensional space, each motion resulted in circular shapes with some distance between them.

We can see that relatively simple motions, such as golf swing, walking, and running, form "flat" trajectories. Despite being assigned three dimensions, the flat shapes of each motion suggest that these simple motions can be embedded in a two-dimensional space with some small loss. Although we can instead decide on a three-dimensional space for visualization, we argue that "navigating" a two-dimensional space can offer a more familiar experience. Since many commercial devices, such as smartphones and computer screens, offer a two-dimensional user interface, a user can more readily interact with the proposed system without the need to learn new forms of interaction. Thus, from the practical considerations discussed above, the simplest and most direct approach remains to be the two-dimensional representation.

4.3. Interpretation of Motor Knowledge

Subjects were found to interpret intervention differently after being guided by the PT. This difference can be observed by looking at the LEARNED condition, which is the subject's attempt to repeat the INDUCED movement. The PT's general strategy was observed to be guiding the subject lower and more forward than natural. However, some subjects undershoot while some overshoot the target motion. This behavior can be observed by plotting the body center of mass (CoM) in the sagittal plane (**Figure 9**). Further, the expert-INDUCED CoM trajectories were observed to be different across subjects. These results suggest that both treatment and subject response are personalized in practice.

We observed that the subjects with the lowest LEARNED performance scores (i.e., subjects B, C, E, F in **Table 1**) were performing a posture called "augmented arm," where arms are extended forward at shoulder height. We speculate that these subjects have the same interpretation of the therapist's intended change in motor behavior; thus, the same posture. Since this posture has the tendency to produce lower scores, such postures should be identified and avoided. This finding suggests that "low-score" postures may exist in other exercises as well, and identifying particular bad postures may help in self-correction. This posture may also be a reaction to counteract the slower motion induced by the PT, as it has been demonstrated to reduce standing up time (Kwong et al., 2014).

4.4. Alternative Models

Human motion can be modeled reasonably well in a few other ways. One can look at variants of GPDM, GPLVM, or deep learning techniques, which have the capacity to model a variety of high-dimensional dynamical data. We first discuss some extensions of GPDM and their application to different types of human motion. We also discuss two models which use conceptually different techniques to model dynamics.

GPDM has been demonstrated to work well with different types of whole-body motion. Wang et al. (2008) uses the GPDM to describe walking and golf swing movements, demonstrating the model's capability to model simple cyclic and acyclic motions. Chen et al. (2009) extends this concept by introducing a switching mechanism to account for motion sequences that involve switching dynamics, such as in salsa dancing. To our knowledge, a GPDM-based approach has not yet been applied to sit-to-stand motion, but we should note that Gupta et al. (2008) have applied it to the related stand-to-sit task. Together, these studies demonstrate the applicability of GPDM as a model for describing full-body motion, which can reasonably include sit-to-stand.

Hierarchical GPLVM (HGPLVM) can be considered an alternative implementation of dynamics for GPLVM. The main difference is that GPDM is autoregressive while HGPLVM is not. Instead, HGPLVM takes timestamps as inputs (Lawrence

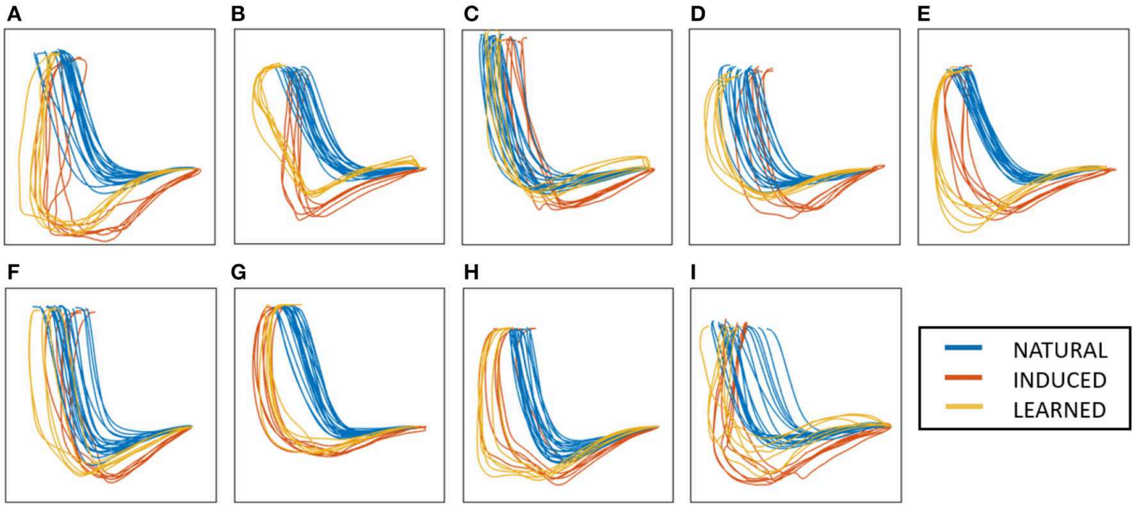


FIGURE 9 | Body center of mass in the sagittal plane for all subjects **(A–I)**.

and Moore, 2007). This is advantageous if uniform sampling is difficult to achieve. But in the controlled environment of telehealth, such a precaution is not necessary. HGPLVM also offers the option to "decompose" the subject model into component parts, allowing an isolated view of selected parts. However, this increases the number of visualized subspaces and requires a decision on the appropriate decomposition for each target activity. While still considering whole-body information, the single latent space representation of GPDM is a more straightforward approach to visualization.

Deep learning models are widely considered as universal approximators, which can work well with a large variety of data (Hornik et al., 1989). Given enough data and resources, deep learning models can exceed the performance of whatever specific-purpose model. In fact, a single network can be demonstrated to generalize well to multiple types of human actions. A trained model can simultaneously perform classification and prediction of novel poses with very little computational cost (Butepage et al., 2017). The main downside with such a model is the amount of resources necessary to perform training. In the context of single subjects with limited sessions, such large amount of resources is simply unavailable.

4.5. Motion Data Format

Motion data is an attractive modality for telehealth since it can be naturally learned and can be measured remotely. However, high-quality captures of specific motor tasks can sometimes be expensive and logistically difficult to obtain. Thus, in several GPLVM-based works on human motion data, no motion experiments were actually performed. Instead, the popular CMU Graphics Lab Motion Capture Database (mocap.cs.cmu.edu) was often used. One practical note regarding the human models in these studies is the format used. The format used in the CMU database contains *joint angle* information instead of the *marker coordinate* information we used in our study.

Although both formats are functionally similar, formats that store joint angles typically need to define a skeleton. The main advantage to this is that bone segments can be calibrated to each user, and limb lengths can be fixed. On the other hand, we can also argue that coordinate-based formats are more *accessible* as sensors and algorithms themselves measure anatomical *points*. Currently, in-home telehealth applications often use the ubiquitous Kinect sensor (Microsoft Corp.) which tracks 3D skeletal landmarks of the users. As computer vision algorithms become more advanced, ordinary images and videos are increasingly used to extract similar coordinate-based anatomical key points as well (Cao et al., 2017).

4.6. Limitations

The consistent results found across all subjects highlight the ability of our proposed feedback framework to perform as intended. However, our approach was only demonstrated to work well in a controlled environment. Tests on a larger variety of users and motor tasks would be necessary to confirm its clinical utility. We discuss some of the methodological limitations, and how these limitations compare to related studies.

The proposed framework was tested on nine healthy subjects, where each subject performed a total of 270 trials for standing up motion (five conditions for six trials each). Other motion studies employing GPDM rather focus on a larger variety of movements with fewer samples each. For context, Wang et al. (2008) used the original GPDM formulation to model three different whole-body motions: two gait cycles from one subject, one gait cycle from four subjects, and four golf swings from one subject. A GPDM extension by Gupta et al. conducts two types of experiments. The first experiment models jumping jack, walking, and climbing a ladder with one subject and one instance each. The second experiment models four different sitting instances for one subject (Gupta et al., 2008).

Our framework was tested on sessions conducted by one professional therapist. Realistically, different PTs may have different individual preferences, and it would be interesting to investigate intervention strategies across multiple PTs. Currently, we made the simplifying assumption that a patient's attending PT can best prescribe the personalized exercises. Identifying the appropriate impairments and conditions for our system would be challenging and is outside the scope of the current study. A separate study with a larger cohort would be necessary for each target motor impairment.

Learning the GPDM involves numerical optimization in estimating the model unknowns $\{\mathbf{X}, \overline{\alpha}, \overline{\beta}\}$. In our study, we needed to set both the number of learning iterations and the number of appended *CP*s. We found that setting a low number for both quantities saves on computational costs. In studies using GPLVM-based methods, the best working settings are generally reported without explanation. Some examples include iteration $T = 15$ for GPLVM (Lawrence, 2005), outer loop iteration $I = 100$ for GPDM (Wang et al., 2008), and no mention for Hierarchical-GPLVM (Lawrence and Moore, 2007). Notably, these methods have been demonstrated to generalize well despite few iterations and training samples.

5. CONCLUSION

This study aims to address some of the problems in existing telehealth systems. We first modified the GPDM algorithm, which allowed us to extract simple yet meaningful feedback mechanisms for self-correction in physiotherapy exercises. These mechanisms allow for whole-body movements and are personalized through expert-induced demonstrations. Our framework is appropriate for telehealth due to its ability to train a sensible model using only a small number of good examples. We confirmed its utility using sit-to-stand motion data, an important physiotherapy exercise.

We imagine that this research can take on two interesting directions. First, incorporating modalities other than motion can be used to extend the current framework. A multi-modal model can be an interesting approach to incorporate more assessment tools used by the PT. Second, our approach was designed with physiotherapy in mind, but it can be reasonably applied to arbitrary motor tasks where expert demonstrations are available, e.g., sports science. Instead of learning expert-induced movements, learning the expert motor skill itself is also an interesting possibility.

DATA AVAILABILITY STATEMENT

The datasets generated for this study are available on request to the corresponding author.

ETHICS STATEMENT

The studies involving human participants were reviewed and approved by the Nara Institute of Science and Technology Ethics Review Committee for research involving human participants. The patients/participants provided their written informed consent to participate in this study.

AUTHOR CONTRIBUTIONS

BL and TT conceived, designed, and performed the experiments. BL developed the software, analyzed the data, prepared tables and figures, and wrote the manuscript. All authors designed the study, contributed to data interpretation and manuscript revision, and read and approved the submitted version.

FUNDING

This work was supported in part by JSPS-KAKENHI grant number 18K19821 and the NAIST Big Data Project.

ACKNOWLEDGMENTS

This work made use of the Sheffield Machine Learning group's publicly-available package of GPLVM codes (https://github.com/SheffieldML/GPmat).

REFERENCES

Anderson, F., Grossman, T., Matejka, J., and Fitzmaurice, G. (2013). "Youmove: enhancing movement training with an augmented reality mirror," in *Proceedings of the 26th Annual ACM Symposium on User Interface Software and Technology* (St. Andrews, UK: ACM), 311–320.

Bahrami, F., Riener, R., Jabedar-Maralani, P., and Schmidt, G. (2000). Biomechanical analysis of sit-to-stand transfer in healthy and paraplegic subjects. *Clin. Biomech.* 15, 123–133. doi: 10.1016/S0268-0033(99)00044-3

Butepage, J., Black, M. J., Kragic, D., and Kjellstrom, H. (2017). "Deep representation learning for human motion prediction and classification," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition* (Honolulu, HI), 6158–6166.

Cao, Z., Simon, T., Wei, S.-E., and Sheikh, Y. (2017). "Realtime multi-person 2d pose estimation using part affinity fields," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition* (Honolulu, HI), 7291–7299.

Chen, J., Kim, M., Wang, Y., and Ji, Q. (2009). "Switching gaussian process dynamic models for simultaneous composite motion tracking and recognition," in *2009 IEEE Conference on Computer Vision and Pattern Recognition* (Miami Beach, FL: IEEE), 2655–2662.

Damianou, A. C., Titsias, M. K., and Lawrence, N. D. (2016). Variational inference for latent variables and uncertain inputs in gaussian processes. *J. Mach. Learn. Res.* 17, 1425–1486. Available online at: http://www.jmlr.org/papers/v17/damianou16a.html

Gupta, A., Chen, T., Chen, F., Kimber, D., and Davis, L. S. (2008). "Context and observation driven latent variable model for human pose estimation," in *2008 IEEE Conference on Computer Vision and Pattern Recognition* (Anchorage, AK: IEEE), 1–8.

Higo, M., and Khan, H. T. (2015). Global population aging: unequal distribution of risks in later life between developed and developing countries. *Glob. Soc. Policy* 15, 146–166. doi: 10.1177/1468018114543157

Hof, A. L. (1996). Scaling gait data to body size. *Gait Posture* 3, 222–223. doi: 10.1016/0966-6362(95)01057-2

Hornik, K., Stinchcombe, M., and White, H. (1989). Multilayer feedforward networks are universal approximators. *Neural Netw.* 2, 359–366. doi: 10.1016/0893-6080(89)90020-8

Kato, N., Tanaka, T., Sugihara, S., and Shimizu, K. (2015). Development and evaluation of a new telerehabilitation system based on VR technology using multisensory feedback for patients with stroke. *J. Phys. Ther. Sci.* 27, 3185–3190. doi: 10.1589/jpts.27.3185

Kwong, P. W., Ng, S. S., Chung, R. C., and Ng, G. Y. (2014). Foot placement and arm position affect the five times sit-to-stand test time of individuals with chronic stroke. *Biomed. Res. Int.* 2014:636530. doi: 10.1155/2014/636530

Lang, C. E., MacDonald, J. R., and Gnip, C. (2007). Counting repetitions: an observational study of outpatient therapy for people with hemiparesis post-stroke. *J. Neurol. Phys. Ther.* 31, 3–10. doi: 10.1097/01.NPT.0000260568.31746.34

Lang, C. E., MacDonald, J. R., Reisman, D. S., Boyd, L., Kimberley, T. J., Schindler-Ivens, S. M., et al. (2009). Observation of amounts of movement practice provided during stroke rehabilitation. *Archiv. Phys. Med. Rehabil.* 90, 1692–1698. doi: 10.1016/j.apmr.2009.04.005

Lawrence, N. (2005). Probabilistic non-linear principal component analysis with gaussian process latent variable models. *J. Mach. Learn. Res.* 6, 1783–1816. Available online at: http://www.jmlr.org/papers/v6/lawrence05a.html

Lawrence, N. D., and Moore, A. J. (2007). "Hierarchical gaussian process latent variable models," in *Proceedings of the 24th International Conference on Machine Learning* (Corvallis, OR:ACM), 481–488.

Lee, A. C., Davenport, T. E., and Randall, K. (2018). Telehealth physical therapy in musculoskeletal practice. *J. Orthop. Sports Phys. Ther.* 48, 736–739. doi: 10.2519/jospt.2018.0613

Motion Analysis Corporation (2006). *EVaRT 5.0 User's Manual*. Santa Rosa, CA: Motion Analysis Corporation.

Piqueras, M., Marco, E., Coll, M., Escalada, F., Ballester, A., Cinca, C., et al. (2013). Effectiveness of an interactive virtual telerehabilitation system in patients after total knee arthroplasty: a randomized controlled trial. *J. Rehabil. Med.* 45, 392–396. doi: 10.2340/16501977-1119

Schopp, L., Johnstone, B., and Merrell, D. (2000). Telehealth and neuropsychological assessment: new opportunities for psychologists. *Prof. Psychol. Res. Pract.* 31:179. doi: 10.1037/0735-7028.31.2.179

Sunaryadi, Y. (2016). "The role of augmented feedback on motor skill learning," in *6th International Conference on Educational, Management, Administration and Leadership* (Bandung: Atlantis Press).

Tang, A., and Dillman, K. (2013). *Towards Next-Generation Remote Physiotherapy With Videoconferencing Tools*. Technical report, University of Calgary.

Tang, R., Yang, X.-D., Bateman, S., Jorge, J., and Tang, A. (2015). "Physio@ home: exploring visual guidance and feedback techniques for physiotherapy exercises," in *Proceedings of the 33rd Annual ACM Conference on Human Factors in Computing Systems* (Seoul: ACM), 4123–4132.

Tully, E. A., Fotoohabadi, M. R., and Galea, M. P. (2005). Sagittal spine and lower limb movement during sit-to-stand in healthy young subjects. *Gait Posture* 22, 338–345. doi: 10.1016/j.gaitpost.2004.11.007

United Nations (2019). *World Population Ageing 2019: Highlights* New York, NY: United Nations, Department of Economic and Social Affairs, Population Division.

Velloso, E., Bulling, A., and Gellersen, H. (2013). "Motionma: motion modelling and analysis by demonstration," in *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems* (Paris: ACM), 1309–1318.

Wang, J., Fleet, D., and Hertzmann, A. (2008). Gaussian process dynamical models for human motion. *IEEE Trans. Pattern Anal. Mach. Intell.* 30:283. doi: 10.1109/TPAMI.2007.1167

Yang, U., and Kim, G. J. (2002). Implementation and evaluation of "just follow me": an immersive, vr-based, motion-training system. *Presence Teleoper. Virt. Environ.* 11, 304–323. doi: 10.1162/105474602317473240

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.

PERSPECTIVE
published: 28 April 2020
doi: 10.3389/fpubh.2020.00117

A "Third Wheel" Effect in Health Decision Making Involving Artificial Entities: A Psychological Perspective

Stefano Triberti[1,2]*, *Ilaria Durosini*[2] *and Gabriella Pravettoni*[1,2]

[1] *Department of Oncology and Hemato-Oncology, University of Milan, Milan, Italy,* [2] *Applied Research Division for Cognitive and Psychological Science, IEO, European Institute of Oncology IRCCS, Milan, Italy*

In the near future, Artificial Intelligence (AI) is expected to participate more and more in decision making processes, in contexts ranging from healthcare to politics. For example, in the healthcare context, doctors will increasingly use AI and machine learning devices to improve precision in diagnosis and to identify therapy regimens. One hot topic regards the necessity for health professionals to adapt shared decision making with patients to include the contribution of AI into clinical practice, such as acting as mediators between the patient with his or her healthcare needs and the recommendations coming from artificial entities. In this scenario, a "third wheel" effect may intervene, potentially affecting the effectiveness of shared decision making in three different ways: first, clinical decisions could be delayed or paralyzed when AI recommendations are difficult to understand or to explain to patients; second, patients' symptomatology and medical diagnosis could be misinterpreted when adapting them to AI classifications; third, there may be confusion about the roles and responsibilities of the protagonists in the healthcare process (e.g., Who *really* has authority?). This contribution delineates such effects and tries to identify the impact of AI technology on the healthcare process, with a focus on future medical practice.

Keywords: decision making, artificial intelligence, ehealth, patient-doctor relationship, technology acceptance, healthcare process, patient-centered medicine

Edited by:
Koichi Fujiwara,
Nagoya University, Japan

Reviewed by:
Takanori Fujita,
Keio University, Japan
Yoko Kiriyama,
MICIN, Japan

****Correspondence:***
Stefano Triberti
stefano.triberti@unimi.it

Specialty section:
This article was submitted to
Digital Public Health,
a section of the journal
Frontiers in Public Health

Received: *03 December 2019*
Accepted: *23 March 2020*
Published: *28 April 2020*

Citation:
Triberti S, Durosini I and Pravettoni G (2020) A "Third Wheel" Effect in Health Decision Making Involving Artificial Entities: A Psychological Perspective. Front. Public Health 8:117. doi: 10.3389/fpubh.2020.00117

INTRODUCTION

In the last few years, Artificial Intelligence (AI) has been on the rise, and some think that this technology will define the contemporary era as automation and factory tools defined the industrial revolutions, or as computers and the web characterized recent decades (1–3). These technologies, based on machine learning, promise to become more than simple "tools"; rather, they will be interlocutors of human operators that can help in complex tasks involving reasoning and decision making. The expression "machine learning" refers to a branch of computer science devoted to developing algorithms able to learn from experience and the external environment, improving performance over time (4–6). More specifically, algorithms are able to detect associations, similarities, and patterns in data, allowing predictions to be made on the likelihood of uncertain outcomes.

AIs and machine learning are present in a number of commonly used technologies, such as email, social media, mobile software, and digital advertising. However, the near future of AI is not that it will continue to work outside of the end users' awareness, as it mostly does nowadays; on the

TABLE 1 | A resume of the main areas for AI implementation in healthcare and medicine.

AI function in healthcare	Description; AI is…	Examples
Diagnosis	Employed as a diagnostic support tool; it analyzes clinical/pathological data to identify the disease	(11, 16, 20, 21)
Treatment (identification)	Involved in identification of treatment, often patient-specific solutions (genomics, precision medicine); it could participate in providing early interventions to delay the onset of chronic conditions (pre-emptive medicine)	(22–25)
Health management/patient engagement	Featured in devices that collect data on patient health status and provide recommendations for everyday care (eHealth, Digital Therapeutics, Ambient Intelligence)	(26–28)
Health Systems organization support/simulation	Used in agent-based simulations that model care coordination capabilities, providing insights for organizational improvements	(29, 30)

contrary, AI promises to become an active collaborator with human operators in a number of tasks and activities. AIs are able to analyze enormous quantities of data of various contents and formats, even where it is dynamically changing (Big Data). AIs identify associations and differences between data and provide human operators with outputs that are impossible to achieve by humans alone, at least in the same amount of time.

An example of such outputs are medical diagnoses and the identification of therapy regimens to be administered to patients. Health professionals (physicians especially) will increasingly interact with AIs to get information on their patients that will hopefully be more exact, specific, and based on objective data (7–9). Diagnostic decision support could be considered the main application area for AI-based innovation in medical practice (10, 11). Basically, machine learning devices are trained to classify stimuli based on initial examples. For instance, tumor types can be identified by the comparison of patient's TAC with information coming from scientific literature (12, 13); the same can be done with pictures of skin lesions (14), optical coherence tomography in the case of sight diseases (15–17), or the integration of clinical observations and medical tests for other diseases (18, 19). While diagnosis is recognized by many as the main area for AI implementation in medicine and healthcare, others could be envisaged, as summarized in **Table 1**.

However, the study of AI in healthcare in its social-psychological aspects is still an underrepresented area. One important field is that of "Explainable Artificial Intelligence" (commonly abbreviated in XAI), namely the research on AI's transparency and ability to explain its own elaboration processes. The American Agency for Advanced Research Projects for Defense (DARPA) launched a program on XAI (31), and the European Parliament demands a "right to explanation" in automated decision making (32). Indeed, one issue with AI implementation in professional practice regards the fact that it is supposed to be used by non-professionals: doctors, marketers, or military personnel are not expected to become experts in informatics or AI development, yet they will have to interact with artificial entities to make important decisions in their fields. While one could easily agree with the analyses and outputs of AIs, *trusting them* and taking responsibility for decisions that will affect the "real world" is no easy task. For this reason, XAI is identified by many scholars as a priority for technological innovation. Miller and colleagues (33, 34) maintain that AI developers and engineers should turn to social sciences in order to understand what is an explanation, and how it could be effectively implemented within AIs' capacities. For example, Vellido (35) proposed that AIs learn to make their processes transparent via visual aids that help a human user to understand how a given conclusion has been reached; Pravettoni and Triberti (36) highlighted that explanation is rooted in interaction and conversation, so that a complete, sophisticated XAI would be reached when artificial entities were able to communicate with human users in a realistic manner (e.g., answering questions, learning basic forms of perspective-taking, etc.). In any case, besides working on AI-human interfaces, another field of great interest is that of AI's impact on professional practice or the prediction of possible organizational, practical, and social issues that will emerge in the context of implementation.

Though the contribution of AIs to medical practice is promising, their impact on the clinician-patient relationship is still an understudied topic. From a psychosocial point of view, it is possible that new technologies will influence the relationship between clinicians and patients in several ways. Indeed, the introduction of AI into the healthcare context is changing the ways in which care is offered to patients: the information given by AIs on diagnosis, treatment, and drugs will be used to make decisions in any phase of the healthcare journey (e.g., choices on treatment or lifestyle changes, deciding to inform relatives of one's health status, communication of bad news).

According to a patient-centered perspective, such care choices should be made by the patient and the doctor within a mutual collaboration, which points to the popular concept of *shared decision making*. This concept has become fundamental in the debate on patient-centered approaches to care, with the number of scientific publications on the subject rising more than 600% from 2000 to 2013 (37). Reviews show that the communication process and relationship quality among doctors and patients has a significant effect on patients' well-being and quality of life, so that the proper communication style can alleviate the traumatic aspects of illness (38–40). However, when prefiguring the adoption of AIs participating in diagnosis and therapy identification, it is possible that the same concept of "shared decision" should be updated, taking into consideration the contribution of artificial entities.

WHO TO SHARE DECISION MAKING WITH?

Shared decision making has been proposed as an alternative paradigm to the "paternalistic" one (41, 42). The latter model dominated disease-centered medicine, with the physician being authoritative and autonomous, giving recommendations to patients without taking into consideration their full understanding, personal needs, and feelings. While the "paternalistic" physician intended to act in the best interest of the patient, such an approach may be ineffective or counterproductive in the end, because the patient may not understand nor follow the recommendations (43). Shared decision making is a process by which patients and health professionals discuss and evaluate the options for a particular medical decision, in order to find the best available treatment that is based on knowledge that is accessible and comprehensible for both and satisfies both needs (44–46). During this process, the patient is made aware of diagnostic and treatment pathways, as well as of related risks and benefits; also, the patient's point of view is taken into consideration in terms of preferences and personal concerns (47–49).

In other words, shared decision making entails a process of communication and negotiation between the health professional and the patient in which both medical information (e.g., diagnosis, therapy, prognosis) and patient's concerns (e.g., doubts and request for clarification, lifestyle changes, worries for the future, etc.) are exchanged.

In the near future, where AI is expected to take a role in medical practice, it is important to understand its influence on shared decision making and on the patient-doctor relationship as a whole. In most of the health systems around the world, the patient has the right to be informed about which tools, resources, and approaches are being employed to treat his or her case; a patient will have to know that the diagnosis or even the medical prescriptions first came out of a machine, not through the human doctor's effort. Presumably, just the knowledge of the presence of a "machine" in the healthcare process could influence the attitudes of doctors and patients: from a psychological point of view, attitudes toward something may develop before any direct experience of it (e.g., as a product of hearsay, social norms, etc.), and influence subsequent conduct (50). Indeed, while medicine itself is inherently open to innovation and technology, some health professionals harbor negative attitudes toward technology for care (51, 52), the main reasons being the risk they feel for patient de-humanization (53, 54) or the fear that tools they are not confident in mastering may be used against them in medical controversies (55). Similarly, patients who do not feel confident in using technology ("computer self-efficacy") benefit less from eHealth resources than other patients who do feel confident (56, 57), and technological systems for healthcare are not expected to work as desired if development is not tailored to users' actual needs and context of use (26). Though these data regard types of technologies different from AIs, they clearly show that technological innovation in the field of health is hardly a smooth process. While it is clear how the "technical" part of medicine (e.g., improving diagnosis correctness) would benefit from AIs and machine learning devices, their impact on the patient-doctor relationship is mostly unknown.

Furthermore, with the development of eHealth (58, 59) and the diffusion of interactive AIs as commonly used tools (e.g., home assistants), it is possible that chronic patients (e.g., patients with obesity, arthritis, anorexia, heart disease, diabetes) will be assisted by AIs in their everyday health management. Indeed, for example, the American Food and Drugs Administration (FDA) has made "significant strides in developing policies that are appropriately tailored to ensure that safe and effective technology reaches users" [US FDA (60), p.2], promoting the development of Software as a Medical Device (SaMD), devices that play a role in diagnosis or treatment (not only health or wellness management). FDA-approved digital therapeutics include, for example, reSET developed by PEAR Therapeutics (61), which delivers cognitive-behavioral therapy to patients suffering from substance abuse (62). It is possible that similar future resources will include AIs that directly interact with patients based on natural language processing.

In other words, AIs will not be just a "new app" on doctors' devices, but active interlocutors, able to deliver diagnosis, prognosis, and intervention materials to both the doctor and the patient. According to Topol (7), AI's implementation in care could potentially have positive effects, but this depends mainly on doctors' attitudes: for example, if AI were to take on administrative and technical tasks in medicine, doctors would have the occasion to recover the "lost time" for consultation with and empathic listening to their patients, so to improve shared decision making. In this sense, AI would become an active go-between among care providers and patients.

This considered, it becomes fundamental to understand whether we should expect structural changes in the same context of shared decision making and medical consultation. Will patients interact with AIs directly? Will doctors encounter difficulties and obstacles in adapting their work practices to include technologies able to participate in diagnosis and treatment? Are patient-doctor decisions to be shared with artificial entities too?

At the present time, the scientific literature lacks research data to fully respond to these questions. However, by considering the literature on health providers' reactions to technological innovation and the psychology of medicine, it is possible to prefigure some social-psychological phenomena that could occur in the forthcoming healthcare scenarios in order to prepare to manage undesirable side-effects.

A "THIRD WHEEL" EFFECT

In common language, the expression "third wheel" refers to someone who is superfluous with respect to a couple. Typically, the focus of the expression is on this third person, who unintentionally finds him or herself in the company of a couple of lovers and feels excluded and out of place. On the other side of the relationship, the couple may be unaware of the stress caused to the third wheel, or they may feel awkward and uneasy because of the unwanted presence. In other words, a third wheel is someone who is perceived as an adjunct, something unnecessary, who may spoil the mood and negatively influence others' experience. Despite just being a popular idiom, this expression is sometimes used in psychological research to describe relationship issues as experienced by research participants (63): new technology, specifically social media, has been called a third wheel as well because of its possible negative influence on relationship quality (64, 65).

We propose to employ the expression "third wheel" to highlight an emergent phenomenon relating to the implementation of artificial entities in real-life contexts: while technologies become more and more autonomous, able to talk, to "think," and to actively participate in decision making, their role within complex relationships may be unclear to the human interlocutors, and new obstacles to decision making could arise.

We identified three main ways a third wheel effect may appear in medical consultation aided by artificial intelligence: *decision paralysis, or a risk of delay, "Confusion of the Tongues," and role ambiguity*. In the next sections, these will be described in detail.

Decision Paralysis, or a Risk of Delay

As previously stated, current AIs are not transparent in their elaboration processes; that is, their interlocutors may have no clear representation of how AIs have reached a given conclusion: this could generate "trust issues," especially when important decisions should be taken on the basis of these conclusions. According to Topol and his seminal book *Deep Medicine* (7), one positive consequence for AI implementation in medical practice that we could hope for is giving back time to doctors to reserve to empathic consultation and patient-centered medicine. Indeed, if AIs were to take on technical and administrative tasks in medicine, doctors could devote their attention to patients as individuals and improve the "human side" of their profession. However, we should take into account that doctors using AIs will need to contextualize and justify their role within practice and the relationship with the patient. It could be said that doctors will become "mediators" between their artificial allies and the patients: AI's conclusions and recommendations should be reviewed by the doctor, approved and refined, and explained to the patient, answering his or her questions. On the other side, future technologies could include opportunities for direct interaction between AIs and patients: for example, digital therapeutics, or eHealth applications providing assistance to patients and caregivers in the management and treatment of chronic diseases, could potentially include access to the AI providing diagnosis and therapy guidelines. However, we can foresee that a patient would still need approval and guidance from the human doctor for modifications to the treatment schedule, medication intake, specific changes to lifestyle, and everyday agenda.

Such a "mediation" role could be time-consuming and, at least at an organizational level, generate decision paralysis or delays. Imagine that a hospital tumor board has to make a decision on a patient's diagnosis and only half of the board members agree with the AIs' recommendation; or, that a patient receives an important indication from the AI (e.g., stop taking a medication because wearable devices registered unwanted side effects), but he or she struggles to get in contact with his or her doctor to gain reassurance that this is the right thing to do.

These are examples of the implementation of AIs giving rise to a risk of delay. Though the technical processes could be accelerated, organizational and practical activities could be affected by the complex inclusion of an additional figure in the decision making process.

"Confusion of the Tongues"

The psychoanalyst Sandor Ferenczi used the expression "confusion of the tongues" to identify the obstacles inherent to communication between adults and children, who are inexorably heterogeneous in their mental representations of relationships and emotional experience. Since then, it has proved an effective expression to refer to interlocutors misinterpreting one another without knowledge.

The expression could be useful when we consider the utilization of AI in medical diagnosis, especially when the latter should be communicated to the patient. The physician is not a simple "translator" of information from the AI to the patient; on the contrary, he or she should play an active role during the process. Let us consider an example: an AI requires that the information on the patient's state is entered according to formats, categories, and languages that it is able to understand and analyze (e.g., data); however, it is possible that not all the relevant information for diagnosis could be transformed as such. How can doctors enter an undefined symptom, a general malaise, or a vague physical discomfort, if the patient himself or herself is hardly able to describe it? Even if trained in the understanding of natural language, the AI will not be able to integrate such information in its original form; this is not related to some sort of malfunction; rather, the AI does not have access to the complex and subtle emotional intelligence abilities that a human doctor can employ when managing a consultation with a patient. Specifically, one risk is that doctors would try to adapt symptoms to AIs' language and capabilities, for example by forcing the information coming from the patient into predefined categories; this could be related to an exaggerated faith in the technology itself, which could lead human users to overestimate its abilities (66).

This could lead patients not being motivated to report doubts, feelings, and personal impressions; indeed, patients can feel when doctors are not really listening to them (67, 68) and could experience a number of negative emotions ranging from anger to demoralization and a sense of abandonment (69–71). Such experiences have a detrimental effect both on the success of shared decision making and on therapy effectiveness because the patient will not adhere to the recommendations (72, 73). In other words, in this way, the new technology would become a source of patient reification, neglecting important elements that only humans' emotional intelligence can grasp.

Role Ambiguity

When the press started to write about Artificial Intelligence in medicine, a number of authoritative medical sources expressed a firm belief: AI will not replace doctors; it will only help them to do their jobs better, especially by analyzing complex medical data. However, we still have no clear idea about the perception of AIs *on the side of patients*; though it is obvious that AIs will not take over doctors' work, what will the patients think?

According to a recent survey by PwC on 11,000 patients from twelve different countries, 54% of the interviewees were amenable to the idea of being cured by artificial entities, 38% were against it, and rest were uncertain. The highest rates of acceptance can be traced to developing countries, which are open to any innovation in medicine, while countries used to high-level care systems were more critical.

This points to the need for AI innovation to be communicated and explained to patients in the right way, by justifying its added value but also by avoiding the risk that technology takes the place of human doctors *in patients' perception*. For example, as shown by some of the first implementations of the AI system Watson for Oncology by IBM (74, 75), it may happen that the diagnosis provided by AI does not mirror completely the ideas and assessments implemented by the doctor. It is possible that physicians, patients, and AIs will provide different narratives of diagnosis, prognosis, and treatment. This situation of ambiguity and disagreement could lead the patient to experience uncertainty, not knowing what opinion to follow, who really has authority, and who is actually working to help him or her.

When the doctor has to explain the role of AI in the consultation, he or she will have to reassure the patient that the recourse to such a technology is a desirable strategy to employ to provide the best possible consultation and treatment. However, in the perception of the patient, this communication may contain an implicit message, for example, that *someone else* is doing the doctor's work. A recent case was reported in the news worldwide where a patient and his family received a terminal diagnosis from the doctor on a moving robot interface: the family was shocked by the experience and perceived the use of the machine as an insensitive disservice (76).

This is an extreme example where a machine has been introduced in a delicate phase of the healthcare process: even if the machine was not acting autonomously, the effect was disastrous. Obviously, it is fundamental to build an empathic relationship with patients, especially with those dealing with the reality of death and grief (77, 78); while speaking through a machine is "technically the same" as in person, a grieving patient or his caregiver could reasonably feel talking to a robot to be a tragically absurd situation. This example shows how the implementation of technology should be analyzed not only in terms of functionality and technical effectiveness but also from the point of view of patients (79), taking into account their reaction and its consequences for patient health engagement as well as their commitment to shared decision making.

But if the AI is good as my doctor, or maybe even better, whom should I trust? A similar issue exists in medicine already: when multidisciplinary care is offered, patients may experience anxiety and confusion because they have to schedule appointments with several doctors and are not sure who to refer to with specific questions or who to listen to when recommendations are (or appear to them) contradictory (80); patients may find it difficult to trust health providers when the recommendation received is unexpected or counterintuitive (81) and they sometimes consult multiple health professionals, searching for infinite alternative options, as if the cure were some goods to buy, a maladaptive conduct known as "doctor shopping" (82). In other words, when there is disagreement, doubt, or ambiguity in diagnosis and treatment, its effect on the patient's perception and behavior should be taken into consideration and adequately managed within the consultation.

On the side of the doctor, the description of symptoms, diagnosis, and prognosis given by the AI could be more clear and understandable than the patient's; indeed, AI uses medical language and adopts the perspective of a medical professional, relying on objective data and scientific literature. While a patient could often experience difficulties when trying to explain his or her experience, AI could provide a different "narrative" of the diagnosis that the doctor would perceive as more comprehensible and reassuring. In this case, it is possible that the patient's testimony would be undermined or partially ignored, this way losing trace of the nuances and peculiarity of the actual patient's situation, which only a fine-grained analysis of the subjective testimony could detect.

To sum up, AI could potentially take the role of doctors in patients' perception or the opposite.

DISCUSSION

In this contribution, we tried to identify possible dysfunctional effects of AI's inclusion in medical practice and consultation, conceptualizing them as multiple forms of a "third wheel" effect; besides prefiguring them, it is possible to sketch solutions to the issues to be explored by means of future research.

The three forms of the "third wheel" effect may affect three important areas of the medical consultation: *organizational*, *communicational*, and *socio-relational* aspects, respectively (see **Figure 1** for a summary of the concept).

First, doctors and patients could experience a *decision paralysis*: decisions could be delayed when AIs' recommendations are difficult to understand or to explain to patients. *Decision paralysis* may affect the o*rganizational* aspects of healthcare contexts. It refers to how AI technology will be implemented in healthcare systems that may struggle to adapt their timings, procedures, and organizational boundaries to innovation. Tackling these issues entails making plans for the management of AI implementation that take into consideration not only the benefits of AI for the "technical" aspects of medical activities but also the behavior of organizational units toward AI outcomes and how these outcomes fit among any care practice processes.

Second, the presence of AIs could lead to a "*confusion of the tongues*" between doctors and patients, because patients' health information could be lost or transformed when adapted to AI's classifications. *"Confusion of the tongues*" affects *communication between doctors and patients*. It refers to the actual possibility for patients and doctors to understand each other and enact a desirable process of shared decision making. The solution to these possible issues involves the design of training resources for doctors that make them aware of how AI implementation could be perceived by patients; desirable

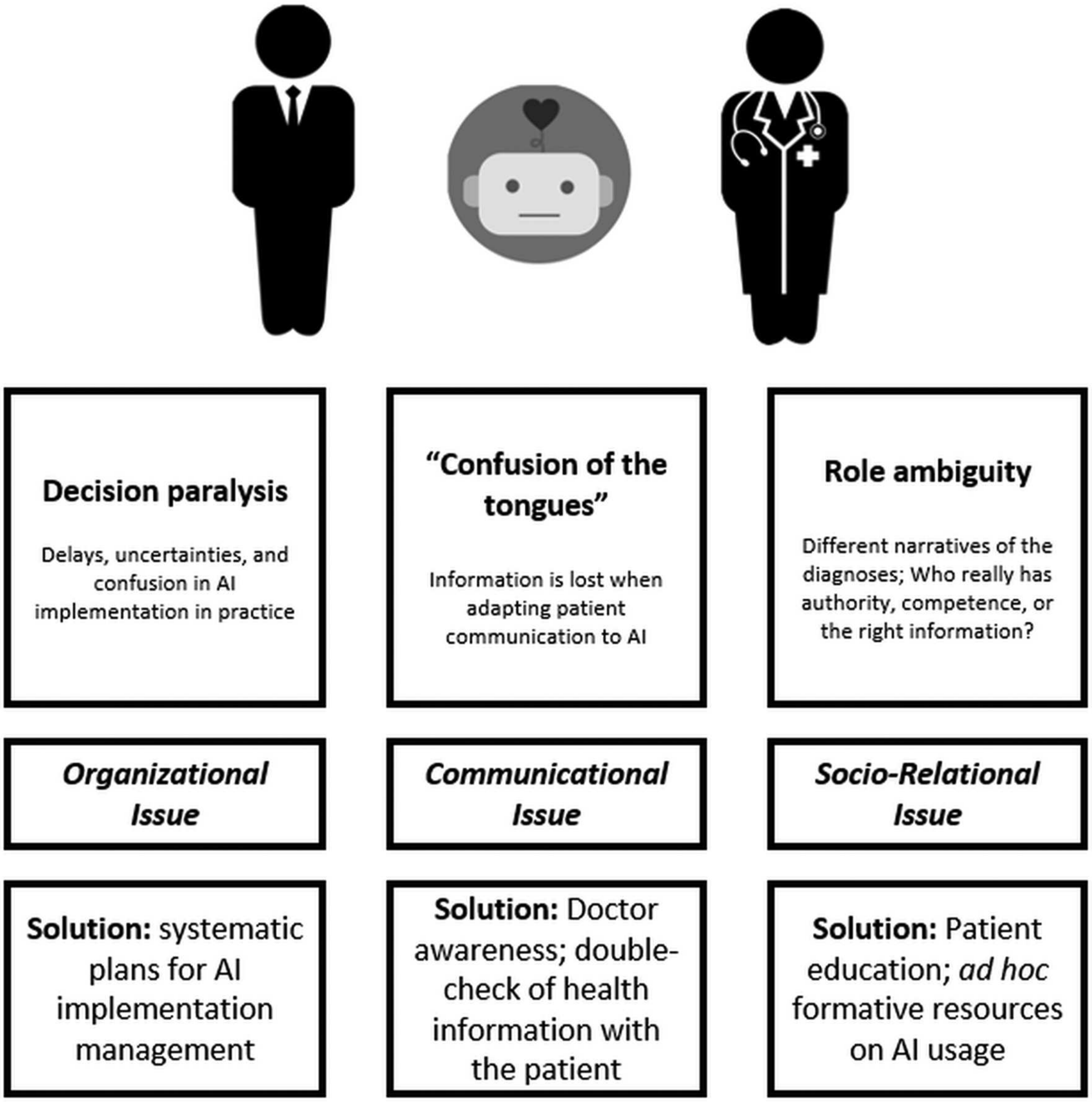

Decision paralysis
Delays, uncertainties, and confusion in AI implementation in practice
"Confusion of the tongues"
Information is lost when adapting patient communication to AI
Role ambiguity
Different narratives of the diagnoses; Who really has authority, competence, or the right information?
Organizational Issue
Communicational Issue
Socio-Relational Issue
Solution: systematic plans for AI implementation management
Solution: Doctor awareness; double-check of health information with the patient
Solution: Patient education; ad hoc formative resources on AI usage

FIGURE 1 | A summary of the "third wheel" effect with possible related solutions.

practices within patient-doctor communication would include double-checking health-related information to address possible confusion arising from the delivery of relevant information mediated by AI.

Lastly, the involvement of AIs could cause confusion regarding roles in patient-doctor relationships when ambiguity or disagreement arises about treatment recommendations. *Role ambiguity* acts on *socio-relational* aspects in healthcare contexts. It refers to the unwanted effects on trust and quality of relationship related to the addition of an artificial interlocutor within the context. These relational aspects are important prerequisites for achieving a desirable healthcare collaboration. Therefore, solutions to role ambiguity issues would entail proper patient education on the usage of AI in their own healthcare journey, especially when intelligent technology resources interact with them directly, mediating treatment (e.g., eHealth). Future studies on the ethical implementation of AI in medical treatment should consider patients' *perception* of these tools and forecast under which conditions patients may feel "put aside" by their doctor because health advice and treatment are delivered by autonomous technologies.

The identification of the psychosocial effects of AI on medical practice is speculative in nature: we should wait until these technologies become actual protagonists in a renovated approach to clinical practice in order to collect data about their effects on the scenario. As a limitation of the present study, we did not report research data; rather, we tried to sketch possible correlates of AI implementation in healthcare based on the literature in health psychology and the social science of technology implementation issues. We believe that consideration of such established phenomena may help pioneers of AI in healthcare to forecast (and possibly manage in advance) issues that will characterize AI implementation as well. Future studies may employ technology acceptance measures to explore health professionals' and patients' attitudes toward artificial intelligence. Moreover, qualitative research methods (e.g., ethnographic observation) could be employed within the pioneer contexts where AIs start to be used in medical consultation, in order to capture the possible obstacles to practice consistent with the third wheel effect prefigured here.

AUTHOR CONTRIBUTIONS

ST conceptualized the ideas presented in the article and wrote a first draft. ID helped to refine the theoretical framework and edited the manuscript. GP contributed with important intellectual content and supervised the whole process. All authors contributed to revision, read and approved the submitted version.

FUNDING

ST and GP were supported by MIUR - Italian Ministry of University and Research (Departments of Excellence Italian Law n.232, 11th December 2016) for University of Milan. ID was supported by Fondazione Umberto Veronesi.

REFERENCES

1. Barrat J. *Our Final Invention: Artificial Intelligence and the End of the Human Era*. New York, NY: St. Martin Press. (2013).
2. Li B, Hou B, Yu W, Lu X, Yang C. Applications of artificial intelligence in intelligent manufacturing: a review. *Front Inform Technol Electron Eng*. (2017) 18:86–96. doi: 10.1631/FITEE.1601885
3. Makridakis S. The forthcoming artificial intelligence (AI) revolution: its impact on society and firms. *Futures*. (2017) 90:46–60. doi: 10.1016/j.futures.2017.03.006
4. Angermueller C, Pärnamaa T, Parts L, Stegle O. Deep learning for computational biology. *Mol Systems Biol*. (2016) 12:878. doi: 10.15252/msb.20156651
5. Ghahramani Z. Probabilistic machine learning and artificial intelligence. *Nature*. (2015) 521:452–9. doi: 10.1038/nature14541
6. Lawrynowicz A, Tresp V. Introducing Machine Learning. In: Lehmann J, Volker J, editors. *Perspectives on Ontology Learning*, Berlin: IOS Press. (2011)
7. Topol E. *Deep Medicine*. New York, NY: Basic Books. *(*2019).
8. Wartman SA, Combs CD. Reimagining medical education in the age of AI. *AMA J Ethics*. (2019) 21:146–52. doi: 10.1001/amajethics.2019.146
9. Wooster E, Maniate J. Reimagining our views on medical education: part 1. *Arch Med Health Sci*. (2018) 6:267–9. doi: 10.4103/amhs.amhs_142_18
10. Amato F, López A, Peña-Méndez EM, Vanhara P, Hampl A, Havel J. Artificial neural networks in medical diagnosis. *J Appl Biomed*. (2013) 11:47–58. doi: 10.2478/v10136-012-0031-x
11. Jiang F, Jiang Y, Zhi H, Dong Y, Li H, Ma S, et al. Artificial intelligence in healthcare: past, present and future. *Stroke Vasc Neurol*. (2017) 2:230–43. doi: 10.1136/svn-2017-000101
12. Fakoor R, Nazi A, Huber M. Using deep learning to enhance cancer diagnosis and classification. In *Proceedings of the international conference on machine learning (Vol. 28)*. New York, NY: ACM. (2013).
13. Hu Z, Tang J, Wang Z, Zhang K, Zhang L, Sun Q. Deep learning for image-based cancer detection and diagnosis – A survey. *Pattern Recognit*. (2018) 83:134–49. doi: 10.1016/j.patcog.2018.05.014
14. Esteva A, Kuprel B, Novoa RA, Ko J, Swetter SM, Blau HM, et al. Dermatologist-level classification of skin cancer with deep neural networks. *Nature*. (2017) 542:115–8. doi: 10.1038/nature21056
15. De Fauw J, Ledsam JR, Romera-Paredes B, Nikolov S, Tomasev N, Blackwell S, et al. Clinically applicable deep learning for diagnosis and referral in retinal disease. *Nat Med*. (2018) 24:1342–50. doi: 10.1038/s41591-018-0107-6
16. Gargeya R, Leng T. Automated identification of diabetic retinopathy using deep learning. *Ophthalmology*. (2017) 124:962–9. doi: 10.1016/j.ophtha.2017.02.008
17. Gulshan V, Peng L, Coram M, Stumpe MC, Wu D, Narayanaswamy A, et al. (2016). Development and validation of a deep learning algorithm for detection of diabetic retinopathy in retinal fundus photographs. *JAMA*. 316:2402–10. doi: 10.1001/jama.2016.17216
18. Abiyev RH, Ma'aitah MKS. Deep convolutional neural networks for chest diseases detection. *J Health Eng*. (2018) 2018:4168538. doi: 10.1155/2018/4168538
19. Litjens G, Sánchez CI, Timofeeva N, Hermsen M, Nagtegaal I, Kovacs I, et al. Deep learning as a tool for increased accuracy and efficiency of histopathological diagnosis. *Sci Rep*. (2016) 6:26286. doi: 10.1038/srep 26286
20. Bahaa K, Noor G, Yousif Y. The Artificial Intelligence Approach for Diagnosis, Treatment Modelling in Orthodontic. In S. Naretto, editor, *Principles in Contemporary Orthodontics*, InTech (2011).
21. Ramesh AN, Kambhampati C, Monson JR, Drew PJ. Artificial intelligence in medicine. *Ann R Coll Surg Engl*. (2004) 86:334–8. doi: 10.1308/147870804290
22. Gubbi S, Hamet P, Tremblay J, Koch C, Hannah-Shmouni F. Artificial intelligence and machine learning in endocrinology and metabolism: the dawn of a new era. *Front Endocrinol*. (2019) 10:185. doi: 10.3389/fendo.2019.00185
23. Somashekhar SP, Sepúlveda MJ, Puglielli S, Norden AD, Shortliffe EH, Rohit Kumar C, Ramya Y. Watson for oncology and breast cancer treatment recommendations: agreement with an expert multidisciplinary tumor board. *Ann Oncology*. (2018) 29:418–23. doi: 10.1093/annonc/mdx781
24. Liu C, Liu X, Wu F, Xie M, Feng Y, Hu C. Using artificial intelligence (Watson for Oncology) for treatment recommendations amongst Chinese patients with lung cancer: feasibility study. *J Med Internet Res*. (2018) 20:e11087. doi: 10.2196/1108
25. Itoh H, Hayashi K, Miyashita K. Pre-emptive medicine for hypertension and its prospects. *Hypertens Res*. (2019) 42:301–5. doi: 10.1038/s41440-018-0177-3
26. Triberti S, Barello S. The quest for engaging AmI: patient engagement and experience design tools to promote effective assisted living. *J Biomed Inform*. 63:150–6. doi: 10.1016/j.jbi.2016.08.010
27. Davenport T, Kalakota R. The potential for artificial intelligence in healthcare. *Future Healthc J*. (2019) 6:94. doi: 10.7861/futurehosp.6-2-94
28. Triberti S, Durosini I, Curigliano G, Pravettoni G. Is explanation a marketing problem? the quest for trust in artificial intelligence and two conflicting solutions. *Public Health Genomics*. (2020). doi: 10.1159/000 506014. [Epub ahead of print].
29. Kalton A, Falconer E, Docherty J, Alevras D, Brann D, Johnson K. Multi-agent-based simulation of a complex ecosystem of mental health care. *J Med Syst*. (2016) 40:1–8. doi: 10.1007/s10916-015-0374-4
30. Hamet P, Tremblay J. Artificial intelligence in medicine. *Metabolism*. (2017) 69:S36–40. doi: 10.1016/j.metabol.2017. 01.011

31. Gunning D. *Explainable Artificial Intelligence (XAI): Program Update Novmeber 2017*. Defense Advanced Research Projects Agency (DARPA) (2017). Retrieved from: https://www.darpa.mil/attachments/XAIProgramUpdate.pdf
32. Edwards L, and Veale M. Enslaving the algorithm: from a "right to an explanation" to a "right to better decisions"?. *IEEE Secur Priv.* (2018) 16:43–54. doi: 10.1109/MSP.2018.2701152
33. Miller T. Explanation in artificial intelligence: insights from the social sciences. *Artif Intell.* (2019) 267, 1–38. doi: 10.1016/j.artint.2018.07.007
34. Miller T, Hower P, Sonenberg L. Explainable ai: beware of inmates running the asylum or: how i learnt to stop worrying and love the social and behavioural sciences. In: *Proc. IJCAI Workshop Explainable AI (XAI), 2017,*. (2017) pp. 36–42. doi: 10.1016/j.foodchem.2017.11.091
35. Vellido A. The importance of interpretability and visualization in machine learning for applications in medicine and health care. *Neural Comput Appl.* (2019) 1–15, doi: 10.1007/s00521-019-04051-w
36. Pravettoni G, Triberti S. *Il Medico 4.0*. Milan: EDRA. (2019).
37. Légaré F, and Thompson-Leduc P. Twelve myths about shared decision making. *Patient Educ Couns.* (2014) 96:281–6. doi: 10.1016/j.pec.2014.06.014
38. Barry MJ, Edgman-Levitan S. Shared decision making — the pinnacle of patient-centered care. *N Engl J Med.* (2012) 366, 780–1. doi: 10.1056/NEJMp1109283
39. Simpson M, Buckman R, Stewart M, Maguire P, Lipkin M, Novack D, Till J. Doctor-patient communication: the Toronto consensus statement. *BMJ.* (1991) 303:1385–7. doi: 10.1136/bmj.303.6814.1385
40. Stewart MA. Effective physician-patient communication and health outcomes: a review. *CMAJ.* (1995) 152, 1423–1433
41. Bragazzi NL. From P0 to P6 medicine, a model of highly participatory, narrative, interactive, and "augmented" medicine: Some considerations on Salvatore Iaconesi's clinical story. *Patient Prefer Adherence.* (2013) 7:353–9. doi: 10.2147/PPA.S38578
42. Müller-Engelmann M, Keller H, Donner-Banzhoff N, and Krones T. Shared decision making in medicine: the influence of situational treatment factors. *Patient Educ Couns.* (2011) 82:240–6. doi: 10.1016/j.pec.2010.04.028
43. Spencer KL. Transforming patient compliance research in an era of biomedicalization. *J Health Soc Behav.* (2018) 59. 170–84. doi: 10.1177/0022146518756860
44. Ford S, Schofield T, Hope T. What are the ingredients for a successful evidence-based patient choice consultation?: a qualitative study. *Soc Sci Med.* (2003) 56:589–602. doi: 10.1016/S0277-9536(02)00056-4
45. Marzorati C, and Pravettoni G. Value as the key concept in the health care system: how it has influenced medical practice and clinical decision-making processes. *J Multidiscip Healthc.* (2017) 10:101–6. doi: 10.2147/JMDH.S122383
46. Fioretti C, Mazzocco K, Riva S, Oliveri S, Masiero M, Pravettoni G. (2016). Research studies on patients' illness experience using the narrative medicine approach: a systematic review. *BMJ Open.* 6:e011220. doi: 10.1136/bmjopen-2016-011220
47. Chawla NV, Davis DA. Bringing big data to personalized healthcare: A patient-centered framework. *J Gen Intern Med.* (2013) 28:660–5. doi: 10.1007/s11606-013-2455-8
48. Corso G, Magnoni F, Provenzano E, Girardi A, Iorfida M, De Scalzi AM, et al. Multicentric breast cancer with heterogeneous histopathology: a multidisciplinary review. *Future Oncol.* (2020) 16:395–412. doi: 10.2217/fon-2019-0540
49. Riva S, Antonietti A, Iannello P, Pravettoni G. What are judgment skills in health literacy? A psycho-cognitive perspective of judgment and decision-making research. *Patient Prefer Adherence.* (2015) 9:1677–86. doi: 10.2147/PPA.S90207
50. Terry DJ, and Hogg MA. *Attitudes, Behavior, and Social Context: The Role of Norms and Group Membership*. Hove, UK: Psychology Press (1999).
51. Huryk LA. Factors influencing nurses' attitudes towards healthcare information technology. *J Nurs Manag.* (2010) 18:606–12. doi: 10.1111/j.1365-2834.2010.01084.x
52. Jacobs RJ, Iqbal H, Rana AM, Rana Z, Kane MN. (2017). Predictors of osteopathic medical students' readiness to use health information technology. *J Am Osteopath Assoc.* 117:773. doi: 10.7556/jaoa.2017.149
53. Kossman SP, Scheidenhelm SL. Nurses' perceptions of the impact of electronic health records on work and patient outcomes. *CIN – Comput Inform Nurs.* (2008) 26:69–77. doi: 10.1097/01.NCN.0000304775.40531.67
54. Weber S. A qualitative analysis of how advanced practice nurses use clinical decision support systems. *J Am Acad Nurse Pract.* (2007) 19:652–67. doi: 10.1111/j.1745-7599.2007.00266.x
55. Liberati EG, Ruggiero F, Galuppo L, Gorli M, González-Lorenzo M, Maraldi M, et al. What hinders the uptake of computerized decision support systems in hospitals? a qualitative study and framework for implementation. *Implemen Sci.* (2017) 12:113. doi: 10.1186/s13012-017-0644-2
56. Cho J, Park D, Lee HE. Cognitive factors of using health apps: systematic analysis of relationships among health consciousness, health information orientation, eHealth literacy, and health app use efficacy. *J Med Internet Res.* (2014) 16:e125. doi: 10.2196/jmir.3283
57. Kelley H, Chiasson M, Downey A, Pacaud D, Payton FC, Paré G, et al. The clinical impact of ehealth on the self-management of diabetes: a double adoption perspective. *J Assoc Inf Syst.* (2011) 12:208–34. doi: 10.17705/1jais.00263
58. Afra P, Bruggers CS, Sweney M, Fagatele L, Alavi F, Greenwald M, et al. Mobile software as a medical device (SaMD) for the treatment of epilepsy: development of digital therapeutics comprising behavioral and music-based interventions for neurological disorders. *Front Human Neurosci.* 12:171. doi: 10.3389/fnhum.2018.00171
59. Sverdlov O, van Dam J, Hannesdottir K, Thornton-Wells T. Digital therapeutics: an integral component of digital innovation in drug development. *Clin Pharmacol Ther.* (2018) 104:72–80, doi: 10.1002/cpt.1036
60. US Food and Drug Administration. *Proposed Regulatory Framework for Modifications to Artificial Intelligence/Machine Learning (AI/ML) Based Software as a Medical Device (SAMD)—Discussion Paper and Request for Feedback*. (2019). Available online at: https://www.fda.gov/files/medical%20devices/published/US-FDA-Artificial-Intelligence-and-Machine-Learning-Discussion-Paper.pdf
61. Pear Therapeutics Inc. *reSET for Substance Use Disorder*. (2018). Retrieved from: https://peartherapeutics.com/reset/ (accessed April 16, 2020).
62. Budney AJ, Borodovsky JT, Marsch LA, Lord SE. Technological innovations in addiction treatment. In: Danovitch I, Mooney L, editors. *The Assessment and Treatment of Addiction*. St. Louis, MO: Elsevier (2019). p. 75–90.
63. Cosson B, Graham E. I felt like a third wheel': Fathers' stories of exclusion from the 'parenting team. *J Family Studi.* (2012) 18:121–9. doi: 10.5172/jfs.2012.18.2-3.121
64. Clayton RB. The third wheel: The impact of Twitter use on relationship infidelity and divorce. *Cyberpsychol Behav Soc Netw.* (2014) 17:425–30. doi: 10.1089/cyber.2013.0570
65. Halpern D, Katz JE, Carril C. The online ideal persona vs. the jealousy effect: two explanations of why selfies are associated with lower-quality romantic relationships. *Telemat Inform.* (2017) 34:114–23. doi: 10.1016/j.tele.2016.04.014
66. Jochemsen H. (2008). Medical practice as the primary context for medical ethics. In D. Weisstub, G. Diaz Pintos Editors, *Autonomy and human rights in health care. An international perspective*. Dordrecht: Springer.
67. Charon R. Narrative medicine as witness for the self-telling body. *J Appl Commun Res.* 37:118–31. doi: 10.1080/00909880902792248
68. Smith SK, Dixon A, Trevena L, Nutbeam D, McCaffery KJ. Exploring patient involvement in healthcare decision making across different education and functional health literacy groups. *Soc Sci Med.* (2009) 69:1805–12. doi: 10.1016/j.socscimed.2009.09.056
69. Harris CR, Darby RS. Shame in physician-patient interactions: patient perspectives. *Basic App Soc Psychol.* (2009) 31:325–34. doi: 10.1080/01973530903316922
70. Kee JWY, Khoo HS, Lim I, and Koh MYH. (2017). Communication skills in patient-doctor interactions: learning from patient complaints. *Health Prof Educ.* 4:97–106. doi: 10.1016/j.hpe.2017.03.006
71. Miaoulis G, Gutman J, Snow MM. Closing the gap: the patient-physician disconnect. *Health Mark Q.* (2009) 26:56–8. doi: 10.1080/07359680802473547

72. Ozawa S, Sripad P. How do you measure trust in the health system? A systematic review of the literature. *Soc Sci Med.* 91:10–4. doi: 10.1016/j.socscimed.2013.05.005
73. Taber JM, Leyva B, Persoskie A. Why do people avoid medical care? a qualitative study using national data. *J Gen Intern Med.* 30:290–7. doi: 10.1007/s11606-014-3089-1
74. Rajkomar A, Dean J, Kohane I. Machine learning in medicine. *New Engl J Med.* (2019) 380:1347–58. doi: 10.1161/CIRCULATIONAHA.115.001593
75. Ross C, Swetlitz I, Thielking M. *IBM Pitched Watson as a Revolution in Cancer Care: it's Nowhere Close*. Boston. (2017) Retrieved from https://www.statnews.com/2017/09/05/watson-ibm-cancer/ (accessed April 16, 2020)
76. Nichols G. Terminal patient learns he's going to die from a robot doctor. (2019) Retrieved at: https://www.zdnet.com/article/terminal-patient-learns-hes-going-to-die-from-a-robot-doctor/ (accessed Mar. 26, 2019).
77. Durosini I, Tarocchi A, Aschieri F. Therapeutic assessment with a client with persistent complex bereavement disorder: a single-case time-series design. *Clin Case Stud*. (2017) 16:295–312. doi: 10.1177/1534650117693942
78. Rosner R, Pfoh G, Kotoučová M. Treatment of complicated grief. *Euro J Psychotraumatol*. (2011) 2:7995. doi: 10.3402/ejpt.v2i0.7995
79. Wiederhold BK. Can artificial intelligence predict the end of life... And do we really want to know? *Cyberpsychol. Behav. Soc. Netw*. 22:297. doi: 10.1089/cyber.2019.29149.bkw
80. Kedia SK, Ward KD, Digney SA, Jackson BM, Nellum AL, McHugh L, et al. "One-stop shop": lung cancer patients' and caregivers' perceptions of multidisciplinary care in a community healthcare setting. *Transl Lung Cancer Res*. (2015) 4:456–64. doi: 10.3978/j.issn.2218-6751.2015.07.10
81. Briet JP, Hageman MG, Blok R, Ring D. When do patients with hand illness seek online health consultations and what do they ask? *Clinical Orthopaedics and Related Research* (2014) 472:1246–50. doi: 10.1007/s11999-014-3461-9
82. Yeung RY, Leung GM, McGhee SM, Johnston JM. Waiting time and doctor shopping in a mixed medical economy. *Health Econ*. (2004) 13:1137–44. doi: 10.1002/hec.871

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 Public Health

ORIGINAL RESEARCH
published: 19 May 2020
doi: 10.3389/fpubh.2020.00178

Over- and Under-sampling Approach for Extremely Imbalanced and Small Minority Data Problem in Health Record Analysis

Koichi Fujiwara[1]*, *Yukun Huang*[2], *Kentaro Hori*[2], *Kenichi Nishioji*[3], *Masao Kobayashi*[3], *Mai Kamaguchi*[3] and *Manabu Kano*[2]

[1] *Department of Material Process Engineering, Nagoya University, Nagoya, Japan,* [2] *Department of Systems Science, Kyoto University, Kyoto, Japan,* [3] *Health Care Division, Japanese Red Cross Kyoto Daini Hospital, Kyoto, Japan*

A considerable amount of health record (HR) data has been stored due to recent advances in the digitalization of medical systems. However, it is not always easy to analyze HR data, particularly when the number of persons with a target disease is too small in comparison with the population. This situation is called the imbalanced data problem. Over-sampling and under-sampling are two approaches for redressing an imbalance between minority and majority examples, which can be combined into ensemble algorithms. However, these approaches do not function when the absolute number of minority examples is small, which is called the extremely imbalanced and small minority (EISM) data problem. The present work proposes a new algorithm called boosting combined with heuristic under-sampling and distribution-based sampling (HUSDOS-Boost) to solve the EISM data problem. To make an artificially balanced dataset from the original imbalanced datasets, HUSDOS-Boost uses both under-sampling and over-sampling to eliminate redundant majority examples based on prior boosting results and to generate artificial minority examples by following the minority class distribution. The performance and characteristics of HUSDOS-Boost were evaluated through application to eight imbalanced datasets. In addition, the algorithm was applied to original clinical HR data to detect patients with stomach cancer. These results showed that HUSDOS-Boost outperformed current imbalanced data handling methods, particularly when the data are EISM. Thus, the proposed HUSDOS-Boost is a useful methodology of HR data analysis.

Keywords: health record analysis, imbalanced data problem, boosting, over- and under-sampling, stomach cancer detection

OPEN ACCESS

Edited by:
Aek Palakorn Achananuparp,
Singapore Management University,
Singapore

Reviewed by:
Juan J. Rodriguez,
University of Burgos, Spain
Zhijun Yin,
Vanderbilt University Medical Center,
United States

****Correspondence:***
Koichi Fujiwara
fujiwara.koichi@
hps.material.nagoya-u.ac.jp

Specialty section:
This article was submitted to
Digital Public Health,
a section of the journal
Frontiers in Public Health

Received: *01 January 2020*
Accepted: *22 April 2020*
Published: *19 May 2020*

Citation:
Fujiwara K, Huang Y, Hori K, Nishioji K, Kobayashi M, Kamaguchi M and Kano M (2020) Over- and Under-sampling Approach for Extremely Imbalanced and Small Minority Data Problem in Health Record Analysis. Front. Public Health 8:178. doi: 10.3389/fpubh.2020.00178

1. INTRODUCTION

Digitalization of medical information is rapidly expanding due to advances in information technologies, and many governments and medical institutions worldwide are promoting the adoption of electronic health record (EHR) systems. An EHR system is a container for storing the collection of patient and population health information in a digital format and for sharing them over networks (1–3). A health record (HR) includes a wide range of items, such as patient demographics, medical history, medical images, prescription, laboratory test results, vital signs, and billing. According to the U.S. Department of Health and Human Services, more than 80 percent of hospitals in the U.S. had adopted EHR systems by 2014 (4). In Japan, 77.5% of 400-bed hospitals had introduced EHR systems by 2016, according to a survey by the Ministry of Health, Labour and Welfare (MHLW).

The use of EHR systems would improve the quality and efficiency of medical care, for example, by facilitating smooth transition of patients between hospitals, preventing unnecessary treatments and tests, and optimizing medical resources (5). Analysis of a significant amount of HR data will contribute to improving clinical decision-making, discovering hidden relationships between diseases and patient lifestyles, and predicting clinical endpoints (3).

It is beneficial to detect signs of a disease in its early stages without special examinations. From the viewpoint of machine learning, rare disease detection is formulated as a binary classification problem: persons with or without the disease. However, the majority of people will not contract a disease unless the target disease becomes prevalent, such as the cold or the flu. In this case, the objective data become imbalanced because the number of patients with the target disease is small while that of others is large.

Examples observed from the target rare event are referred to as minority class examples, and examples from frequent events are called majority class examples. Coping with the imbalance between majority and minority classes is a challenging problem for standard machine learning algorithms since most of them are designed for balanced data (6, 7). These algorithms that optimize model parameters based on classification accuracy tend to ignore the minority class. Consider a dataset with 99 majority examples and one minority example. A typical algorithm may classify all examples into the majority class because a classification accuracy of 99% is achieved. An accuracy of 99% means a highly-accurate classifier for the balanced data problem; however, such a classifier is unsatisfactory, since the detection of minority examples is of crucial importance in most imbalanced data problems. Although some methodologies for coping with the imbalanced data problem have been proposed, they do not always function well, particularly when the absolute number of minority examples is too small. In this work, such a situation is defined as an extremely imbalanced and small minority (EISM) data problem. HR data analysis frequently faces the EISM data problem.

The present work proposes a new boosting-based algorithm that combines heuristic under-sampling (HUS) and distribution-based sampling (DOS) to overcome the binary classification problem of EISM data, particularly for HR data analysis. The proposed method is referred to as boosting combined with HUS and distribution-based sampling (HUSDOS-Boost). HUS selects majority examples that may be important for subsequent weak classifier learning based on the former boosting results, and DOS generates multiple artificial minority examples whose variables are generated randomly in accordance with the distribution of the minority class. Through using these two sampling methods simultaneously, an artificially balanced training dataset is generated for weak classifier learning. In HUSDOS-Boost, multiple weak classifiers are constructed using classifications and regression trees (CARTs) (8). Finally, they are combined into a strong classifier for binary classification using the boosting method.

This paper is organized as follows: section 2 provides an overview of conventional algorithms for handling the imbalanced data problem. To cope with the EISM problem, HUSDOS-Boost is proposed in section 3. Section 4 evaluates the performance of the proposed HUSDOS-Boost through application to eight imbalanced datasets and discusses its characteristics. Section 5 reports the result of applying the proposed method to original clinical HR data. The objective here is to detect patients with stomach cancer from the HR data. Also, this section discusses variables relevant to stomach cancer development derived from the variable importance. Conclusion and future works are presented in section 6.

2. RELATED WORKS

Various methodologies for coping with the imbalanced data problem have been investigated because the imbalanced data problem is not limited to the medical field (9), and many real-world issues involve learning from imbalanced data, such as fraud detection (10) and oil spill detection (11). The imbalanced data problem arises due to characteristics of severe events like natural disasters. This phenomenon is sometimes called the power law (12).

This section explains existing methodologies for dealing with the imbalanced data problem, which are classified into six approaches–anomaly detection approach, cost-sensitive approach, rule-based approach, sampling approach, ensemble learning approach, and hybrid approach, which is a combination of the sampling approach and the ensemble learning approach.

2.1. Anomaly Detection Approach

One approach to deal with the imbalanced data problem is formulated as anomaly detection, which is also called one-class learning. One class support vector machine (OCSVM) and local outlier factor (LOF) are well-known anomaly detection algorithms (13, 14). Fujiwara et al. (15) used multivariate statistical process control (MSPC) for epileptic seizure prediction, which is a well-known anomaly detection method originally used in process control (16, 17). When interested in the discovery of hidden factors related to disease development from HR data, the importance of each variable to the outcome should be calculated. Such importance is not always calculated in an anomaly detection approach, although some methods have been proposed (18, 19).

2.2. Cost-Sensitive Approach

The main concept of cost-sensitive approaches is to introduce different miss-classification costs for different classes. For instance, if an algorithm incorrectly classifies a healthy person as a patient in a health check, the impact of misdiagnosis is not crucial. In contrast, a patient may lose an opportunity for treatment if he/she is diagnosed as healthy. In this example, the misclassification cost of the latter case is much higher than that of the former case. In general, the misclassification cost of the minority examples must be higher than that of the majority examples (20). Cost-sensitive support vector machine (C-SVM) is a well-known cost-sensitive algorithm, which introduces different costs for different classes into the support vector machine (SVM) (21).

2.3. Rule-Based Approach

Rule-based approaches find classification rules from the dataset. A major methodology of the rule-based approach is a decision tree. In the decision tree, a measure is needed to find the classification rules, of which information gain is widely used (22, 23). Some measures have been proposed in order to cope with the imbalance data problem. Liu et al. (24) proposed a class confidence proportion (CCP) measure which uses Fisher's exact test to prune branches that are not statistically significant. In addition, the rule-based approach can be combined with another machine learning method. Batuwita and Palade (25) proposed fuzzy-ruled SVM (FSVM) with the cost-sensitive approach, referred to as FSVM-CIL (FSVM with class imbalance learning), which copes well with the imbalanced data problem particularly when the data contains outliers.

2.4. Sampling Approach

The imbalanced numbers of examples between the majority class and the minority class are modified through sampling methods (9). Under-sampling deletes majority examples from the dataset so that the numbers of examples between different classes become balanced, of which random under-sampling (RUS) is a well-known method (26). Since under-sampling shrinks the data size, less time is necessary for learning. The disadvantage is that discarding majority examples may lead to losing useful information of the majority class.

Over-sampling is carried out to add minority examples to the dataset in order to achieve a balance, in which the existing minority examples are replicated, or artificial minority examples are generated. Random over-sampling (ROS) replicates the existing minority examples randomly and adds them to the dataset. However, it may cause overfitting because learning algorithms tend to focus on replicated minority examples. To avoid overfitting, over-sampling methods which generate artificial minority examples are preferred. Synthetic minority over-sampling technique (SMOTE) is a commonly used over-sampling method that randomly selects minority examples and creates artificial minority examples via random interpolation between the selected examples and their nearest neighbors (27). Some modifications of SMOTE for enhancing its performance by modifying minority example selection have been proposed. For instance, adaptive synthetic sampling (ADASYN) adaptively changes the number of artificial minority examples following the density of majority examples around the original minority example (28).

2.5. Ensemble Learning Approach

In order to use ensemble algorithms, like boosting and bagging, it is necessary to construct multiple weak classifiers by means of any learning algorithm and to integrate them into a final strong classifier. Although ensemble algorithms were not originally designed for handling imbalanced data problems, they perform relatively well in many imbalanced data problems (29). Random forest (RF) and Adaptive Boosting (AdaBoost) are well-known methods of ensemble algorithms (30–32). Moreover, these methods can calculate the importance of variables (33), which may contribute to discovering hidden factors of disease development in HR data analysis.

2.6. Hybrid Approach

Sampling approaches can be combined with ensemble learning algorithms, such as boosting and bagging, because ensemble learning algorithms tend to outperform other machine learning algorithms when dealing with the imbalanced data problem (9). Such combinations are called hybrid algorithms. Under-sampling or over-sampling methods for balancing classes are used for weak classifier learning in boosting or bagging. RUSBoost is a well-known hybrid algorithm that combines RUS and boosting (26). A hybrid approach method adopting a sampling method and hyper ensemble learning, which is referred to as hyperSMURF, has been proposed (34). Hyper ensemble learning is an meta-ensemble learning framework that combines classification results of multiple ensemble learning classifiers.

However, hybrid algorithms do not always function well, particularly when the objective data is EISM.

3. HUSDOS-BOOST

The present work proposes a new method for coping with the imbalanced data problem, in particular, with the EISM data problem. The proposed HUSDOS-Boost combines HUS and distribution-based over-sampling (DOS) with the AdaBoost framework.

To deal with the EISM problem, such as detecting rare diseases from HR data, both under-sampling and over-sampling can be used. Although a large number of minority examples need to be generated by over-sampling, such manipulation may lead to overfitting because many similar minority examples exist in the dataset. To avoid overfitting, under-sampling, which reduces the number of majority examples, should be used in addition to over-sampling so that a class balance is achieved with the generation of a small number of artificial minority examples.

Let $S = \{(\boldsymbol{x}_n, y_n)\}(n = 1, \cdots, N)$ be the dataset and $\boldsymbol{x}_n$ and $y_n = \{-1, 1\}$ denote variables and class labels, respectively. In the imbalanced data, $S^{maj} = \{(\boldsymbol{x}_n, y_n)|y_n = 1\}$ and $S^{min} = \{(\boldsymbol{x}_n, y_n)|y_n = -1\}$ are the majority and the minority datasets, respectively, and $S = S^{maj} \cup S^{min}$. $N^{maj} = |S^{maj}|$.

3.1. AdaBoost

Although there are some variations in the algorithms in the AdaBoost framework, AdaBoost.M1 is described here. The present work aims to detect a specific disease from HR data, which is formulated as a binary classification problem. In this case, AdaBoost.M1 and AdaBoost.M2 result in the same algorithm, and the former is simpler than the latter (35).

A procedure of AdaBoost.M1 is described in Algorithm 1. In step 1, the boosting weights of each example, $D_{1,n}(n = 1, \cdots, N)$, are initialized to $1/N$. After initialization, weak classifier learning is repeated in steps 2–8. Step 3 trains the tth weak classifier w_t so that the following objective function J_t is minimized:

$$J_t = \sum_{n=1}^{N} D_{t,n} I(h_{t,n} \neq y_n) \tag{1}$$

where $I(h_{n,t} \neq y_n)$ is an indicator function which returns 1 if $h_{n,t} \neq y_n$ and 0 otherwise. The error ε_t is calculated in steps 4 and 5. Steps 6 and 7 update a parameter β_t and the boosting weights $D_{t,n}$:

$$D_{t+1,n} = \frac{D_{t,n}}{Z_t} \times \begin{cases} \beta_t & \text{if} \quad h_{t,n} = y_n \\ 1 & \text{otherwise} \end{cases} \tag{2}$$

where Z_t is a normalization constant. After T iterations, the final classifier $H(\boldsymbol{x})$ is built as a weighted vote of the T weak classifiers as follows:

$$H(\boldsymbol{x}) = \arg\max_{y \in Y} \sum_{t\,:\,h_t = y} \log(1/\beta_t). \tag{3}$$

3.2. Heuristic Under-sampling

Although random under-sampling (RUS) randomly extracts a part of the majority examples for weak classifier learning (26), the drawback is that it does not consider the contribution that each majority example makes to the classification.

The proposed HUS selects majority examples according to sampling weights $SW_{t,n}(t = 1, \cdots T; n = 1, \cdots, N^{maj})$ which are updated based on the estimation results in each boosting iteration. The initial sampling weight $SW_{1,n}$ for the majority examples $\boldsymbol{x}_m \in S^{maj}$ is set to $1/N^{maj}$. After the tth boosting iteration, HUS updates the sampling weights $SW_{t,n}$ based on the tth estimation result $h_{t,n} = w_t(\boldsymbol{x}_n)$ as follows:

$$\begin{aligned} SW_{t+1,n} &= \frac{SW_{t,n}}{Z_{SW_t}} \\ &\times \begin{cases} \beta_t & \text{if } \boldsymbol{x}_n \in \hat{S}_t^{maj} \wedge h_{t,n} = y_m \\ 1/\beta_t & \text{if } \boldsymbol{x}_n \in \hat{S}_t^{maj} \wedge h_{t,n} \neq y_m \\ 1 & \text{if } \boldsymbol{x}_n \in S^{maj} \wedge \boldsymbol{x}_n \notin \hat{S}_t^{maj} \end{cases} \end{aligned} \tag{4}$$

where $\hat{S}_t^{maj}$ is the tth learning set sampled from S^{maj}, and Z_{SW_t} is a normalization constant.

Algorithm 1: AdaBoost.M1

```
1: Initialize the boosting weights $D_{n,1} = 1/N$ for $\boldsymbol{x}_n \in S$.
2: for $t = 1, \dots, T$ do
3:     Train the $t$th weak classifier $f_t$ so as to minimize $J_t$.
4:     Get estimate of $\boldsymbol{x}_n \in S$: $h_{t,n} = f_t(\boldsymbol{x}_n)$.
5:     Calculate the error of $h_{t,n}$, $\varepsilon_t$:
           $\varepsilon_t = \sum_{n=1}^{N} D_{t,n} I(h_{t,n} \neq y_n)$
6:     Set $\beta_t = \varepsilon_t/(1 - \varepsilon_t)$.
7:     Update the boosting weights $D_{t+1,n}$ using Eq.(2).
8: end for
9: return The final classifier $H(\boldsymbol{x})$.
```

This update rule means that the sampled and misclassified majority examples have a higher probability of being sampled in the subsequent training set $\hat{S}_{t+1}^{maj}$, while the sampled and correctly classified examples have a lower probability of being sampled. That is, majority examples that may be important for improving classification performance tend to be sampled for the subsequent weak classifier learning. Note that the sampling weights $SW_{t,n}$ are different from the boosting weights $D_{t+1,n}$, although their update rules use the same parameter β_t.

We refer to a method in which the random under-sampling in RUSBoost is replaced with HUS as HUSBoost.

3.3. Distribution-Based Over-sampling

Over-sampling methods that generate artificial minority examples increase the amount of information for weak classifier learning. This study proposes distribution-based over-sampling (DOS), which generates artificial values for the variables based on their distributions.

Categorical and continuous variables are considered here. Categorical variables are generated by following the proportion of each attribute in the minority class, $p_k = N_k/N_a$, where N_a and N_k are the number of examples in the minority class and the number of examples that have the attribute k, respectively. For example, it is assumed that the number of "male" is 15 and that of "female" is 9 in "gender," and the generated values in "gender" have a probability of $15/24$ of being "male" and $9/24$ of being "female."

Continuous variables are generated by following the continuous distribution estimated from the minority examples. When we assume that a variable "height" follows the Gaussian distribution $N(\mu, \sigma^2)$, its mean μ and variance σ^2 need to be estimated. Then, artificial values for 'height' are generated by following $N(\mu, \sigma^2)$.

Correlated variables may be generated by chance in the process of over-sampling, and such samples may cause multicollinearity in multiple regression (36). The multicollinearity problem is a phenomenon in which the estimated regression coefficients in a multiple regression model greatly fluctuate in response to small changes in training data when there is correlation among input variables. The regression coefficients are estimated using the normal equation: $\boldsymbol{b} = (\boldsymbol{X}^T\boldsymbol{X})^1\boldsymbol{X}\boldsymbol{y}$, where $\boldsymbol{X}$ is an input matrix and $\boldsymbol{y}$ is an output vector. The matrix $(\boldsymbol{X}^T\boldsymbol{X})$ becomes ill-conditioned when there is correlation among input variables, which lead to unstable inverse matrix calculation (37). On the other hand, the learning process of CART does not contain the inverse matrix calculation. Thus, the proposed HUSDOS-Boost avoids the multicollinearity problem even if the correlated variables are generated by over-sampling.

3.4. HUSDOS-Boost

Algorithm 2 shows the proposed HUSDOS-Boost algorithm, which combines AdaBoost.M1 with both HUS and DOS. HUSDOS-Boost with AdaBoost.M1 can be easily modified to an algorithm using AdaBoost.M2.

In step 1, the boosting weights of each example $D_{1,n}$ and the sampling weights of each majority example $SW_{1,n}$ are initialized to $1/N$ and $1/N^{maj}$, respectively. After initialization, T weak classifiers are iteratively trained in steps 2–12. In step 3, HUS is applied to select N_u majority examples for the tth majority training set $\hat{S}_t^{maj}$. On the other hand, DOS generates N_o artificial minority examples and adds them to S^{min} to construct the tth minority training set $\hat{S}_t^{min}$ in step 4. The numbers of selected majority examples by HUS and added minority examples by DOS, N_u and N_o, should be determined by considering the desired ratio of the majority examples to the minority examples. After the tth training set $\hat{S}_t$ is constructed, the tth weak classifier is trained in step 6. Note that the range of summation in the objective function is modified from Equation (1) in Algorithm 1:

$$\hat{J}_t = \sum_{n|y_n \in \hat{S}_t} D_{t,n} I(h_{t,n} \neq y_n). \tag{5}$$

The tth error ε_t is calculated in steps 7–8. The following steps 9-11 update the parameter β_t, the sampling weights $SW_{t+1,n}$, and the boosting weights $D_{t+1,n}$. After T iterations, the final hypothesis $H(\boldsymbol{x})$ is built as Equation (3).

3.5. Classification and Regression Tree

Although any learning algorithm can be used for the weak classifier in the proposed HUSDOS-Boost, a classification and regression tree (CART) (8) is adopted in this work. In CART, variable importance can be obtained.

A CART model is a binary tree that is obtained by splitting a variable set into two variable subsets recursively so that the cost function for misclassification is minimized. In addition, some leaf nodes are pruned after tree construction to obtain simple tree structures. CART uses the Gini coefficient as the cost function, which is an indicator of uniformity of data distribution. The Gini coefficient of the rth node, $I_G(r)$, is defined as follows:

$$I_G(r) = 1 - \sum_{k=1}^{K} \Big(\frac{n_r^{\{k\}}}{N_r} \Big) \tag{6}$$

where N_r and $n_r^{\{k\}}$ are the numbers of all examples and examples belonging to class k, respectively. K is the number of classes. The decrease in the Gini coefficient due to the splitting of the rth node, $\Delta I_G(r)$, is expressed as

$$\Delta I_G(r) = I_G(r) - \sum_{l=1,2} w_{r_l} I_G(r_l). \tag{7}$$

$I_G(r_l)(l = 1, 2)$ are the Gini coefficients of the child nodes of the rth node. w_{r_l} is defined as $w_{r_l} = N_{r_l}/N_r$, where N_{r_l} denotes the number of examples in the lth child node. The split that gives the largest decrease should be searched. Thus, $\Delta I_G(r)$ also indicates the variable importance for classification in CART (32).

Since a strong classifier is the weighted sum of multiple CART models in HUSDOS-Boost, the variable importance of the pth variable, VI_p, is defined as the weighted sum of the decreases due to the pth variable splitting:

$$\mathrm{VI}_p = \frac{1}{Z_{\mathrm{VI}}} \sum_t \log(1/\beta_t) \Delta I_G^t(p) \tag{8}$$

where $\Delta I_G^t(p)(t = 1, \cdot, T)$ is the Gini coefficient decrease due to the pth variable splitting in the tth CART model, and Z_{VI} is a normalization constant.

Algorithm 2: HUSDOS-Boost with AdaBoost.M1

```
1:  Initialize the boosting weights $D_{n,1} = 1/N$ for $\boldsymbol{x}_n \in S$, and
    the sampling weights $SW_{1,n} = 1/N^{maj}$ for $\boldsymbol{x}_n \in S^{maj}$.
2:  for $t = 1, \ldots, T$ do
3:      Apply HUS with $SW_{t,n}$ to $S^{maj}$ to generate $\hat{S}_t^{maj}$ with a size
        $N_u$.
4:      Apply DOS to $S^{min}$ to generate $\hat{S}_t^{min}$ with a size $N_o$, where
        $S^{min} \subset \hat{S}_t^{min}$.
5:      $\hat{S}_t = \hat{S}_t^{maj} \cup \hat{S}_t^{min}$.
6:      Train the $t$th weak classifier $f_t$ from $\hat{S}_t$ so as to minimize $\hat{J}_t$.
7:      Get hypothesis of $\boldsymbol{x}_n \in S$: $h_{t,n} = f_t(\boldsymbol{x}_n)$.
8:      Calculate the error of $h_{t,n}$, $\varepsilon_t$:
              $\varepsilon_t = \sum_{n : h_{t,n} \neq y_n} D_{t,n}$.
9:      Set $\beta_t = \varepsilon_t/(1 - \varepsilon_t)$.
10:     Update the boosting weights $D_{t+1,n}$ by Eq.(2).
11:     Update the sampling weights $SW_{t+1,n}$ by Eq.(4).
12: end for
13: return The final hypothesis $H(\boldsymbol{x})$.
```

4. CASE STUDY

This section investigates the performance and the characteristics of the proposed HUSDOS-Boost through its application to eight imbalanced datasets collected from the UCI Machine Learning repository (38). In this case study, random forest (RF), AdaBoost, SMOTE, ADASYN, RUSBoost, HUSBoost were tested for comparison.

4.1. Datasets

This case study used the following eight imbalanced datasets, which cover a wide variety of data sizes, imbalance ratios of the majority class to the minority class, and application domains.

- **Covertype**: Dataset for forest cover type estimation based on cartographic data, which consists of seven classes (27). "Ponderosa Pine" and "Cottonwood/Willow" were selected as the majority and minority classes.
- **Satimage**: Dataset for soil type classification from multi-spectral image data measured by a satellite (27). The smallest class "red soil" was the minority class, and other classes were considered the majority class.
- **Segment**: Dataset for object type prediction from outdoor image segmentation data (26). There are five classes, and the number of examples in each class is the same. "brick face" was selected as the minority class, and the rest was considered the majority class.

TABLE 1 | Dataset Characteristics.

Dataset	#Var	#Minority	#Majority	Ratio [%]
Covertype	54	2,747	35,754	7.13
Satimage	19	626	5,809	9.73
Segment	36	330	1,980	14.3
Pageblocks	10	115	5,358	2.10
E. coli	7	77	259	22.9
CTG	21	53	2,073	2.56
Abalone	8	42	689	5.75
Yeast	8	30	1,464	1.35

- **Pageblocks**: Dataset for block type classification of a document page layout, which consists of five classes. "graphic" with 115 examples was selected as the minority class, and the rest was considered the majority class.
- ***E. coli***: Dataset for protein localization site prediction consisting of eight classes. "Inner membrane without signal sequence" was the minority class, and the others were considered the majority class (39).
- **CTG**: Dataset of fetal heart rate (FHR) prediction from cardiotocography. There are ten types of FHR, and "type 3," whose size is the smallest, was selected as the minority class, and the rest were considered the majority class.
- **Abalone**: Dataset for abalone age estimation using physical measurements of an abalone. The ages of the abalones range from 1 to 29 in the dataset. The ages of 9 and 18 were selected as the majority and the minority classes, respectively (40).
- **Yeast**: Dataset for predicting cellular localization sites, which consists of ten classes (27). The class "VAC" with only 30 examples was chosen as the minority class, and others were considered the majority class.

Table 1 shows the characteristics of eight datasets, in which #Var, #Minority, and #Majority denote the numbers of input variables, minority examples, and majority examples in each dataset, respectively, and Ratio is their imbalance ratio: #Minority/(#Majority + #Minority). Note that datasets in **Table 1** are sorted in descending order of #Minority.

4.2. Experimental Procedure

The classification performances of RF, AdaBoost, SMOTE, ADASYN, RUSBoost, HUSBoost, hyperSMURF, and the proposed HUSDOS-Boost were evaluated using the imbalanced datasets described in section 4.1.

In SMOTE, the number of artificial minority examples generated by over-sampling was the same as the original number of majority examples for obtaining a perfectly balanced dataset, and a CART model was constructed. RUSBoost and HUSBoost sampled the same number of majority examples as that of minority examples by under-sapling. In the proposed HUSDOS-Boost, the number of artificial minority examples generated by DOS was the same as the original number of minority examples, and the number of sampled majority examples by HUS was twice that of the original minority examples. Thus, $N_u = N_o =$ #Minority in steps 3–4 in Algorithm 2. The weak classifier used in RF, AdaBoost, RUSBoost, HUSBoost, and HUSDOS-Boost was a CART model, and the maximum number of their constructed weak classifiers was 100. hyperSMURF used RF for hyper ensemble learning.

Each dataset was randomly divided into ten subsets, of which nine were used for modeling while the remaining one was used for validation. Modeling and validation were repeated ten times so that all subsets became the validation dataset once. The above procedure was repeated ten times for precise performance evaluation.

The computer configuration used in this case study was as follows: CPU: Intel Core i7-9700K (3.60GHz × 8 cores), RAM: 32GB, OS: Windows 10 Pro (64 bit), and the R language was used.

4.3. Performance Metrics

In standard machine learning problems, the overall accuracy is a metric for performance evaluation: however, it is not appropriate in this case study because an accuracy of 99% is achieved when the imbalance ratio is 1:99 and a stupid classifier discriminates all of the examples as the majority class.

The geometric mean (G-mean) of the sensitivity and the specificity was used in this work:

$$\mathrm{G_{mean}} = \sqrt{\text{sensitivity} \times \text{specificity}.} \tag{9}$$

The G-mean measures the classification performance of a classifier for minority class examples as well as majority class examples, simultaneously. A low value of the G-mean indicates that the classifier is highly biased toward one class and vice-versa. Thus, the G-mean is an appropriate metric for evaluating the imbalanced data problem.

In addition, an area under the curve (AUC) of a receiver operating characteristic (ROC) curve and the area under the precision-recall curve (AUPRC) were used for evaluating the averaged performances of classifiers.

The average CPU time per modeling calculation was measured for each method.

4.4. Results and Discussion

Table 2 shows the sensitivity, the specificity,the G-mean, AUC, and AUPRC of each method in eight imbalanced datasets. The bold fonts indicate the best scores in the seven algorithms.

RF and AdaBoost, which do not employ sampling methods, achieved high specificities while their sensitivities were lower than the three algorithms with sampling methods, which resulted in low G-means. SMOTE, which uses over-sampling and which are not an ensemble algorithm, performed modestly. ADASYN improved the performance of SMOTE, which showed that adaptive changes in the number of artificial minority examples is certainly effective. These results indicate that sampling method are effective in the imbalanced data problem.

RUSBoost, which uses random under-sampling and boosting, achieved the highest G-means in four datasets whose number of minority samples are the first to the fourth largest among the eight datasets. However, AUC and AUPRC of RUSBOOST achieved modest values, which means that its

TABLE 2 | Performances of seven methods.

Dataset	Metrics	RF	AdaBoost	SMOTE	ADASYN	RUSBoost	HUSBoost	hyperSMURF	HUSDOSBoost
Cover type	Sensitivity	0.65±0.02	0.87±0.01	0.71±0.02	0.74±0.01	0.98±0.00	0.81±0.01	**0.99±0.00**	0.83±0.01
	Specificity	**1.00±0.00**	0.99±0.00	0.97±0.00	0.92±0.00	0.96±0.00	0.99±0.00	0.84±0.01	0.97±0.00
	G-mean	0.81±0.02	0.93±0.00	0.83±0.01	0.82±0.01	**0.97±0.00**	0.90±0.00	0.91±0.01	0.90±0.00
	AUC	0.99±0.00	**1.00±0.00**	0.87±0.00	0.92±0.01	0.99±0.00	0.99±0.00	0.98±0.00	0.98±0.00
	AUPRC	0.90±0.00	**0.96±0.00**	0.53±0.01	0.43±0.01	0.93±0.01	0.92±0.00	0.83±0.01	0.87±0.01
Satimage	Sensitivity	0.52±0.02	0.63±0.02	0.68±0.02	0.89±0.01	0.91±0.02	0.70±0.01	**0.94±0.01**	0.75±0.01
	Specificity	**0.99±0.00**	0.98±0.00	0.93±0.01	0.80±0.01	0.86±0.00	0.95±0.00	0.83±0.00	0.93±0.00
	G-mean	0.72±0.01	0.79±0.01	0.79±0.01	0.84±0.01	**0.89±0.01**	0.82±0.01	0.88±0.00	0.83±0.00
	AUC	**0.96±0.00**	0.78±0.01	0.87±0.01	0.85±0.01	0.55±0.12	0.95±0.00	**0.96±0.00**	0.94±0.00
	AUPRC	0.78±0.01	0.18±0.00	0.47±0.02	0.32±0.01	0.09±0.03	**0.75±0.01**	0.74±0.01	0.71±0.00
Segment	Sensitivity	**0.99±0.00**	**0.99±0.00**	0.96±0.01	0.91±0.01	**0.99±0.01**	**0.99±0.01**	**0.99±0.00**	**0.99±0.00**
	Specificity	**1.00±0.00**	**1.00±0.00**	0.99±0.00	0.99±0.00	0.99±0.00	1.00±0.00	0.99±0.00	0.99±0.00
	G-mean	**0.99±0.00**	**0.99±0.00**	0.98±0.01	0.94±0.01	**0.99±0.00**	**0.99±0.00**	**0.99±0.00**	**0.99±0.00**
	AUC	**1.00±0.00**	**1.00±0.00**	0.98±0.01	0.96±0.01	**1.00±0.00**	**1.00±0.00**	**1.00±0.00**	**1.00±0.00**
	AUPRC	**1.00±0.00**	**1.00±0.00**	0.93±0.03	0.89±0.01	**1.00±0.00**	**1.00±0.00**	**1.00±0.00**	**1.00±0.00**
Pageblocks	Sensitivity	0.65±0.03	0.65±0.06	0.71±0.06	0.87±0.02	**0.95±0.03**	0.90±0.02	0.90±0.02	0.90±0.02
	Specificity	**1.00±0.00**	**1.00±0.00**	0.99±0.00	0.92±0.01	0.94±0.01	0.97±0.00	0.97±0.00	0.97±0.00
	G-mean	0.81±0.02	0.80±0.04	0.84±0.03	0.89±0.01	**0.94±0.01**	0.93±0.01	0.93±0.01	0.93±0.01
	AUC	0.97±0.01	0.98±0.00	0.91±0.04	0.93±0.02	0.58±0.11	**0.99±0.00**	**0.99±0.00**	**0.99±0.00**
	AUPRC	0.80±0.01	0.76±0.02	0.55±0.07	0.26±0.04	0.03±0.01	0.75±0.02	**0.77±0.01**	0.74±0.03
Ecoil	Sensitivity	0.77±0.03	0.76±0.04	0.87±0.06	0.92±0.02	0.91±0.06	0.91±0.04	**0.97±0.00**	0.92±0.03
	Specificity	0.94±0.01	0.94±0.02	0.87±0.02	0.87±0.02	0.87±0.02	**0.89±0.01**	0.76±0.02	**0.89±0.01**
	G-mean	0.85±0.02	0.84±0.03	0.87±0.03	**0.90±0.01**	0.89±0.03	**0.90±0.02**	0.86±0.01	**0.90±0.02**
	AUC	0.95±0.01	0.95±0.01	0.93±0.03	0.92±0.01	0.95±0.02	0.95±0.01	0.95±0.01	**0.96±0.01**
	AUPRC	0.86±0.03	0.86±0.03	0.77±0.09	0.74±0.04	0.85±0.04	0.85±0.03	0.83±0.04	**0.87±0.03**
CTG	Sensitivity	0.55±0.05	0.67±0.07	0.65±0.08	0.79±0.05	0.93±0.07	0.89±0.02	0.87±0.04	**0.92±0.02**
	Specificity	**1.00±0.00**	1.00±0.00	0.98±0.01	0.96±0.00	0.91±0.01	0.96±0.00	0.98±0.00	0.96±0.00
	G-mean	0.74±0.03	0.82±0.04	0.80±0.05	0.87±0.03	0.92±0.03	0.93±0.01	0.92±0.02	**0.94±0.01**
	AUC	0.99±0.01	0.96±0.08	0.92±0.04	0.86±0.04	0.70±0.11	0.97±0.00	**0.98±0.00**	**0.98±0.01**
	AUPRC	0.78±0.02	0.65±0.55	0.45±0.06	0.42±0.08	0.09±0.04	0.72±0.04	0.68±0.03	0.73±0.03
Abalone	Sensitivity	0.15±0.05	0.37±0.05	0.46±0.10	0.60±0.09	**0.69±0.07**	0.57±0.02	0.76±0.03	0.67±0.10
	Specificity	**1.00±0.00**	0.99±0.01	0.92±0.02	0.83±0.01	0.74±0.03	0.87±0.02	0.80±0.01	0.86±0.01
	G-mean	0.38±0.07	0.61±0.04	0.65±0.07	0.71±0.05	0.72±0.04	0.70±0.02	**0.78±0.02**	0.76±0.05
	AUC	0.82±0.02	0.81±0.05	0.74±0.08	0.73±0.05	0.67±0.11	0.82±0.01	0.83±0.01	**0.84±0.03**
	AUPRC	0.44±0.05	0.44±0.08	0.27±0.11	0.24±0.08	0.19±0.08	0.37±0.06	0.40±0.07	**0.42±0.06**
Yeast	Sensitivity	0.00±0.00	0.03±0.04	0.07±0.03	0.23±0.10	0.60±0.10	0.43±0.10	0.31±0.08	**0.56±0.08**
	Specificity	1.00±0.00	1.00±0.00	0.98±0.01	0.85±0.02	0.57±0.03	0.84±0.01	**0.90±0.01**	0.74±0.01
	G-mean	0.00±0.00	0.14±0.14	0.25±0.07	0.44±0.10	0.58±0.04	0.60±0.07	0.53±0.07	**0.64±0.04**
	AUC	0.62±0.02	0.62±0.06	0.59±0.14	0.55±0.04	0.54±0.08	0.67±0.02	**0.70±0.02**	0.66±0.03
	AUPRC	0.05±0.03	0.08±0.05	0.06±0.04	0.03±0.01	0.03±0.03	**0.09±0.02**	0.05±0.01	0.06±0.04

averaged performance is not so high. HUS-Boost that combines HUS and boosting kept rather high AUC and AUPRC when the imbalance ratio of a dataset was low although other performance metrics were modest. This indicated that HUS was effective when the imbalance ratio is low. hyperSMURF, which adopts hyper ensemble learning, achieved high performance on average even when the number of minority examples was rather small.

The proposed HUSDOS-Boost, which utilizes both over-sampling and under-sampling in addition to boosting, achieved the best G-means in five datasets whose numbers of minority samples are the third to the eighth largest. These results suggest that HUSDOS-Boost achieves higher performance than RUSBoost and HUSBoost when the imbalance ratio of a dataset is not particularly low, but the absolute number of minority examples contained in a dataset is minimal. In addition, HUSDOS-Boost also kept high AUC and AUPRC when the imbalance ratio was low, which means that its averaged performance does not deteriorate when the number of minority examples is minimal. Thus, the use of both HUS and distribution-based over-sampling is certainly effective.

To verify this point, we compared RUSBoost and HUSDOS-Boost through another experiment using datasets with intentionally reduced minority examples. The minority examples in Covertype, Satimage, Segment, and Pageblocks, which have more than 100 minority examples, were eliminated randomly. The numbers of reduced minority examples in these datasets were 20, 30, 40, 50, 60, and 70. The procedure described in section 4.2 was applied to these reduced datasets. **Figure 1** shows the G-means of RUSBoost and HUSDOS-Boost for the reduced datasets. The proposed HUSDOS-Boost performed better than RUSBoost when the number of minority examples was 20 and 30 regardless of #Var, and the performance of HUSDOS-Boost was almost the same as RUSBoost when the number of minority examples was more than 40. Thus, over-sampling, as well as under-sampling, should be used when the number of minority examples is small. It is concluded that the proposed HUSDOS-Boost is more appropriate than RUSBoost for solving the EISM data problem.

To evaluate the effects of the number of examples generated by over-sampling N_o, we investigated the performance of HUSDOS-Boost and SMOTE with $N_o = \{2,3,4,5,6\}\times$ #Minority using eight datasets. The number of examples sampled by under-sampling N_u is fixed to #Minority. **Figure 2** illustrates the G-means of HUSDOS-Boost and SMOTE calculated for each N_o. The $\triangledown$ marks in the figures denote the pairs for which a significant difference was not confirmed by the t-test ($\alpha = 0.05$). These results show that the proposed HUSDOS-Boost achieved a higher performance than SMOTE regardless of which N_o was selected, and that the performance did not improve even when the number of artificial examples generated by over-sampling became large in most cases, which indicates that an excessive number of similar minority examples do not contribute to classifier learning.

The influence of the number of majority examples sampled by under-sampling on classifier learning was checked. We tested HUSDOS-Boost and RUSBoost with $N_u = \{2,3,4,5,6\}\times$ #Minority using eight datasets and $N_o =$ #Minority. Their G-means are illustrated in **Figure 3**, which shows that their performances deteriorated as N_u became large. Thus, the numbers of majority examples used for classifier learning should be balanced with the numbers of minority examples.

The average CPU times of each of the seven methods required for one strong classifier learning are reported in **Table 3**. In almost all datasets, RF was the fastest, in which multiple CARTs are constructed using a bagging approach in parallels. SMOTE was the second-fastest. Although SMOTE roughly doubled the number of examples for learning through over-sampling in this case study, just one CART model was built. Thus, the total amount of calculation was not significant. AdaBoost performed the worst because it uses all examples for weak classifier learning, and the learning process has to be performed in series. In hyperSMURF, the CPU times did not decrease so much when the number of examples became small because it constructed multiple RFs as hyper ensemble learning. The CPU times of RUSBoost were modest. Although RUSBoost is based on boosting in the same manner as AdaBoost, the number of examples used for weak classifier learning is significantly reduced due to under-sampling. Since RUSBoost was much faster than HUSBoost and the computational burdens of HUSBoost and

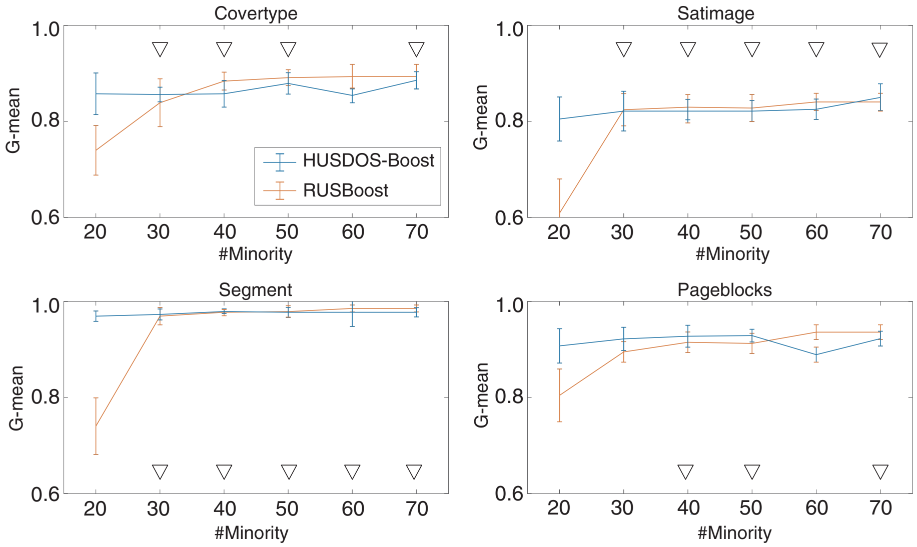


FIGURE 1 | G-means of HUSDOS-Boost and RUSBoost vs. #Minority.

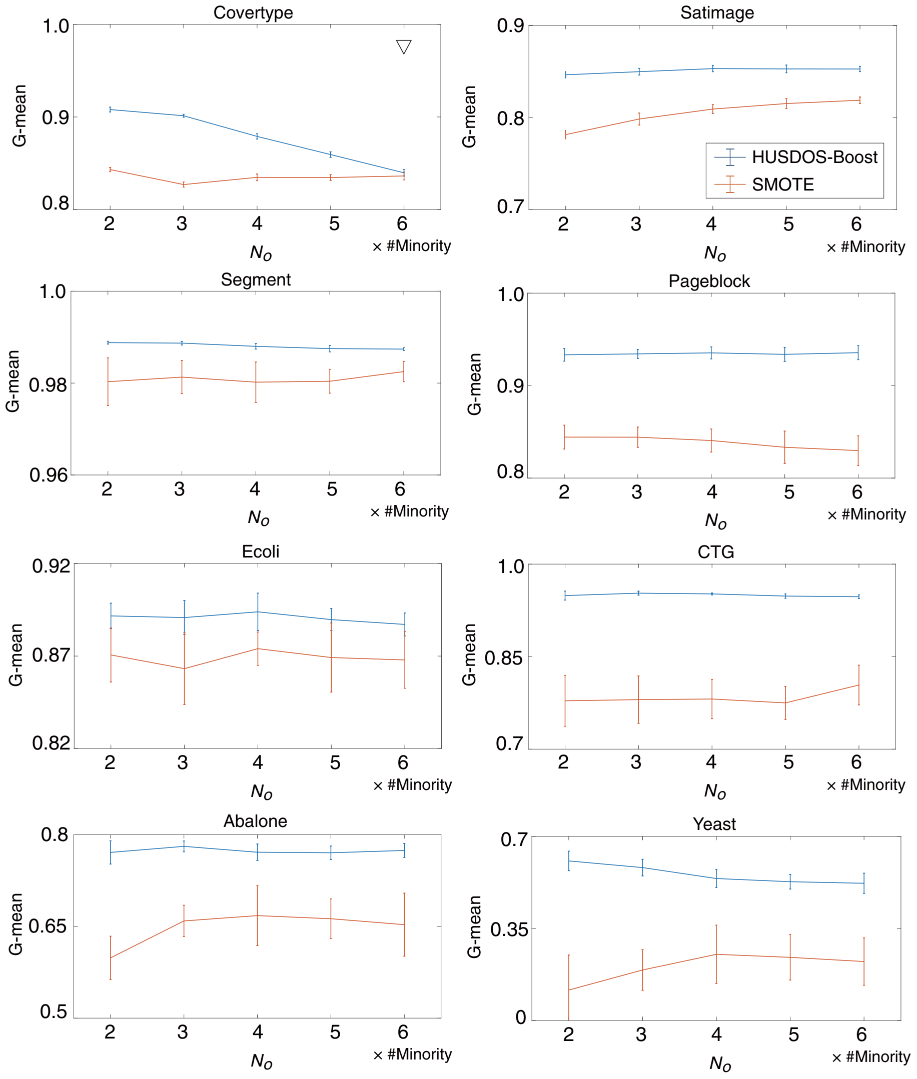


FIGURE 2 | G-means of HUSDOS-Boost and SMOTE vs. N_o.

HUSDOS-Boost were almost at the same level, heuristics undersampling requires heavy computational burden although it is more effective than random under-sampling for the imbalanced data problem.

The variable importance is discussed in the following section 5.

5. STOMACH CANCER SCREENING FROM CLINICAL HEALTH RECORD DATA

Early detection of stomach cancer is essential for its prognosis; however, stomach cancer detection is a typical EISM data problem. The lifetime morbidity risk of stomach cancer is 11% in males and 5% in females, and newly diagnosed patients per year is about 0.1–0.2% of the population in Japan. Hence, the number of patients with stomach cancer in the HR data is small, while those without stomach cancer is large. Although it is challenging to find stomach cancer at early stages due to lack of subjective symptoms, stomach cancer detection from HR data would be beneficial. The 5-year survival rate of stomach cancer is 82% for stage I while it is 8% for stage IV in Japan.

This section reports the result of applying the proposed HUSDOS-Boost to original clinical HR data to detect patients with stomach cancer. In addition, possible factors of stomach cancer development estimated by the variable importance of HUSDOS-Boost are discussed.

5.1. Health Examination Data

The clinical HR data were collected from the Japanese Red Cross Kyoto Daini Hospital, which provides comprehensive health examination menus. The Research Ethics Committee of the Japanese Red Cross Kyoto Daini Hospital approved the use

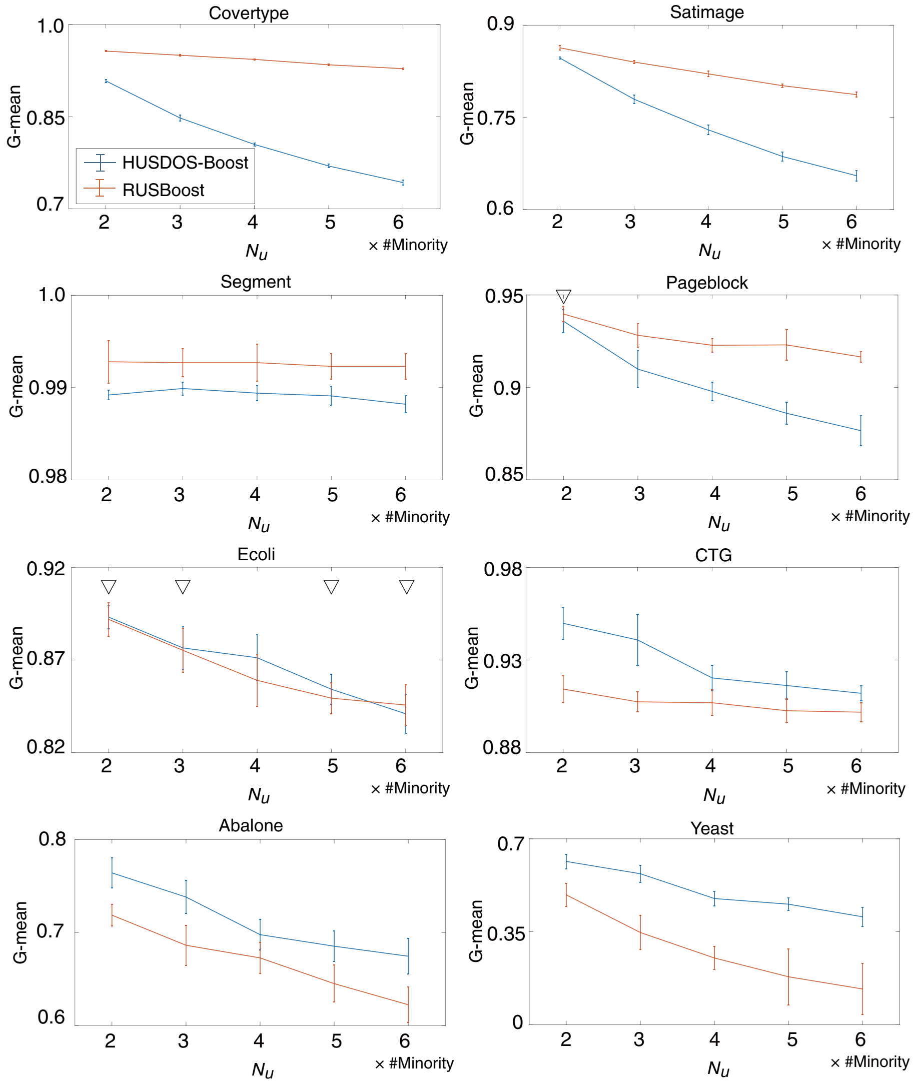


FIGURE 3 | G-means of HUSDOS-Boost and RUSBoost vs. N_u.

and analysis of the HR data. Written informed consent was not obtained in this study.

The original HR data were collected between 2014 and 2015, on more than 100 items, including observations, body measurements, blood examination, medical history, and lifestyle. Since some records belonged to the same person collected in both years, we extracted records measured in the year that stomach cancer was initially diagnosed as patient records and the latest records of persons without stomach cancer as healthy records. Persons who had other types of cancer or a prior stomach operation were eliminated from the analysis. The item "gastroscopy result" was not used as an input variable for stomach cancer detection because it is almost identical to the outcome. In addition, the item "family history of stomach cancer" was eliminated. *Helicobacter pylori* is an essential risk factor for stomach cancer development, in which its main infection path is a family member. Only continuous and binary variables were analyzed here because descriptive variables such as "observations" were difficult to analyze.

Finally, the objective data consisted of 7,379 healthy person records (male: 3,890, female: 3,489, age: 56.6 ± 11.6 years old) and 16 patient records (male: 10, female: 6, age: 68.8 ± 10.8 years old); that is, its imbalance ratio was 0.2%. Twelve out of sixteen patients had tubular adenocarcinoma, and the other four patients had either stage IA or IB signet ring cell carcinoma. Forty-one items were adopted as input variables, which are shown in **Table 4**. "Type" in this table denotes a variable type: a numerical variable (N) and a binary variable (B). No. 1 "Gender" was male/female, and No. 38-41, which asked about lifestyle habits, was yes/no. The data contained about 13% missing values because examination menus vary for each person.

TABLE 3 | CPU times (s).

	RF	AdaBoost	SMOTE	ADASYN	RUSBoost	HUSBoost	hyperSMURF	HUSDOSBoost
Cover type	**41.4±3.57**	4,635±1,582	119.0±13.9	66.4±10.1	285.6±18.1	3,978±90.6	402.7±14.8	3800±328
Satimage	**5.05±0.23**	321.0±12.0	6.08 ±0.88	11.4±0.46	130.6±9.34	191.7±2.99	47.89±1.22	206.1±3.96
Segment	**0.68±0.03**	122.3±3.11	1.09 ±0.18	0.76±0.02	101.5±6.69	39.9±1.71	38.9±0.93	44.7±1.16
Pageblocks	**1.15±0.18**	146.6±3.25	0.90 ±0.10	1.64±0.02	99.3±1.47	36.3±1.13	37.0±0.87	39.4±0.86
Ecoil	**0.09±0.02**	105.5±3.66	0.20 ±0.01	0.21±0.01	97.3±3.41	4.90±0.18	36.5±2.21	5.92±0.20
CTG	**0.68±0.06**	127.2±3.61	0.48 ±0.03	1.14±0.02	97.9±2.59	11.4±0.29	36.8±2.30	13.3±0.47
Abalone	**0.20±0.19**	108.8±2.77	0.21 ±0.01	0.32±0.01	97.0±4.41	6.00±0.20	36.5±1.97	7.14±0.34
Yeast	**0.32±0.03**	113.8±2.32	0.27 ±0.03	0.52±0.00	97.8±5.76	6.53±0.24	36.3±0.83	6.88±1.11

TABLE 4 | Input variables.

No.	Description	Type	No.	Description	Type
1	Gender	B	22	Uric acid	N
2	Age	N	23	Na	N
3	Height	N	24	K	N
4	Weight	N	25	Cl	N
5	Degree of obesity	N	26	Ca	N
6	Body fat percentage	N	27	Cholesterol	N
7	C-reactive protein	N	28	Neutral fat	N
8	Total protein	N	29	HDL cholesterol	N
9	Albumin	N	30	Amylase	N
10	A/G ratio	N	31	LDL cholesterol	N
11	Bilirubin	N	32	White blood cell count	N
12	ALP	N	33	Red blood cell count	N
13	γ GTP	N	34	Hemoglobin content	N
14	GOT	N	35	Hematocrit	N
15	GPT	N	36	Platelet count	N
16	LDH	N	37	fasting blood sugar level	N
17	Cholinesterase	N	38	Habit of quick eating	B
18	ZTT	N	39	Habit of meal before sleep	B
19	BUN	N	40	Habit of breakfast	B
20	Creatinine	N	41	Habit of smoking	B
21	eGFR	N			

5.2. Procedure

The present work applied RF, AdaBoost, SMOTE, ADASYN, RUSBoost, HUSBoost, hyperSMURF, and the proposed HUSDOS-Boost to the HR data for stomach cancer detection. Before analysis, missing values in the HR dataset needed to be input appropriately.

Multiple imputations were used for missing value imputation, which generates multiple complete datasets by replacing missing values with plausible values generated from the posterior distribution of missing values and aggregates them into the final complete dataset (41). We used multiple imputations using chained equations (MICE), which is a standard methodology for coping with HR data with missing values (42). MICE approximates the posterior distribution by regressing it on all other remaining variables. Categorical variables (No. 1 and 38-41) were digitized.

The input data were randomly divided into ten subsets, of which nine were used for modeling while the remaining one was used for validation. Modeling and validation were repeated ten times so that all subsets became the validation dataset once. The above procedure was repeated ten times for precise performance evaluation. The experimental settings of seven methods were the same as section 4.

5.3. Results

Table 5 shows the sensitivities, the specificities, the G-means, AUC, and AUPRC in which the bold fonts indicate the best score in the seven algorithms. RF, AdaBoost, and SMOTE did not function because their sensitivities stayed zero while their specificities were almost one. Thus, these algorithms classified all records as healthy. ADASYN improved the classification performance of SMOTE. On the other hand, the performance of hyperSMURF was not improved.

RUSBoost achieved the highest sensitivity, and HUSDOS-Boost and HUS-Boost were the second and the third best. On the other hand, the specificity of HUSDOS-Boost was higher than RUSBoost. Accordingly, the proposed HUSDOS-Boost achieved the best G-mean and AUC. This result agrees with the result of the case study described in section 4.4. Since the number of patients in the HR data was smaller than 30, the G-mean of HUSDOS-Boost was higher than that of RUSBoost.

AUPRC, however, was almost zero in all algorithms in the HR data. **Figures 4**, **5** are the ROC and PR curves drawn by RUSBoost and HUSDOS-Boost. Their sensitivity (recall) and specificity were not low, and their precision was close to zero, which indicates that many false positives were detected. In this data, the number of cancer patients was extremely small (0.02%) and consequently the number of true positives became small in comparison with that of false positives. This result suggests that AUPRC is not always appropriate for classification performance evaluation of the EISM data problem.

Although, at the present moment, HUSDOS-Boost cannot be applied to stomach cancer detection using the HR data due to its unsatisfactory performance, the result above suggests the future applicability of the proposed HUSDOS-Boost to patient detection by means of HR data analysis, particularly when the number of patient records in the HR data is extremely small.

TABLE 5 | Stomach cancer detection results.

	RF	AdaBoost	SMOTE	ADASYN	RUSBoost	HUSBoost	hyperSMURF	HUSDOSBoost
Sensitivity	0.00±0.00	0.00±0.00	0.00±0.00	0.30±0.08	**0.76±0.06**	0.46±0.05	0.14±0.02	0.59±0.07
Specificity	**1.00±0.00**	**1.00±0.00**	**1.00±0.00**	0.93±0.00	0.61±0.01	0.87±0.00	0.98±0.00	0.80±0.00
G-mean	0.00±0.00	0.00±0.00	0.00±0.00	0.53±0.07	0.68±0.02	0.63±0.03	0.36±0.02	**0.69±0.04**
AUC	0.54±0.02	0.62±0.02	0.58±0.02	0.61±0.01	0.56±0.03	0.76±0.00	0.75±0.01	**0.79±0.00**
AUPRC	0.00±0.00	0.01±0.00	0.00±0.00	0.01±0.00	0.00±0.00	0.01±0.00	**0.02±0.01**	0.01±0.00

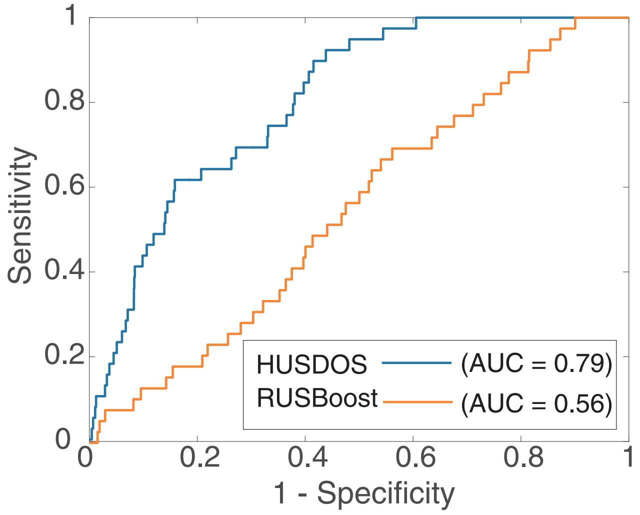


FIGURE 4 | ROC of HUSDOS-Boost and RUSBoost.

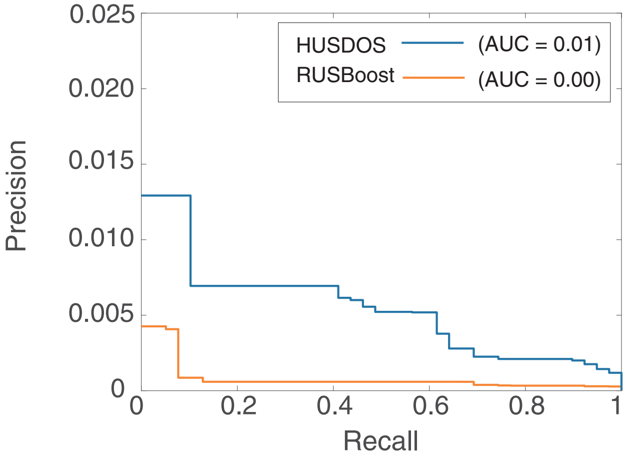


FIGURE 5 | PRC of HUSDOS-Boost and RUSBoost.

5.4. Variable Importance

The variable importance of stomach cancer detection was calculated using RUSBoost and HUSDOS-Boost, which achieved high G-means. **Figure 6** shows the variable importance derived by RUSBoost and HUSDOS-Boost, respectively. "Age" and "amylase" had high importance in both methods.

Age is a factor in stomach cancer development, wherein the morbidity of stomach cancer increases in people over 40 years of age. The mean age of patients was different from healthy persons in the HR data as described in section 5.1. Both methods correctly isolated the factor of stomach cancer from the HR data.

The mean values of amylase were different between patients and healthy persons in the HR data: 88.0 ± 35.8 IU/l of healthy persons and 113.6 ± 45.0 IU/l of patients. They were significantly different ($p = 0.0075$, Effect size: $d = 0.66$, and Power: $1 - \beta = 0.57$); however, the power was rather low due to the sample size of patients being very small. Although salivary gland disorders or pancreatic diseases are suspected when the value of amylase is high, the amylase value becomes high in the elderly population due to the deterioration of amylase clearance in the kidney with age (43). There was the possibility that the values of amylase showed the difference in the mean age between patients and healthy persons. Of course, this result might suggest an unknown relationship between abnormality in amylase and stomach disease, which is difficult to confirm.

Here, we calculated variable importance for another purpose in order to validate the accuracy of the variable importance. Classifiers that detect persons experiencing gastric resection were built by RUSBoost and HUSDOS-Boost, which were utilized for variable importance calculation. Two hundred seven persons experienced gastric resection and did not have stomach cancer at the time of the health examination. The G-means of the classifiers constructed by RUSBoost and HUSDOS-Boost were 0.80 ± 0.01 and 0.77 ± 0.00, respectively. The classification performance of RUSBoost was higher than the proposed HUSDOS-Boost because the number of minority examples, in this case, was more than 40.

Both methods showed that "Age" and "Ca" have the first and the second highest importance for detecting persons with gastric resection. Although there are several causes of persons experiencing gastric resection, they usually occur after middle age. In the HR data, ages of persons with and without gastric resection were 64.9 ± 10.3 and 56.0 ± 11.4, respectively.

In order to confirm the effect of "Age" on the result, we tried to detect stomach cancer without "Age," whose results are shown in **Table 6**. The detection performance in every method deteriorated when "Age" was not used. This indicated that "Age" certainly contributed to stomach cancer detection. In

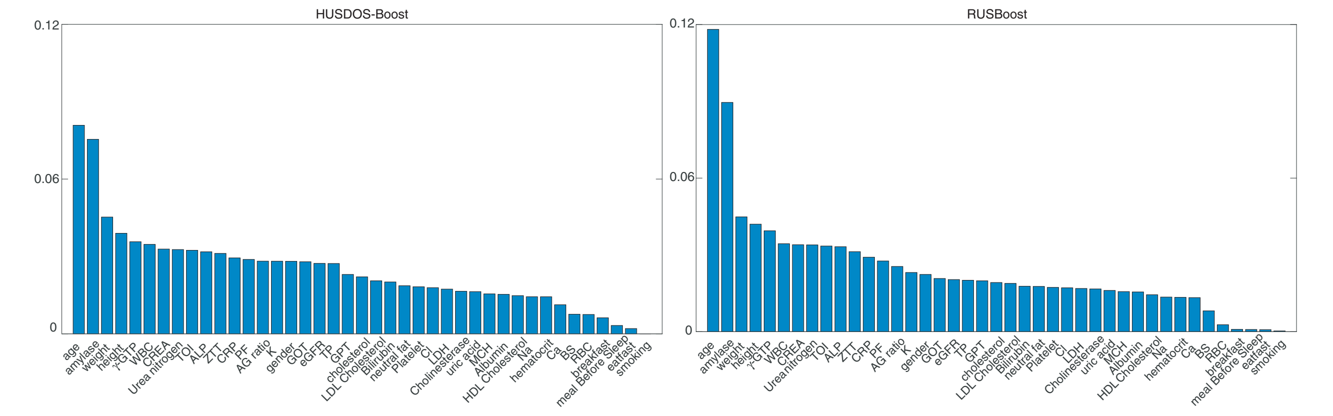


FIGURE 6 | Variable importance: HUSDOS-Boost **(left)** and RUSBoost **(right)**.

TABLE 6 | Stomach cancer detection results without "Age."

	RF	AdaBoost	SMOTE	ADASYN	RUSBoost	HUSBoost	hyperSMURF	HUSDOSBoost
Sensitivity	0.00±0.00	0.00±0.00	0.00±0.00	0.25±0.02	0.72±0.11	0.29±0.06	0.11±0.02	**0.47±0.04**
Specificity	1.00±0.00	1.00±0.00	1.00±0.00	0.94±0.00	0.57±0.01	0.90±0.00	**0.99±0.00**	0.82±0.00
G-mean	0.00±0.00	0.00±0.00	0.00±0.00	0.48±0.02	0.64±0.04	0.51±0.06	0.32±0.03	**0.62±0.03**
AUC	0.54±0.01	0.60±0.02	0.55±0.01	0.59±0.02	0.54±0.03	0.69±0.01	0.70±0.01	**0.71±0.01**
AUPRC	0.00±0.00	0.01±0.00	0.00±0.00	0.00±0.00	0.00±0.00	**0.01±0.00**	**0.01±0.00**	**0.01±0.00**

addition, the proposed HUSDOS-Boost still achieved the best detection performance.

It is well-known that absorption of Ca decreases after gastric resection (44). The Ca values of persons with gastric resection were lower than persons without gastric resection in the HR data: 9.05 ± 0.31 mg/dL and 8.98 ± 0.33 mg/dL of persons with gastric resection, respectively, which were significantly different ($p = 0.026$, Effect size: $d = 0.22$, and Power: $1 - \beta = 0.88$). These results agree with pathological knowledge about the effect of gastric resection. Therefore, this case study shows that variable importance can be applied in the future to the discovery of hidden factors of disease development from HR data.

5.5. Limitations

Limitations include properties of the collected data, such as the fact that all records were from a single hospital and that all records were from the Japanese population. Accordingly, more studies using health records collected from other hospitals are required to confirm our results.

6. CONCLUSION AND FUTURE WORKS

The present work proposed a new boosting-based method for handling EISM data by combining HUS and DOS. The case study using eight imbalanced datasets showed that the proposed HUSDOS-Boost achieved comparable performance to RUSBoost when the number of minority examples was more than 40 and that HUSDOS-Boost achieved the best performance when the number of minority examples was smaller than 30. The proposed HUSDOS-Boost was sufficiently fast for learning.

We applied HUSDOS-Boost to the clinical HR data for detecting patients with stomach cancer. The application result showed that the G-mean of HUSDOS-Boost was 0.69. The possible factors of stomach cancer development derived from the variable importance were discussed.

In future works, the hierarchical Bayes model will be introduced to estimate the distribution parameter in DOS in order to improve the over-sampling performance. We will apply the proposed method to clinical HR data to detect other diseases.

DATA AVAILABILITY STATEMENT

The health examination data will be made available by the corresponding author to colleagues who propose a reasonable scientific request after approval by the institutional review board of the Japanese Red Cross Kyoto Daini Hospital.

ETHICS STATEMENT

The studies involving human participants were reviewed and approved by the institutional review board of the Japanese Red Cross Kyoto Daini Hospital. The ethics committee waived the requirement of written informed consent for participation.

AUTHOR CONTRIBUTIONS

KF, YH, and KH and contributed to algorithm development and clinical data analysis, as well as writing the manuscript. KN, MKam, and MKob collected and organized the data analyzed in this study and interpreted the analysis results. MKan managed study implementation, critically reviewed and edited the manuscript, and gave final approval for submission.

FUNDING

This work was supported in part by JST PRESTO JPMJPR#1859.

REFERENCES

1. Gunter TD, Terry NP. The emergence of national electronic health record architectures in the United States and Australia: models, costs, and questions. *J Med Internet Res*. (2005) 7:e3. doi: 10.2196/jmir.7.1.e3
2. Kierkegaard P. Electronic health record: wiring Europe's healthcare. *Comput Law Secur Rev*. (2011) 27:503–15. doi: 10.1016/j.clsr.2011.07.013
3. Wu PY, Cheng CW, Kaddi CD, Venugopalan J, Hoffman R, IEEE, et al. -Omic and electronic health record big data analytics for precision medicine. IEEE Trans Biomed Eng. (2017) 64:263–73. doi: 10.1109/TBME.2016.2573285
4. [Dataset] The US Office of the National Coordinator for Health Information Technology. Office-Based Physician Electronic Health Record Adoption (2016). Available online at: dashboard.healthit.gov/quickstats/pages/physician-ehr-adoption-trends.php
5. Bell B, Thornton K. From promise to reality: achieving the value of an EHR. *Healthc Financ Manage*. (2011) 65:50–6.
6. Sun Y, Wong AKC, Kamel MS. Classification of imbalanced data: a review. *Intern J Pattern Recognit Artif Intell*. (2009) 23:687–719. doi: 10.1142/S0218001409007326
7. Ali A, Shamsuddin SM, Ralescu A. Classification with class imbalance problem: a review. *Int J Advance Soft Compu Appl*. (2015) 7:176–204.
8. Loh WY. Classification and regression trees. *WIREs Data Mining Knowledge Discov*. (2011) 1:14–23. doi: 10.1002/widm.8
9. He H, Garcia EA. Learning from imbalanced data. *IEEE Trans Knowl Data Eng*. (2009) 21:1041–4347. doi: 10.1109/TKDE.2008.239
10. Phua C, Alahakoon D, Lee V. Minority report in fraud detection: classification of skewed data. *ACM SIGKDD Explor*. (2004) 6:50–9. doi: 10.1145/1007730.1007738
11. Kubat M, Holte RC, Matwin, S. Machine learning for the detection of oil spills in satellite radar images. *Mach Learn*. (1998) 30:195–215. doi: 10.1023/A:1007452223027
12. Malamud BD, Turcotte DL. The applicability of power-law frequency statistics to floods. *J Hydrol*. (2006) 322:168–80. doi: 10.1016/j.jhydrol.2005.02.032
13. Manevitz LM, Yousef M. One-class SVMs for document classification. *J Mach Learn Res*. (2001) 2:139–54.
14. Breunig MM, Kriegel HP, Ng RT, Sander J. LOF: identifying density-based local outliers. In: *Proc ACM SIGMOD 2000 Int Conf On Management of Data*. Dallas, TX (2000). doi: 10.1145/342009.335388
15. Fujiwara K, Miyajima M, Yamakawa T, Abe E, Suzuki Y, Sawada Y, et al. Epileptic seizure prediction based on multivariate statistical process control of heart rate variability features. *IEEE Trans Biomed Eng*. (2016) 63:1321–32. doi: 10.1109/TBME.2015.2512276
16. Kano M, Hasebe S, Hashimoto I, Ohno H. A new multivariate statistical process monitoring method using principal component analysis. *Comput Chem Eng*. (2001) 25:1103–13. doi: 10.1016/S0098-1354(01)00683-4
17. MacGregor JF, Kourti T. Statistical process control of multivariate processes. *Control Eng Pract*. (1995) 3:403–14. doi: 10.1016/0967-0661(95)00014-L
18. Westerhuis JA, Gurden SP, Smilde AK. Generalized contribution plots in multivariate statistical process monitoring. *Chemom Intell Lab Syst*. (2000) 51:95–114. doi: 10.1016/S0169-7439(00)00062-9
19. Yue HH, Qin SJ. Reconstruction-based fault identification using a combined index. *Ind Eng Chem Res*. (2001) 40:4403–14. doi: 10.1021/ie000141+
20. Bach FR, Heckerman D, Horvitz E. Considering cost asymmetry in learning classifiers. *J Mach Learn Res*. (2006) 7:1713–41.
21. Akbani R, Kwek S, Japkowicz N. Applying support vector machines to imbalanced data sets. In: *Machine Learning: ECML 2004, 15th European Conference on Machine Learning*. Pisa (2004). doi: 10.1007/978-3-540-30115-8_7
22. Quinlan JR. Induction of decision trees. *Mach Learn*. (1986) 1:81–106. doi: 10.1007/BF00116251
23. Quinlan JR. *C4.5*: Programs for Machine Learning. Burlington, VT: Morgan Kaufmann Publishers (2014).
24. Liu W, Chawla S, Cieslak DA, Chawla NV. A robust decision tree algorithm for imbalanced data sets. In: *The 2010 SIAM International Conference on Data Mining*. Columbus,OH (2010). doi: 10.1137/1.9781611972801.67
25. Batuwita R, Palade V. FSVM-CIL: fuzzy support vector machines for class imbalance learning. *IEEE Trans Fuzzy Syst*. (2010) 18:558–71. doi: 10.1109/TFUZZ.2010.2042721
26. Seiffert C, Khoshgoftaar T, Van Hulse J, Napolitano A. RUSBoost: a hybrid approach to alleviating class imbalance. *IEEE Trans Syst Man Cybern A Syst Humans*. (2010) 40:185–97. doi: 10.1109/TSMCA.2009.2029559
27. Chawla NV, Bowyer KW, Hall LO, Kegelmeyer WP. SMOTE: synthetic minority over-sampling technique. *J Artif Intell Res*. (2002) 16:321–57. doi: 10.1613/jair.953
28. He H, Bai Y, Garcia EA, Li S. ADASYN: adaptive synthetic sampling approach for imbalanced learning. In: *IEEE International Joint Conference on Neural Networks, 2008, IJCNN 2008 (IEEE World Congress on Computational Intelligence)*. Hong Kong (2008). p. 1322–28.
29. Galar M, Fernández A, Barrenechea E, Sola HB, Herrera F. A review on ensembles for the class imbalance problem: bagging, boosting-, and hybrid-based approaches. *IEEE Trans Syst Man Cybern C Appl Rev*. (2012) 42:463–84. doi: 10.1109/TSMCC.2011.2161285
30. Freund Y, Schapire RE. A decision-theoretic generalization of on-line learning and an application to boosting. *J Comput Syst Sci*. (1997) 55:119–39. doi: 10.1006/jcss.1997.1504
31. Ho TK. The random subspace method for constructing decision forests. *IEEE Trans Pattern Anal Mach Intell*. (1998) 20:832–44. doi: 10.1109/34.709601
32. Breiman L. Random forests. *Mach Learn*. (2001) 45:5–32. doi: 10.1023/A:1010933404324
33. Louppe G, Wehenkel L, Sutera A, Geurts P. Understanding variable importances in forests of randomized trees. In: *Advances in Neural Information Processing Systems 26*. Stateline, NV (2013).
34. Schubach M, Re M, Robinson PN, Valentini G. Imbalance-aware machine learning for predicting rare and common disease-associated non-coding variants. *Sci Rep*. (2017) 7:2959. doi: 10.1038/s41598-017-03011-5
35. Freund Y, Schapire RE. Experiments with a new boosting algorithm. In: *ICML'96: Proceedings of the Thirteenth International Conference on International Conference on Machine Learning*. Bari (1996).
36. Kumar TK. Multicollinearity in Regression Analysis. *Rev Econ Stat*. (1975) 57:365–6. doi: 10.2307/1923925
37. Belsley DA, Kuh E, Welsch RE. Regression Diagnostics: Identifying Influential Data and Sources of Collinearity. New York: John Wiley & Sons (1980) doi: 10.1002/0471725153
38. [Dataset] UCI Repository of Machine Learning Databases (2018). Available online at: https://archive.ics.uci.edu/ml/index.php
39. Barua S, Islam MM, Yao X, Murase K. MWMOTE-majority weighted minority oversampling technique for imbalanced data set learning. *IEEE Trans Knowl Data Eng*. (2014) 26:405–25. doi: 10.1109/TKDE.2012.232
40. Lim P, Goh CK, Tan KC. Evolutionary cluster-based synthetic oversampling ensemble (ECO-Ensemble) for imbalance learning. *IEEE Trans Cybern*. (2017) 47:2850–61. doi: 10.1109/TCYB.2016.2579658

41. Barnard J, Meng XL. Applications of multiple imputation in medical studies: from AIDS to NHANES. *Stat Methods Med Res*. (1999) 8:17–36. doi: 10.1177/096228029900800103
42. Wells BJ, Chagin KM, Nowacki AS, Kattan MW. Strategies for handling missing data in electronic health record derived data. *EGEMS*. (2013) 1:1035. doi: 10.13063/2327-9214.1035
43. Ueda M, Araki T, Shiota T, Taketa K. Age and sex-dependent alterations of serum amylase and isoamylase levels in normal human adults. *J Gastroenterol*. (1994) 29:189–91. doi: 10.1007/BF02358681
44. Schafer AL, Weaver CM, Black DM, Wheeler AL, Chang H, Szefc GV, et al. Intestinal calcium absorption decreases dramatically after gastric bypass surgery despite optimization of vitamin D status. *J Bone Miner Res*. (2015) 30:1377–85. doi: 10.1002/jbmr.2467

Conflict of Interest: KF is with Quadlytics Inc. as well as Nagoya University. KH and MKan is with Quadlytics Inc. as well as Kyoto University.

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.

ORIGINAL RESEARCH
published: 25 May 2020
doi: 10.3389/fninf.2020.00017

Primary Categorizing and Masking Cerebral Small Vessel Disease Based on "Deep Learning System"

Yunyun Duan[1,2†], Wei Shan[2,3,4†], Liying Liu[2‡], Qun Wang[2,3,4], Zhenzhou Wu[2], Pan Liu[2], Jiahao Ji[2], Yaou Liu[1,2], Kunlun He[5,6*] and Yongjun Wang[2,3*]*

[1] Department of Radiology, Beijing Tiantan Hospital, Capital Medical University, Beijing, China, [2] National Center for Clinical Medicine of Neurological Diseases, Beijing, China, [3] Department of Neurology, Beijing Tiantan Hospital, Capital Medical University, Beijing, China, [4] Beijing Institute for Brain Disorders, Beijing, China, [5] Laboratory of Translational Medicine, Chinese PLA General Hospital, Beijing, China, [6] Key Laboratory of Ministry of Industry and Information Technology of Biomedical Engineering and Translational Medicine, Chinese PLA General Hospital, Beijing, China

OPEN ACCESS

Edited by:
Koichi Fujiwara,
Nagoya University, Japan

Reviewed by:
Tomohiro Mitani,
The University of Tokyo, Japan
Akihiro Shindo,
Mie University Hospital, Japan
Chisae Murata,
National Institute of Advanced Industrial Science and Technology, Japan
Xinfeng Liu,
Jinling Hospital, China

***Correspondence:**
Yaou Liu
yaouliu80@163.com
Kunlun He
kunlunhe@plagh.org
Yongjun Wang
weishantiantan@gmail.com

† These authors have contributed equally to this work

‡ LL is the main technical contributor

Received: 21 December 2019
Accepted: 31 March 2020
Published: 25 May 2020

Citation:
Duan Y, Shan W, Liu L, Wang Q, Wu Z, Liu P, Ji J, Liu Y, He K and Wang Y (2020) Primary Categorizing and Masking Cerebral Small Vessel Disease Based on "Deep Learning System". Front. Neuroinform. 14:17. doi: 10.3389/fninf.2020.00017

Objective: To supply the attending doctor's diagnosis of the persisting of cerebral small vessel disease and speed up their work effectively, we developed a "deep learning system (DLS)" for cerebral small vessel disease predication. The reliability and the disease area segmentation accuracy, of the proposed DLS, was also investigated.

Methods: A deep learning model based on the convolutional neural network was designed and trained on 1,010 DWI b1000 images from 1010 patients diagnosed with segmentation of subcortical infarction, 359 T2* images from 359 patients diagnosed with segmentation of cerebral microbleed, as well as 824 T1-weighted and T2-FLAIR images from 824 patients diagnosed with segmentation of lacune and WMH. Dicw accuracy, recall, and f1-score were calculated to evaluate the proposed deep learning model. Finally, we also compared the DLS prediction capability with that of 6 doctors with 3 to 18 years' clinical experience (8 ± 6 years).

Results: The results support that an appropriately trained DLS can achieve a high-level dice accuracy, 0.598 in the training section over all these four classifications on 30 patients (0.576 for young neuroradiologists), validation accuracy is 0.496 in lacune, 0.666 in WMH, 0.728 in subcortical infarction, and 0.503 in cerebral microbleeds. It is comparable to attending doctor with a few years of experience, regardless of whether the emphasis is placed on the segmentation or detection of lesions with less time-spending compared with manual analysis, about 4.4 s/case, which is dramatically less than doctors about 634 s/case.

Conclusion: The results of our comparison lend support to the case that an appropriately trained DLS can be trusted to the same extent as one would trust an attending doctor with a few years of experience, regardless of whether the emphasis is placed on the segmentation or detection of lesions.

Keywords: cerebral small vessel diseases (CSVD), deep learning system (DLS), categorizing, subcortical infarction, white matter hyperintensity, launce, cerebral microbleed, diagnosis-assistance

INTRODUCTION

The Cerebral Small Vessel Disease (CSVD) is an umbrella term covering a variety of abnormalities related to small blood vessels in the brain, which can be caused by many diseases, such as plaque accumulation in the small vessel, small vessel inflammation, and persisted chronic damage in the small vessel (hypertension) (Go et al., 2012; Rincon and Wright, 2014). Consequently, it could lead to irreversible consequences such as stroke, dementia, mood disturbance, and gait problems. The CSVD can be diagnosed by medical professionals based on magnetic resonance imaging (MRI) (Noguchi et al., 1997; Greenberg et al., 2009; Debette and Markus, 2010). Signs of CSVD on conventional MRI include lacunes, white matter hyperintensities (WMH), recent small subcortical infarcts, prominent perivascular spaces, cerebral microbleeds, and atrophy (Wardlaw et al., 2013). CSVD has been suggested to be an essential source of morbidity associated with ischaemic and

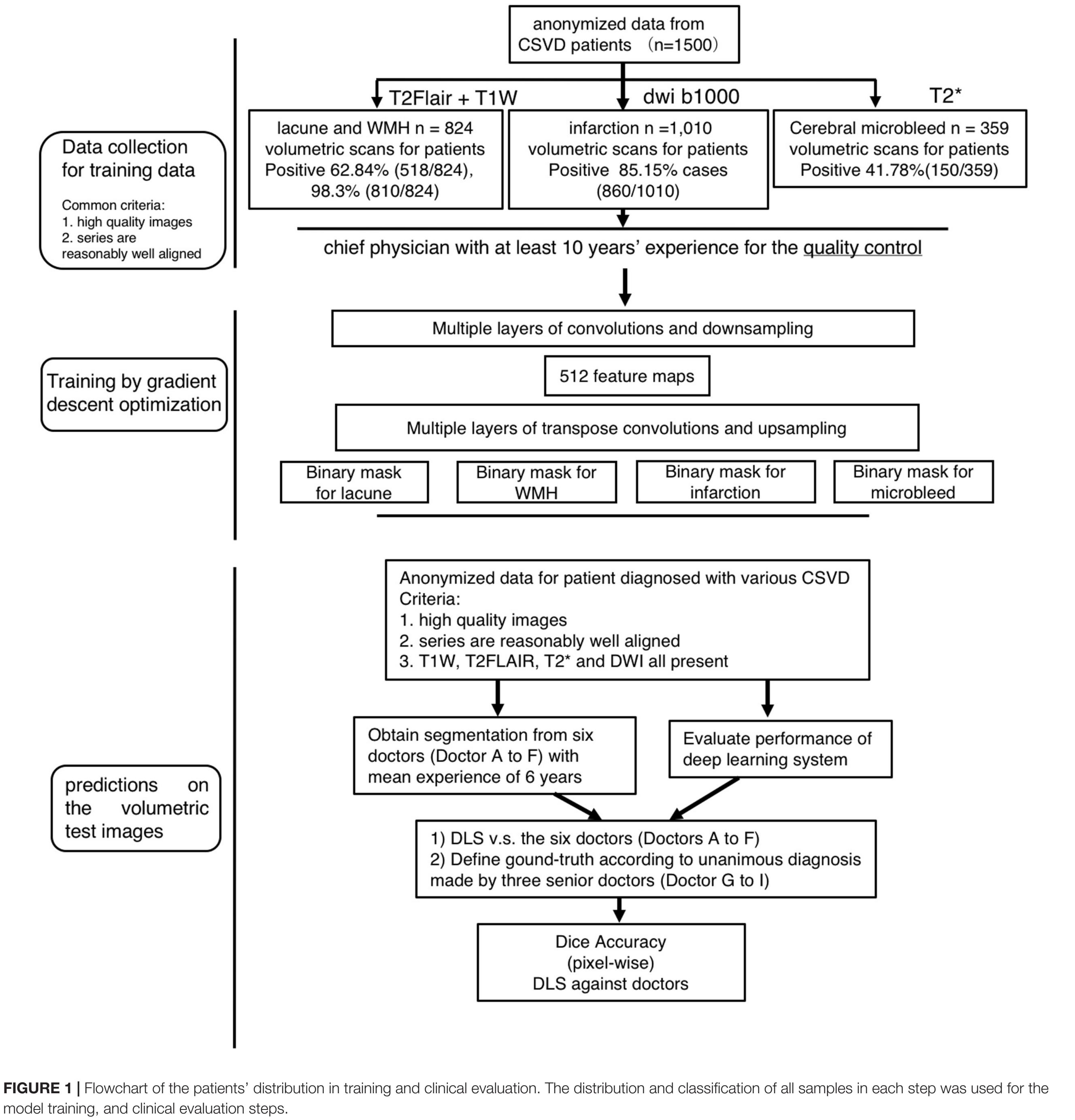

anonymized data from
CSVD patients （n=1500）
T2Flair + T1W
dwi b1000
T2*
Data collection
for training data
Common criteria:
1. high quality images
2. series are
reasonably well aligned
lacune and WMH n = 824
volumetric scans for patients
Positive 62.84% (518/824),
98.3% (810/824)
infarction n =1,010
volumetric scans for patients
Positive 85.15% cases
(860/1010)
Cerebral microbleed n = 359
volumetric scans for patients
Positive 41.78%(150/359)
chief physician with at least 10 years' experience for the quality control
Training by gradient
descent optimization
Multiple layers of convolutions and downsampling
512 feature maps
Multiple layers of transpose convolutions and upsampling
Binary mask
for lacune
Binary mask for
WMH
Binary mask for
infarction
Binary mask for
microbleed
predictions on
the volumetric
test images
Anonymized data for patient diagnosed with various CSVD
Criteria:
1. high quality images
2. series are reasonably well aligned
3. T1W, T2FLAIR, T2* and DWI all present
Obtain segmentation from six
doctors (Doctor A to F) with
mean experience of 6 years
Evaluate performance of
deep learning system
1) DLS v.s. the six doctors (Doctors A to F)
2) Define gound-truth according to unanimous diagnosis
made by three senior doctors (Doctor G to I)
Dice Accuracy
(pixel-wise)
DLS against doctors

FIGURE 1 | Flowchart of the patients' distribution in training and clinical evaluation. The distribution and classification of all samples in each step was used for the model training, and clinical evaluation steps.

hemorrhagic stroke, dementia, and depression (Pantoni, 2010). So it is critical to define the severity of CSVD by a quantitative assessment from MRI, which is relevant to the risk of stroke. However, the severity of CSVD is mainly evaluated by manual semi-quantitative or qualitative methods at present, which is time-consuming, laborious, and subjective (Rensma et al., 2018).

Nowadays, the deep convolutional neural networks (CNN) has proven to be useful and effective in medical applications, such as the classification (Mohsen et al., 2018) and segmentation (Havaei et al., 2017) of brain tumor problem as well as various vessel diseases (Dou et al., 2016; Ghafoorian et al., 2017). Besides, computers are immune to fatigue or emotions and can function 24 h daily. Moreover, a high-quality automatic segment can probably help doctors to speed up their diagnosis, and hence allowing more patients to be processed. In recent studies, deep learning has applied in stroke imaging data in areas including automated featurization, image segmentation, and multimodel prognostication (Huang et al., 2010; Misra et al., 2010; Kamnitsas et al., 2015; Stier et al., 2015). One of the significant strengths of deep learning is that there is no obvious solution that could be obtained manually, such as the prediction of poststroke MRI fluid-attenuated inversion recovery (FLAIR) changes given acute diffusion-weighted imaging‘(DWI) maps (Stier et al., 2015). Currently, the application of deep learning in CSVD is seldom reported. Several deep learning models for segmentation have been applied in three-dimensional images and worked well (Kamnitsas et al., 2017; Aslani et al., 2019). Nevertheless, two-dimensional data is commonly used in clinical practice.

To supply people with consistent and efficient CSVD area segmentation systems and help the young doctors to speed up their workflow, we developed a DLS for automatic area segmentation. Furthermore, to check the reliability of this system, we seek to investigate the relative performance between the proposed DLS system and human doctors on detecting and locating four types of CSVD (lacune, WMH, subcortical infarction and cerebral microbleeds) using T1-weighted, T2*, T2-FLAIR, and DWI b1000 sequences. Prominent perivascular spaces and atrophy were not included in the DLS system, for they are difficult to make a reasonable manual evaluation in conventional two-dimensional images. Specifically, we compare the performance of the proposed DLS with the average performance of six doctors. Accurate lesion segmentation and identification can guarantee objective and accurate quantitative evaluation. The purpose of this study is to validate whether an appropriately-trained DLS can be trusted to the same extent as one would trust a doctor with a satisfying experience. Then the system may be applied to quantify the lesion load of CSVD and further to help establish the risk factor prediction model.

MATERIALS AND METHODS

Standard Protocol Approvals, and Patient Consents

All the patients provided consent for access to the image data in this study. This study was approved by the ethics committee of the Beijing Tiantan Hospital and fulfilled the Helsinki Declaration.

Data Quality Control

For the image quality evaluation, the three-point scale was applied: 1, “poor” (limited image quality that affects diagnosis); 2, “good” (minor artifacts or mildly reduced signal-noise ratio with no effects on diagnosis); and 3, “excellent” (no artifacts and optimal). Only scale 2 or 3 were allowed to be included in this study. More details of manufacturer and resolution information in **Supplementary Figure 1**.

Image Dataset

We obtain 1500 anonymized patients data from Beijing Tiantan hospital and other 12 hospitals across China which are included

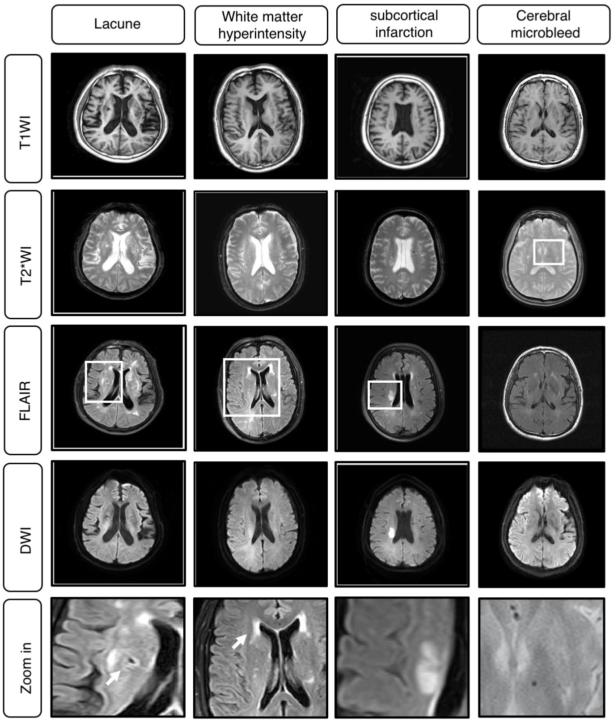


FIGURE 2 | Example cases of Cerebral Small Vessel Disease (CSVD) MRI A. Classical MRI of CSVD including lacune, white matter hyperintensity (WMH), subcortical infarction, cerebral microbleed.

TABLE 1 | The definitions of imaging characteristics for CSVD on MRI.

	Lacunes	**White matter hyperintensity**	**subcortical infarct**	**Cerebral microbleed**
DWI	↓/↔	↔	↑	↔
T1	↓CSF-like	↓/↔	↓	↔
T2-FLAIR	↓/↔	↑	↑	↔
T2*-weighted GRE	↓/↔ if haemorrhage	↔	↔	↔
Diameter	3 to 15 mm	Variable	≤20 mm	2 to 10 mm

↑ signal increased, ↓ signal decreased, ↔ equal signal.

in The Third China National Stroke Registry (CNSR-III). The MRI data include T1 weighted image (T1WI), T2*, T2-FLAIR, DWI b1000 and TOF-MRA., The inclusion criteria of patients were: (Rincon and Wright, 2014) Age older than 18 (Go et al., 2012) Ischemic stroke or Transient ischemic attack (TIA) (Greenberg et al., 2009) Informed consent from patient or legally authorized representative (Primarily spouse, parents, adult Children, otherwise indicated) (Debette and Markus, 2010) The presence of one CSVD sign or more on MRI (Kamnitsas et al., 2015). Patients who had other abnormalities such as hemorrhage or brain tumor on MRI and well-defined macro-vascular stenosis on MRA were excluded.

TABLE 2 | Clinical symptom distribution in the evaluation dataset (*n* = 30).

	Lacune	White matter hyperintensity	subcortical infarct	Cerebral microbleed
Positive symptom	30	27	29	30
Negative symptom	0	3	1	0

In total, we have 824 T1-weighted and T2-FLAIR images from 824 patients with segmentation of lacune and WMH, 1,010 DWI b1000 images from 1010 patients with segmentation of subcortical infarction, as well as 359 T2* images from 359 patients with segmentation of cerebral microbleed. Each volumetric MRI has a vertical spacing of between 6 and 8 mm. For each image, the spacing along the x- and y-direction varies from 0.36 to 1.44 mm between consecutive pixels. The distribution of pixel spacings for each dataset are shown in **Figure 1**. Instead of resizing the images to ensure a uniform pixel spacing, we train the model to be scale-invariant within the reasonable range of resolutions encountered in MRI.

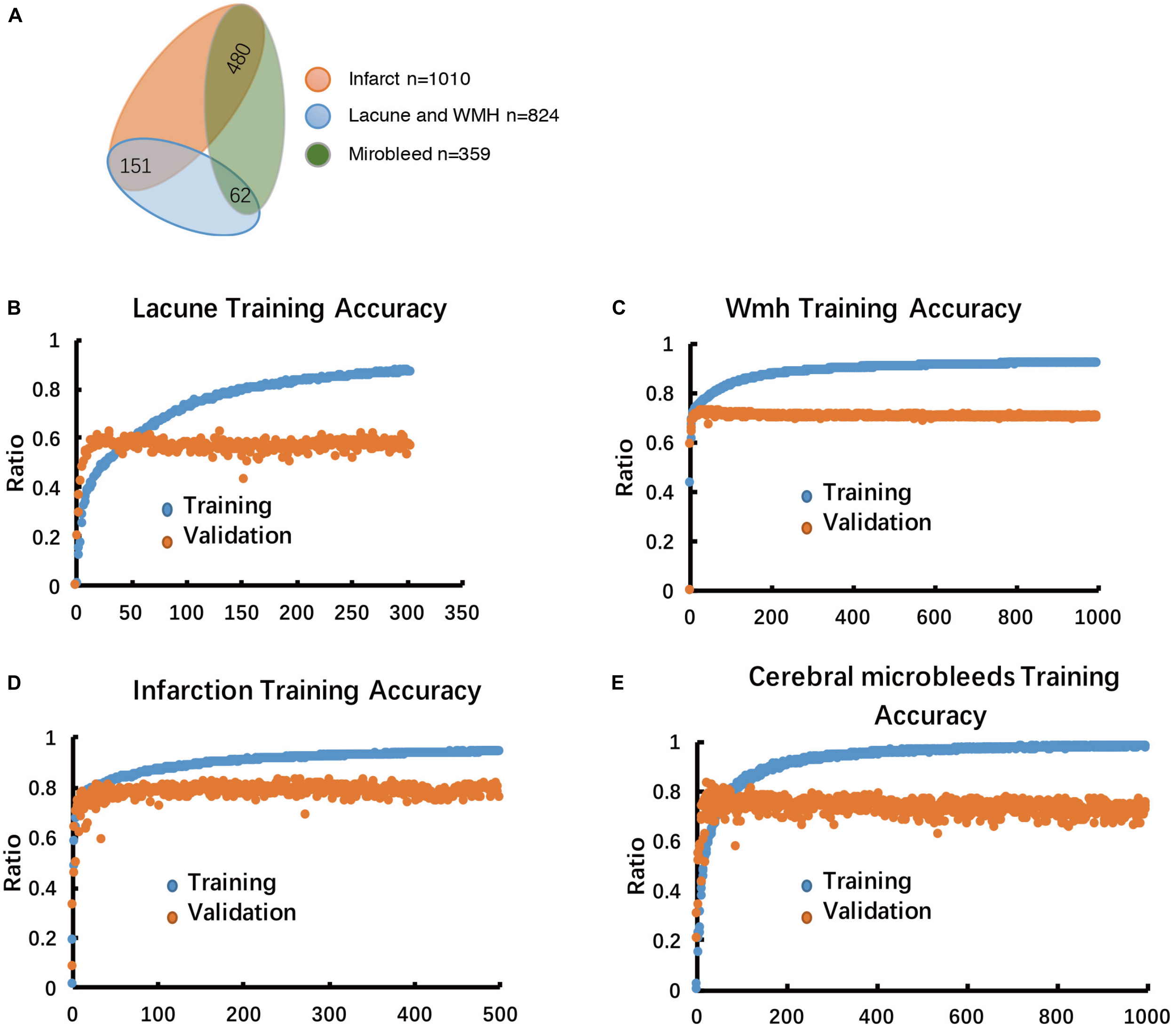


FIGURE 3 | Multiple-label classification of CSVD, and Performance analysis of the model in the training stage. **(A)** Multiple-label classification of CSVD. **(B–E)** Model performance in training accuracy and validation accuracy.

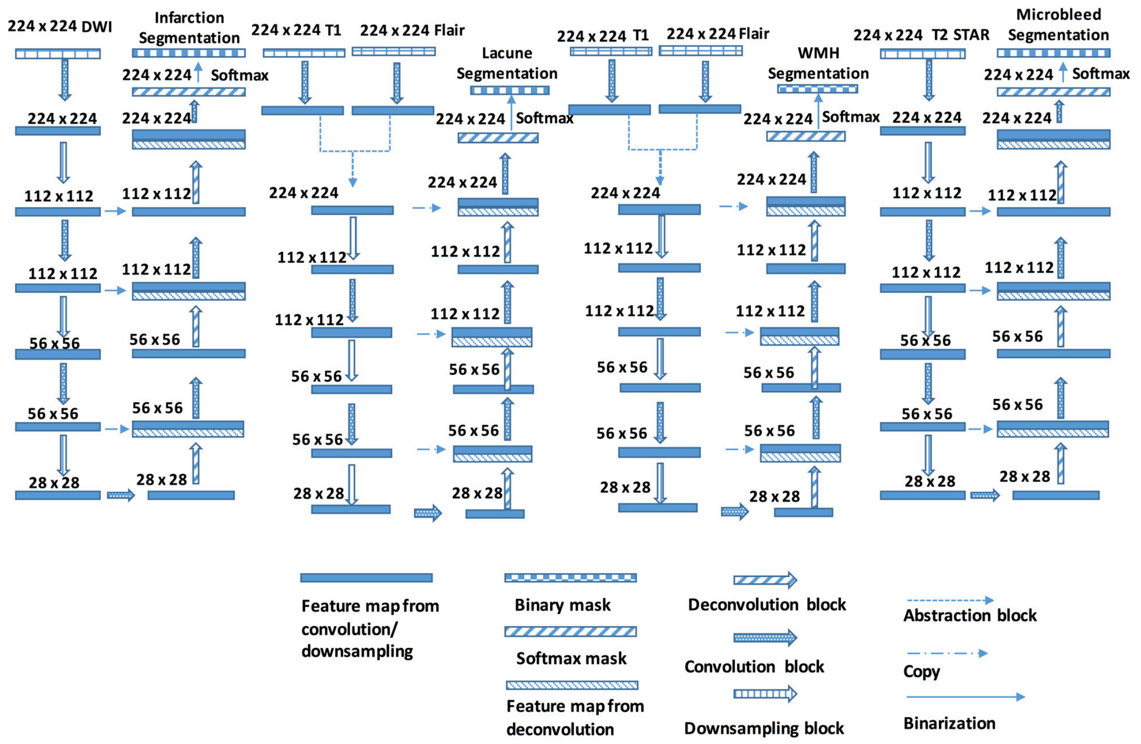


FIGURE 4 | The Structure of the DLS used for CSVD detection and segmentation. Each Encoder block contains one or more convolution steps followed by max-pooling for downsampling. Each time the feature maps are downsampled, the number of output channels is increased. Each Decoder block comprises one deconvolution (transpose convolution) operation that upsamples the size of the feature maps and correspondingly reduces the number of output channels.

The segmentation labels for patients with lacune, WMH, subcortical infarction, or cerebral microbleed are endorsed by two radiologists with 12 years of clinical experience. The lacunes were mainly labeled on T1WI (CSF-like hypointensity) with referred to T2-FLAIR. The segmentation labels with WMH were based on T2-FLAIR. The subcortical infarction was labeled on DWI b1000 images. The segmentation labels with cerebral microbleed were labeled on T2*-weighted GRE, with other sequences as reference. All the segmentation character of MRI illustrated in **Tables 1**, **2** and **Figure 2**.

Evaluation Dataset and Reference Standard

The evaluation dataset comprises 30 patients, with T1-weighted, T2*, T2-FLAIR, and DWI b1000 sequences available for each patient. All these patients' clinical diagnosis must meet the inclusion criteria and each patent's image must have 2 to 4 signs of CSVD, and all of these patients are independed from the previous dataset.

We define the ground truth location of these four possible diseases according to the diagnosis and segmentation label by three senior physicians, with all giving their consensus. These three physicians who set the reference standard on the 30 patients are top expects on radiology in our hospital with 12, 13, and 15 years of experience, respectively.

TABLE 3 | Dice accuracy at pixel-wise criteria and F1 score for four CSVDs.

	Our model	Doctors
Dice Accuracy (Pixel-wise)	0.598	0.576
Region F1 score	0.725	0.691

Dice accuracy at pixel-wise criteria and F1 score for four CSVDs of 30 patients according to the predictions made by six doctors (Neuroradiologists) and by our model, concerning the reference standard. In the pixel-wise evaluation, the images and masks have a resolution of 224 × 224 pixels.

Credentials of Doctors Performing Segmentation on the Evaluation Dataset

After training on 1,500 patients MRI obtained from hospitals, we make predictions on 30 patients chosen by a hospital doctor randomly among patients who had T1-weighted, T2*, DWI and FLAIR sequences in their records. The reference standard is prescribed unanimously by three senior doctors as described previous.

The six doctors in the evaluation test independently performing segmentation on the evaluation dataset include three resident physicians, each with three years of experience, an attending physician with nine years of experience, and two chief physicians with 14 and 18 years of experience, respectively. All the doctors included in the tests are neuroradiologists.

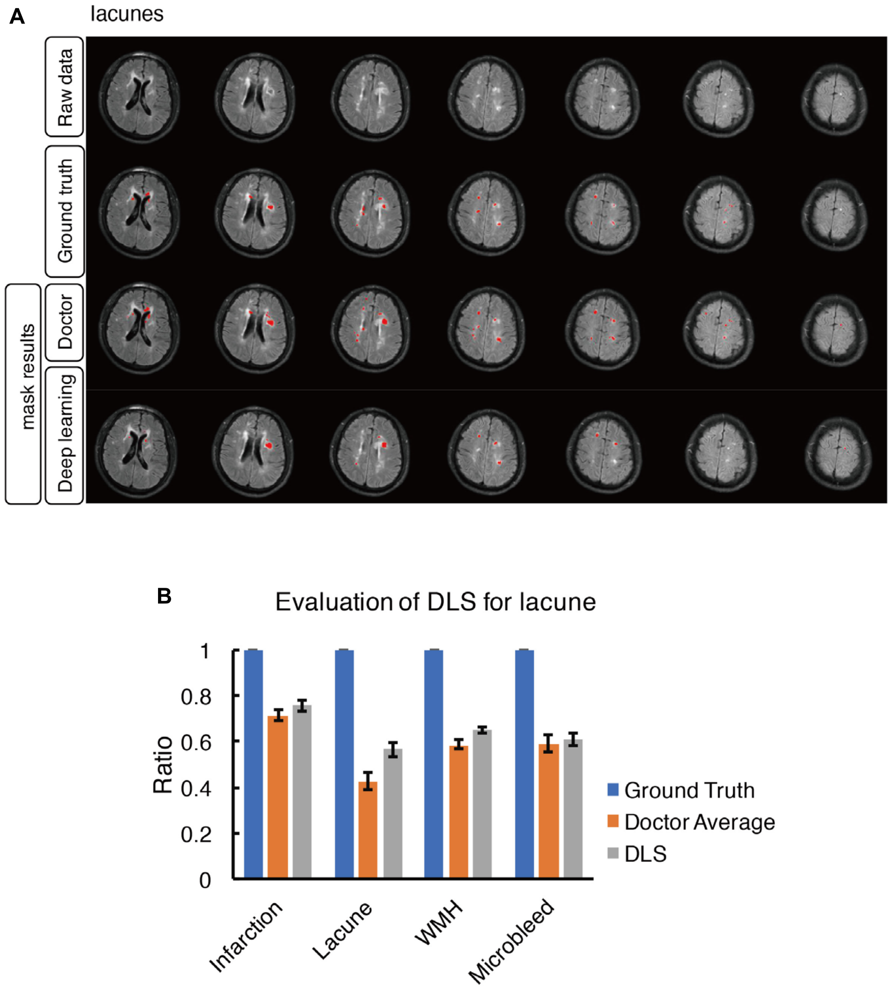


FIGURE 5 | Mask result of lacune in the study, including Raw data, Ground truth, Doctor mask and DLS mask result. **(A)** showed multiple lacunes in the regions of bilateral paraventricular and semi-oval center, represented as well-defined CSF-like hypointensity on T1WI. The four rows are raw data, ground truth, doctor's segmentation label and segmentation prediction from DLS, respectively. **(B)** showed the comparison of accuracy ratio of segmentation label from doctors with 95% confidence interval and DLS.

To ensure that the doctors are evaluated in their best state, they are requested to perform the segmentation to the best of their abilities, without any constraint on time or duration.

Setting up of the Deep Learning Algorithm

The proposed DLS system which consists of four segementor subsystems was trained to learn features from MRI, extracted from four different types of CSVD diseases. To increase the rate of convergence of the network during training, preprocessing was done on each MRI to standardize them across various acquisition parameters. The histogram peaks were normalized and aligned based on the white matter content in the MRI.

The training set was consisted of 1500 patients with conventional MRI T1W, T2*, T2-FLAIR, DWI b1000 image data, including Lacuna data (n = 824 volumetric scans, 98.3% positive cases) and WMH data (n = 824 volumetric scans, 98.3% positive cases), subcortical infarction data (n = 1010 volumetric scans, 85.15% positive cases), Cerebral microbleed data (n = 359 volumetric scans, 41.78% positive cases) (**Figures 1**, **3A**). Each disease was trained independently by one segmentor subsystem of the DLS system. Each training was stopped when the training accuracy was greater than 98% and diverged from validation accuracy by more than 15%, as we think that at such time, the DLS has reached the optimal performance (**Figure 3**).

During the inference stage, patient's MRI sequences are fed to the DLS system as input. Each segementor subsystem grabs its own sequences from the DLS system's input and give a segmentation prediction of a certain disease. Before giving a final output of the DLS, the four segmentation predictions are combined in a way such that WMH, lacune and subcortical infarction are multually exclusive in the pixel level. Please note that it's a multi-label classification problem in the image or patient level.

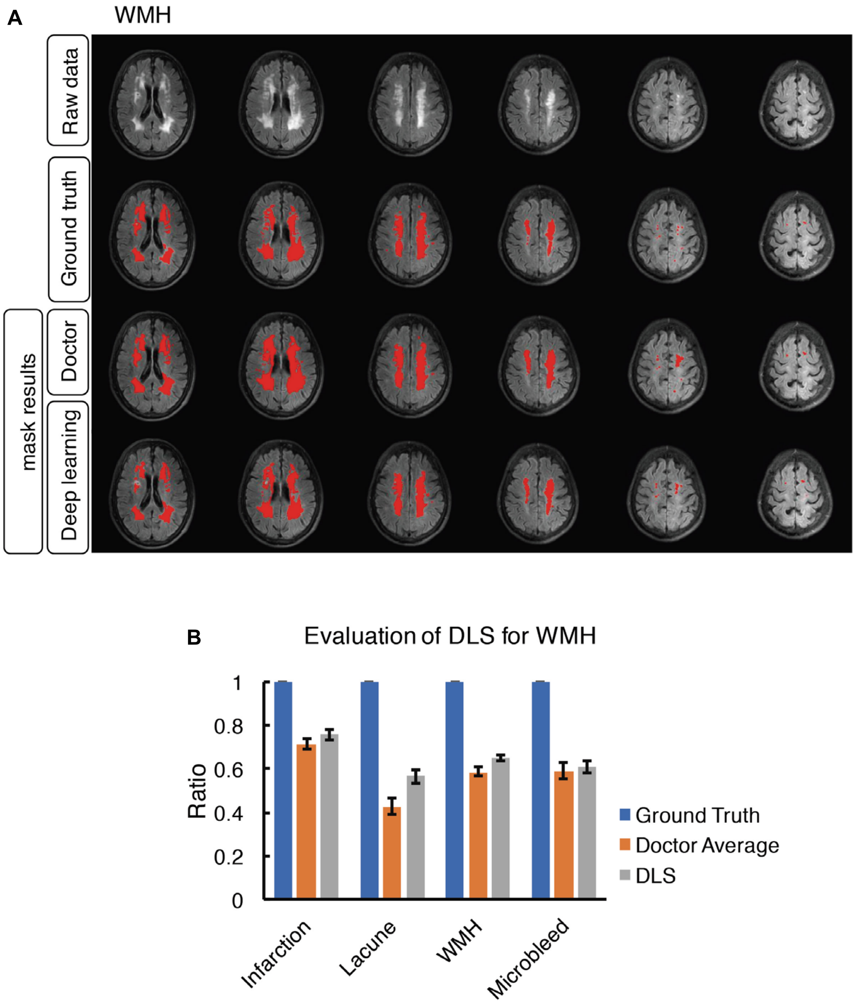


FIGURE 6 | Mask result of WMH in the study, including Raw data, Ground truth, Doctor mask and DLS mask result. **(A)** showed hyperintensity in bilateral white matter regions of paraventricular and the frontal and parietal lobe on T2-FLAIR. The four rows are raw data, ground truth, doctor's segmentation label and segmentation prediction from DLS. **(B)** showed the comparison of accuracy ratio of segmentation label from doctors with 95% confidence interval and DLS.

Network Architecture

The proposed, end-to-end, DLS was composed of four segmentor networks (**Figure 4**). The preprocessing steps consisted of padding to square, resizing, and normalizing. Each segmentor network takes one or more MRI sequences as input and outputs a binary segmentation mask on a corresponding sequence.

All four segmentor networks are based on the widely-used U-Net architecture (Ronneberger et al., 2015; Havaei et al., 2016), and each of them predicts the mask of the disease area of one of four CSVD introduced above. For example, segmentor network #1 is used to detect brain subcortical infarction. It takes DWI b1000 as input and outputs a segmentation mask of subcortical infarction area. Those four segmentor networks are trained and validated independently, which allows the network to be optimized for detecting the CSVD, and are combined to perform the disease area prediction. For each patient, the volumetric MRI are separated into two dimensional images and, after certain preprocessing steps, fed to the DLS to generate masks of diseases area. During the postprocessing steps, the generated WMH mask is subtracted by the generated subcortical infarction mask, because subcortical infarction has a similar signal property to WMH on T2-FLAIR. For the same reason, the generated WMH mask is also subtracted by the generated lacune mask. Then, the generated lacune mask is subtracted by the generated subcortical infarction mask. Finally, the two dimensional segmentation predictions from our model are concatenated to obtain a complete three dimensional segmentation predictions of the patient (More detailed information in **Figure 4**).

Algorithm for Segmented Images

To evaluate the performance of proposed segmentation networks, the commonly used metric known as the dice score (accuracy) was used (Sudre et al., 2017). The dice score is computed for each patient, and the arithmetic mean is taken.

A Free Response Operating Characteristic (FROC) analysis can be obtained in this study (Bandos et al., 2009). Due to binary

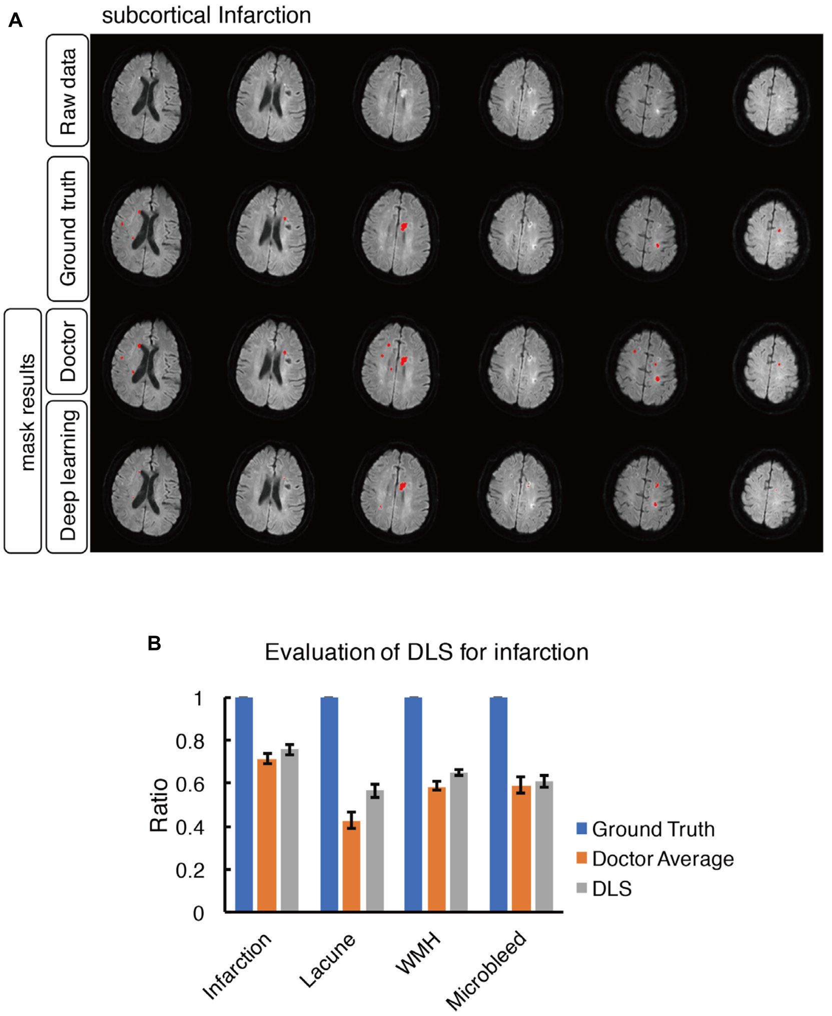


FIGURE 7 | Mask result of subcortical infarction in the study, including Raw data, Ground truth, Doctor mask and DLS mask result. **(A)** The represented images showed multiple recent subcortical infarcts, represented as hyperintensity on DWI in regions of left corpus callosum, bilateral paraventricular and semi-oval center. The four rows are raw data, ground truth, doctor's segmentation label and segmentation prediction from DLS, respectively. **(B)** Showed the comparison of accuracy ratio of segmentation label from doctors with 95% confidence interval and DLS.

rather than probabilistic diagnoses from doctors, rendering the comparison between our model and the doctors irrelevant. Adapting the concept of treating each lesion equally, we do away with the probabilistic element of FROC and compare the F1 score, of our model's predictions after thresholding (Goutte and Gaussier, 2005).

Moreover, Region-wise F1 score also applied in this study, it provides another avenue for us to answer the research question of how the predictions made by a deep learning model compares with that of human doctors (Goutte and Gaussier, 2005).

Another evaluation metric as a less demanding alternative to the dice score was applied in this study. We discretize the reference mask as well as predictions into square grids with spacing approximately equal to the square root of the image dimensions. Each patch, which may be viewed as bins mapped from a neighborhood of pixels, will be classified positive for the disease as long as at least one pixel in that patch is positive, or be classified negative otherwise. This is equivalent to performing a max-pooling followed by resizing back to the original number of pixels. In the limit where the patch is equal to the image size, the segmentation problem becomes converted to a multiple-label classification problem.

Statistical Analysis

The SPSS Statistics 23.0 software package for Windows (IBM Corp., Armonk, NY) was performed for statistical analyses.

RESULTS

It can be observed that our model possibly releases a prediction more faithful to the reference standard, compared to that of the doctors taking part in the clinical evaluation, regardless of

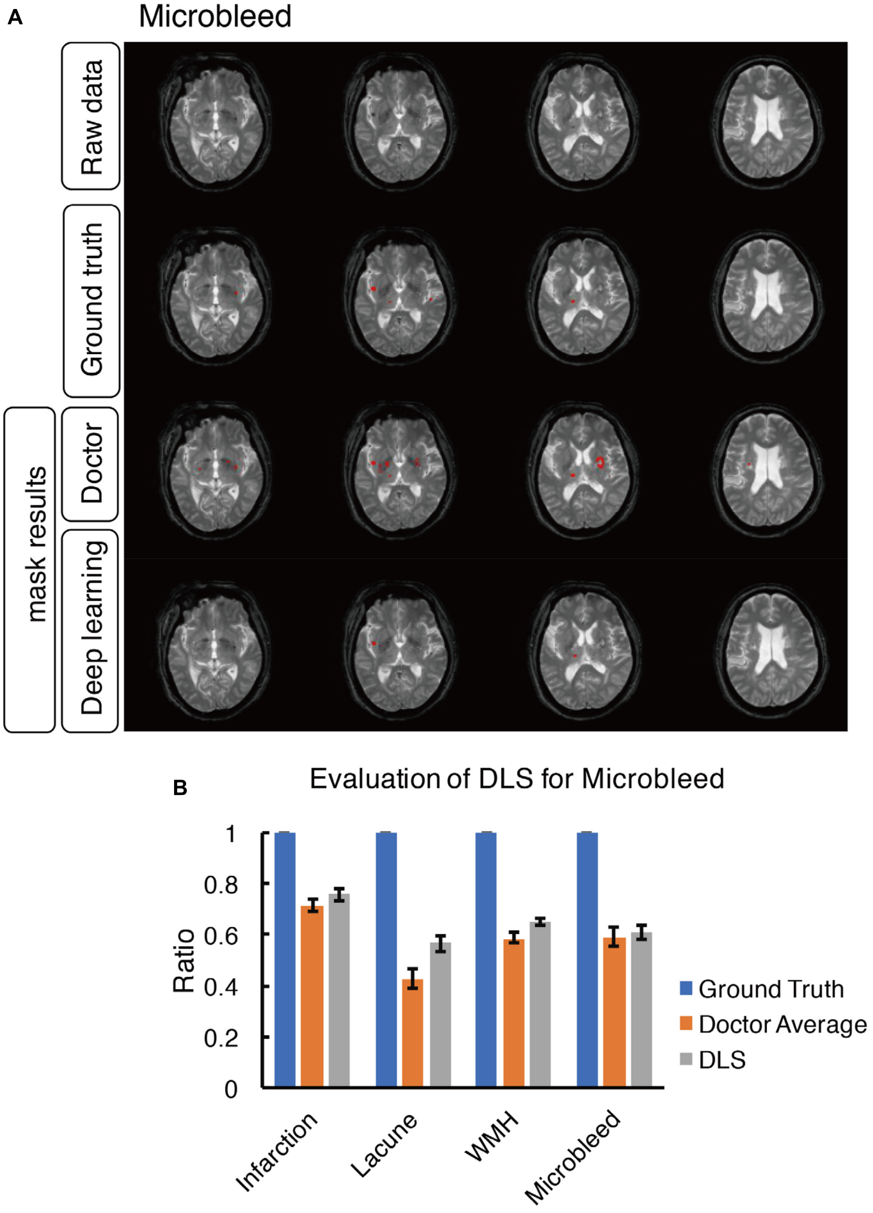


FIGURE 8 | Mask result of cerebral microbleed in the study, including Raw data, Ground truth, Doctor mask and DLS mask result. **(A)** showed cerebral microbleed lesions in the right insula and right thalamus, represented as hypointensity on T2*WI. The four rows are raw data ground truth, doctor's segmentation label and segmentation predation from DLS, respectively. **(B)** showed the comparison of accuracy ratio of segmentation label from doctors with 95% confidence interval and DLS.

whether the emphasis is placed on the segmentation or the detection of lesions. Where detailed pixel-level segmentation of lesions is required, our model's dice accuracy of 59.8% is over two percentage points better than the doctors' dice accuracy of 57.6% (**Table 3**). If the focus is on detect the presence of lesions, our model provides an average F1 score of over 72.5%, more than three percentage points over the doctors' 69.1% (**Table 3**).

Considering each of the four CSVD individually, the dice accuracy, as well as region-wise F1 score achieved by our model, is higher than that of the doctors in the segmentation of lacune, WMH and subcortical infarction, as can be verified from **Table 4** and **Figures 5–8**. Given that our model, as well as the doctors, perform best on the segmentation of subcortical infarction. Depending on how we define success in terms of pixel-wise segmentation or the detection of lesions, and the tolerance for uncertainties of a few pixels, our model, attains a score with 0.728 in dice accuracy and 0.859 in region-wise F1 score, which is consistently similar to the doctors' score with 0.714 in dice accuracy and 0.839 in region-wise F1 score (**Tables 4**, **5**).

TABLE 4 | Comparison of dice accuracy for different CSVDs.

	Lacune	White matter hyperintensity	Infarction	Cerebral microbleed
Doctor A	0.298	0.614	0.717	0.549
Doctor B	0.578	0.670	0.747	0.715
Doctor C	0.506	0.579	0.758	0.613
Doctor D	0.354	0.521	0.754	0.672
Doctor E	0.412	0.596	0.690	0.514
Doctor F	0.388	0.509	0.615	0.456
Average	0.423	0.582	0.714	0.586
Our model	0.496	0.666	0.728	0.503

Comparison of dice accuracy for different CSVDs of 30 patients according to the predictions made by six doctors (Neuroradiologists) and by our model, concerning the reference standard.

TABLE 5 | Comparison of region-wise F1 score for different CSVDs.

	Lacune	White matter hyperintensity	Infarction	Cerebral microbleed
Doctor A	0.375	0.660	0.817	0.785
Doctor B	0.676	0.722	0.905	0.897
Doctor C	0.633	0.661	0.905	0.809
Doctor D	0.41	0.668	0.921	0.797
Doctor E	0.518	0.603	0.836	0.662
Doctor F	0.536	0.518	0.652	0.623
Average	0.525	0.639	0.839	0.762
Our model	0.683	0.644	0.859	0.713

Comparison of region-wise F1 score for different CSVDs of 30 patients according to the predictions made by six doctors (Neuroradiologists) and by our model, with respect to reference standard. Each lesion is treated equally regardless of size, and the prediction is classified as a true positive when one or more pixel overlaps with the reference standard.

For each patient, our DLS system probably can process the images and output a volumetric prediction on the location of four CSVD diseases (lacune, WMH, subcortical infarction and cerebral microbleed) within a mean duration of 4.4 seconds (**Table 6**). The mean time used by each of the six doctors to draw masks of a single patient to produce a volumetric prediction ranges from 330 s to over 1,000 s. Compared to the segmentation independently made by six doctors, the predictions made by our model are over a hundred times faster and attained a higher dice accuracy and region-wise F1 score on average. Our DLS can suggest the diagnosis and draw the segmentation masks for over a hundred patients in the average time used by a doctor to do the same for one patient.

DISCUSSION

In this paper, by using T1-weighted, T2*, T2-FLAIR, and DWI b1000 images, we trained a DLS to draw the presented diseases area of lacune, WMH, subcortical infarction, and cerebral microbleed. We compare its performance with that of six doctors, using the reference standard set unanimously by three senior doctors. The results are evaluated based on the classical dice score, a modified patch-wise dice score, which allows for minor uncertainties in the neighborhood of a few pixels, as well as the region-wise F1 score, which may be a more suitable indication of success in the detection of lesions. The results show that our model can diagnose and draw the segmentation masks of multiple CSVDs more reliably, and over a hundred times faster than doctors with an average experience of eight years.

TABLE 6 | Credentials of doctors and time spent on the segmentation of 30 patients.

	Experience	Job title	Average time spent patient (in seconds, n = 30)
Doctor A	3 years	Resident Physician	1094/case
Doctor B	9 years	Attending Physician	662/case
Doctor C	18 years	Chief Physician	594/case
Doctor D	14 years	Chief Physician	418/case
Doctor E	3 years	Resident Physician	718/case
Doctor F	3 years	Resident Physician	330/case
Average	8 years		636/case
Our model			4.4/case

Credentials of doctors (Neuroradiologists) and time spent on the segmentation of 30 patients, including the drawing of segmentation masks, compared with the time required by our model to perform image processing and prediction.

This indicates that if patients trust the segmentation set by a panel of three senior doctors, they have reason to prefer the advice of our model over the opinion of an average doctor with few years of experience. It is also worthy to note that all six doctors are from Beijing Tiantan Hospital, which is a leading hospital in China and hosts one of the most extensive neurosurgical bases in China. Hence, these doctors are likely to be more rigorously trained than doctors from an average hospital in less affluent parts of China.

The results of our comparison support to the case that appropriately trained DLS can be trusted to the same extent as one would trust a doctor with a few years of experience, regardless of whether the emphasis is placed on the segmentation or detection of lesions. However, we want to emphasize that the proposed DLS is not aimed to replace doctors but meant to serve as a guide to doctors, where inconspicuous anomalies detected by the computer will warrant a closer look.

Limitation

In what follows, we discuss the limitations of our work and recommend possible improvements. First, as the testing dataset gets larger, the DLS is likely to have superior performance. However, it should be noted that, while all annotations made in the dataset have been endorsed by an associate chief physician with at least 10 years of experience, they are initially prepared by junior doctors with relatively less experience. Given that our model has been trained on these data, it is more likely to make predictions similar to these doctors rather than the senior doctors prescribing the reference standard. Had the model been trained on a vast number of images annotated by those senior doctors, its segmentation will likely bear a much closer resemblance to theirs.

Second, our model is compared against the performance of six doctors from a single country, and their average performance may not be representative of the average performance of all doctors globally. More doctors from a variety of hospitals and across different countries can be sought to participate in the clinical evaluation. Additionally, more patients can also be added to the evaluation dataset, so the results of our model, as well as the doctors, can be analyzed with greater certainty.

Third, our model is trained primarily to diagnose only four types of small vessel diseases. Therefore, our results cannot be generalized to compare the reliability of a DLS relative to the overall proficiency of a practicing doctor. Moreover, we do not deny the fact that our model is unable to propose a treatment, unlike a human doctor. Our study can be extended to train models capable of predicting a wide variety of medical anomalies. Besides, artificial intelligence is now progressing toward treatment planning and may be able to recommend solutions to their diagnosis in the future.

We reiterate that the purpose of this study is not to assert that DLS is more reliable than doctors. Instead, it is to propose that an adequately trained deep learning model can supplement the diagnosis of an attending doctor, and that one may heed its advice in the same way as one would respond to the words of a trained doctor.

CONCLUSION AND CONTRIBUTIONS

This study is a preliminary study focusing on lesion segmentation and identification. Previous studies showed the individual feature of CSVD is associated with incident ischemic and hemorrhagic stroke, dementia, and depression. Combinations of two features were more strongly associated with stroke than any specific feature (Pantoni, 2010; Go et al., 2012). So our model covered different types of lesions. According to the current results, the model can obtain lesion recognition at the level of attending physicians, which can significantly reduce the repetitive labor of physicians. For the further clinical application, with the help of this system, it may help clinical doctor fast categorizing and masking cerebral small vessel disease less time consuming, laborious, and subjective. Based on our DLS model, not only the location of the disease can be determined by the segmentation mask, but also the volume of lesions, which is critical in dosage prescription or clinical decision support systems (Belard et al., 2017).

DATA AVAILABILITY STATEMENT

The raw data supporting the conclusions of this article will be made available by the authors, without undue reservation, to any qualified researcher.

ETHICS STATEMENT

The studies involving human participants were reviewed and approved by Beijing Tiantan Ethics Committee. Written informed consent for participation was not required for this study in accordance with the national legislation and the institutional requirements. Written informed consent was not obtained from the individual(s) for the publication of any potentially identifiable images or data included in this article.

AUTHOR CONTRIBUTIONS

YD and WS wrote the initial draft of the manuscript. WS provided both figures and made preliminary revision. LL (Main contributor), PL, JJ, and ZW make contribution in the DLS development and medical test organization. LL, QW, ZW, PL, and JJ made preliminary revision. YL, KH, and YW made crucial revision. All authors together planned the manuscript, critically revised the initial draft, and made final improvements prior to submission.

FUNDING

This study is supported by National Key R&D Program of China (#2018YFC1312903, #2017YFC1310901, and #2017YFC1307500), funded by China National Neurological Clinical Research Centre, Beijing Postdoctoral Research Foundation (ZZ 2019-09), and China Postdoctoral Science Foundation (No. 2019M660719).

SUPPLEMENTARY MATERIAL

The Supplementary Material for this article can be found online at: https://www.frontiersin.org/articles/10.3389/fninf.2020.00017/full#supplementary-material

REFERENCES

Aslani, S., Dayan, M., Storelli, L., Filippi, M., Murino, V., Rocca, M. R., et al. (2019). Multi-branch convolutional neural network for multiple sclerosis lesion segmentation. *Neuroimage* 196, 1–15. doi: 10.1016/j.neuroimage.2019.03.068

Bandos, A. I., Rockette, H. E., Song, T., and Gur, D. (2009). Area under the free-response ROC curve (FROC) and a related summary index. *Biometrics* 65, 247–256. doi: 10.1111/j.1541-0420.2008.01049.x

Belard, A., Buchman, T., Forsberg, J., Potter, B. K., Dente, C. J., Kirk, A., et al. (2017). Precision diagnosis: a view of the clinical decision support systems (CDSS) landscape through the lens of critical care. *J. Clin. Monit. Comput.* 31, 261–271. doi: 10.1007/s10877-016-9849-1

Debette, S., and Markus, H. S. (2010). The clinical importance of white matter hyperintensities on brain magnetic resonance imaging: systematic review and meta-analysis. *BMJ* 341:c3666. doi: 10.1136/bmj.c3666

Dou, Q., Chen, H., Yu, L., Zhao, L., Qin, J., Wang, D., et al. (2016). Automatic detection of cerebral microbleeds from MR images via 3D convolutional neural networks. *IEEE Trans. Med. Imaging* 35, 1182–1195. doi: 10.1109/tmi.2016.2528129

Ghafoorian, M., Karssemeijer, N., Heskes, T., van Uden, I. W. M., Sanchez, C. I., Litjens, G., et al. (2017). Location sensitive deep convolutional neural networks for segmentation of white matter hyperintensities. *Sci. Rep.* 7:5110.

Go, A. S., Mozaffarian, D., Roger, V. L., Benjamin, E. L., Berry, J. D., Borden, W. B., et al. (2012). Heart disease and stroke statistics—2013 update a report from the American heart association. *Circulation* 127, e6–e245.

Goutte, C., and Gaussier, E. (2005). "A probabilistic interpretation of precision, recall and F-score, with implication for evaluation," in *Proceedings of the European Conference on Information Retrieval* (Berlin: Springer), 345–359. doi: 10.1007/978-3-540-31865-1_25

Greenberg, S. M., Vernooij, M. W., Cordonnier, C., Viswanathan, A., Salman, R. A.-S., Warach, S., et al. (2009). Cerebral microbleeds: a guide to detection and interpretation. *Lancet Neurol.* 8, 165–174.

Havaei, M., Davy, A., Warde-Farley, D., Biard, A., Courville, A., Bengio, Y., et al. (2017). Brain tumor segmentation with deep neural networks. *Med. Image Anal.* 35, 18–31. doi: 10.1016/j.media.2016.05.004

Havaei, M., Guizard, N., Chapados, N., and Bengio, Y. (2016). "HeMIS: hetero-modal image segmentation," in *Proceedings of the International Conference on Medical Image Computing and Computer-Assisted Intervention* (Cham: Springer), 469–477. doi: 10.1007/978-3-319-46723-8_54

Huang, S., Shen, Q., and Duong, T. Q. (2010). Artificial neural network prediction of ischemic tissue fate in acute stroke imaging. *J. Cereb. Blood Flow Metab.* 30, 1661–1670. doi: 10.1038/jcbfm.2010.56

Kamnitsas, K., Chen, L., Ledig, C., Rueckert, D., and Glocker, B. (2015). Multi-scale 3D convolutional neural networks for lesion segmentation in brain MRI. *MiCCAi Brain Lesion Work* 2015, 13–16.

Kamnitsas, K., Ledig, C., Newcombe, V. F. J., Simpson, J. P., Kane, A., Menon, D. K., et al. (2017). Efficient multi-scale 3D CNN with fully connected CRF for accurate brain lesion segmentation. *Med. Image Anal.* 36, 61–78. doi: 10.1016/j.media.2016.10.004

Misra, U. K., Kalita, J., Phadke, R. V., Wadwekar, V., Boruah, D. K., Srivastava, A., et al. (2010). Usefulness of various MRI sequences in the diagnosis of viral encephalitis. *Acta Trop.* 116, 206–211. doi: 10.1016/j.actatropica.2010.08.007

Mohsen, H., El-Dahshan, E. S. A., El-Horbaty, E. S. M., and Salem, A.-B. M. (2018). Classification using deep learning neural networks for brain tumors. *Future Comput. Inform. J.* 3, 68–71.

Noguchi, K., Ogawa, T., Inugami, A., Fujita, H., Hatazawa, J., Shimosegawa, E., et al. (1997). MRI of acute cerebral infarction: a comparison of FLAIR and T2-weighted fast spin-echo imaging. *Neuroradiology* 39, 406–410. doi: 10.1007/s002340050433

Pantoni, L. (2010). Cerebral small vessel disease: from pathogenesis and clinical characteristics to therapeutic challenges. *Lancet Neurol.* 9, 689–701. doi: 10.1016/s1474-4422(10)70104-6

Rensma, S. P., van Sloten, T. T., Launer, L. J., and Stehouwer, C. D. A. (2018). Cerebral small vessel disease and risk of incident stroke, dementia and depression, and all-cause mortality: a systematic review and meta-analysis. *Neurosci. Biobehav. Rev.* 90, 164–173. doi: 10.1016/j.neubiorev.2018.04.003

Rincon, F., and Wright, C. (2014). Current pathophysiological concepts in cerebral small vessel disease. *Front. Aging Neurosci.* 6:24. doi: 10.3389/fnagi.2014.00024

Ronneberger, O., Fischer, P., and Brox, T. (2015). "U-net: convolutional networks for biomedical image segmentation," in *Proceedings of the International Conference on Medical Image Computing and Computer-Assisted Intervention* (Cham: Springer), 234–241. doi: 10.1007/978-3-319-24574-4_28

Stier, N., Vincent, N., Liebeskind, D., and Scalzo, F. (2015). Deep learning of tissue fate features in acute ischemic stroke. *IEEE Int. Conf. Bioinform. Biomed.* 5, 1316–1321.

Sudre, C. H., Li, W., Vercauteren, T., Ourselin, S., and Cardoso, M. J. (2017). "Generalised Dice overlap as a deep learning loss function for highly unbalanced segmentations," in *Deep Learning in Medical Image Analysis and Multimodal Learning for Clinical Decision Support*, ed. M. Cardoso (Cham: Springer), 240–248. doi: 10.1007/978-3-319-67558-9_28

Wardlaw, J. M., Smith, E. E., Biessels, G. J., Cordonnier, C., Fazekas, F., Frayne, R., et al. (2013). Neuroimaging standards for research into small vessel disease and its contribution to ageing and neurodegeneration. *Lancet Neurol.* 12, 822–838.

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 Computer Science

ORIGINAL RESEARCH
published: 19 October 2020
doi: 10.3389/fcomp.2020.546917

Extraction of Hierarchical Behavior Patterns Using a Non-parametric Bayesian Approach

Jeric Briones, Takatomi Kubo and Kazushi Ikeda*

Division of Information Science, Nara Institute of Science and Technology, Ikoma, Japan

Extraction of complex temporal patterns, such as human behaviors, from time series data is a challenging yet important problem. The double articulation analyzer has been previously proposed by Taniguchi et al. to discover a hierarchical structure that leads to complex temporal patterns. It segments time series into hierarchical state subsequences, with the higher level and the lower level analogous to words and phonemes, respectively. The double articulation analyzer approximates the sequences in the lower level by linear functions. However, it is not suitable to model real behaviors since such a linear function is too simple to represent their non-linearity even after the segmentation. Thus, we propose a new method that models the lower segments by fitting autoregressive functions that allows for more complex dynamics, and discovers a hierarchical structure based on these dynamics. To achieve this goal, we propose a method that integrates the beta process—autoregressive hidden Markov model and the double articulation by nested Pitman-Yor language model. Our results showed that the proposed method extracted temporal patterns in both low and high levels from synthesized datasets and a motion capture dataset with smaller errors than those of the double articulation analyzer.

Keywords: behavioral pattern, non-parametric Bayesian approach, segmentation, hierarchical structure, dynamics

OPEN ACCESS

Edited by:
Priit Kruus,
Tallinn University of Technology,
Estonia

Reviewed by:
Dimitrije Marković,
Technische Universität Dresden,
Germany
Chee-Ming Ting,
University of Technology Malaysia,
Malaysia

****Correspondence:***
Takatomi Kubo
takatomi-k@is.naist.jp

Specialty section:
This article was submitted to
Digital Public Health,
a section of the journal
Frontiers in Computer Science

Received: *30 March 2020*
Accepted: *14 September 2020*
Published: *19 October 2020*

Citation:
Briones J, Kubo T and Ikeda K (2020)
Extraction of Hierarchical Behavior
Patterns Using a Non-parametric
Bayesian Approach.
Front. Comput. Sci. 2:546917.
doi: 10.3389/fcomp.2020.546917

1. INTRODUCTION

In the big data era, we can easily collect information-rich time series thanks to the advancements in sensing technologies. However, such time series data are not segmented and hence difficult to apply recent machine learning techniques. To segment such data, extraction of temporal patterns in an unsupervised manner is necessary. This has become an active topic in several research fields, such as health care (Zeger et al., 2006), biology (Saeedi et al., 2016), speech recognition (Taniguchi et al., 2016), natural language processing (Heller et al., 2009), and motion analysis (Barbič et al., 2004). Although many methods have been proposed to extract temporal patterns (Keogh et al., 2004), there exists a problem that the number of existing patterns (and consequently, the number of segments) is generally unknown beforehand. To solve this issue, non-parametric Bayesian methods are used to determine the number of patterns (Fox et al., 2008b). Specifically, non-parametric Bayesian methods based on switching AR models, such as the beta process—autoregressive hidden Markov model (BP-AR-HMM) (Fox et al., 2009, 2014), can be used to identify the temporal patterns without specifying the number of patterns beforehand.

Although conventional methods can discover the temporal patterns to segment a time-series sequence, some sequences have a hierarchical structure that makes the segmentation more complex. Motion data, for example, can be seen as a sequence of semantic actions, where each action can be decomposed as a series of *motion primitives* (Viviani and Cenzato, 1985; Zhou et al., 2013; Grigore and Scassellati, 2017). Similarly, speech data consist of words, where each word consists of phonemes. With such a hierarchical structure, usual methods involving switching dynamical systems may not be sufficient since they do not assume the existence of the hierarchical structure. Time series sequences like the examples above should then be analyzed using hierarchical models. Non-parametric Bayesian methods for hierarchical models include the hierarchical hidden Markov model (HHMM) (Fine et al., 1998), the nested Pitman-Yor language model (NPYLM) for sentences (Mochihashi et al., 2009), and the double articulation analyzer (DAA) (Taniguchi and Nagasaka, 2011). However, they are not suitable for analyzing the dynamic patterns. For example, DAA only modeled the time series sequences by fitting segment-wise linear functions to the lower level of the structure. Complex dynamics in the lower level has not been considered in the previous method, despite motion primitives being usually modeled as non-linear functions (Williams et al., 2007; Bruno et al., 2012).

From these backgrounds, it is necessary to develop a method that considers both dynamics and hierarchical structure to extract temporal patterns. To realize such a method, we naively applied BP-AR-HMM and NPYLM in order to model hierarchically-structured sequences with dynamical systems in our previous study (Briones et al., 2018).

In this work, we propose a model that integrates BP-AR-HMM and NPYLM as a unified model. Our method can capture the hierarchical structure of the time series by NPYLM and use dynamical systems (specifically, switching AR models in BP-AR-HMM) to represent the dynamic pattern in the lower level sequences. Also, BP-AR-HMM allows for asynchronous switching of segments across the multiple time series data considered thanks to the beta process in BP-AR-HMM. Compared to our previous two-step approach, the proposed integrated approach is expected to improve segmentation and estimation accuracy. In this study, we tested our method with toy dataset and sequences generated from real motion capture (mocap) sequences with two interacting agents. Such motion sequences are suitable to test the segmentation performance of our method, since interaction switches from time to time (Ryoo and Aggarwal, 2009; Alazrai et al., 2015).

The rest of this paper is organized as follows: section 2 shows our proposed method, with a brief introduction of basic algorithms. Sections 3 and 4 outline the details of the synthetic experiments carried out using two datasets and their corresponding results. Finally, section 5 gives some discussion of the results, including the conclusions.

2. PROPOSED METHOD

We propose to use a hierarchical non-parametric Bayesian approach to extract hierarchical temporal patterns from time series data. Specifically, we use an unsupervised segmentation method, where the extracted segments are used to define the temporal patterns. Our method consists of two non-parametric Bayesian models: BP-AR-HMM (Fox et al., 2009) and NPYLM (Mochihashi et al., 2009) (**Figure 1**).

In the first step, BP-AR-HMM is applied to time series data, to discover low-level temporal patterns or elemental behaviors (EB), which correspond to the motion primitives in the motion analysis. Segmentation is indicated by assigning EB labels at each time step. The obtained EB label sequences for each time series are then summarized, before being used as an input for the second step. In the second step, NPYLM is applied to the (summarized) sequence of EB labels, to detect unit behaviors (UB). Subsequences of EB labels with recurring patterns are grouped together and assigned UB labels. As a consequence, the method outputs a sequence of UBs, each of which is a sequence of EBs represented by AR models. Then, these two steps are iterated a fixed number of times, with the resulting UB labels from the NPYLM step used as initial EB labels for the BP-AR-HMM step of the next iteration.

In the following, we introduce the components of our method: BP-AR-HMM and NPYLM.

2.1. BP-AR-HMM

BP-AR-HMM is an extension of hidden Markov model where each discrete latent variable z_t has an AR model of order r with parameter $\theta_{z_t} = \{\mathbf{A}_{z_t}, \Sigma_{z_t}\}$ (**Figure 2**), and the observed variable $\mathbf{y}_t$ is represented as an AR model with lag order r. This model is a non-parametric Bayesian model with a beta process prior, where an indicator vector over the set of EBs, $\mathbf{f}_i$, is drawn. The EB z_t, the state transition matrix $\pi_j^{(i)}$, and the AR coefficient matrix $\mathbf{A}_k$ are drawn according to $\mathbf{f}_i$, a gamma prior, and a matrix normal prior, respectively (**Figure 2**).

2.1.1. Beta Process (BP)

A beta process prior is placed on the EB indicator vector. This makes it possible to not specify the number of EBs beforehand, and thus allow us to use an infinite-dimensional EB indicator vector $\mathbf{f}$. A beta process is a completely random measure, denoted by

$$B \mid c, B_0 \sim \mathrm{BP}(c, B_0), \tag{1}$$

$$B = \sum_{k=1}^{\infty} \omega_k \delta_{\theta_k}, \qquad \text{with } \alpha = B_0(\Theta), \tag{2}$$

where B_0 is a base measure, c the concentration parameter, and α the mass parameter. The number of active EBs, including which EBs are active, for time series i is determined by a realization of the indicator vector $\mathbf{f}_i \mid B \sim \mathrm{BeP}(B)$, given by

$$\mathbf{f}_i = \sum_k f_{ik} \delta_{\theta_k}, \qquad \text{with } f_{ik} \sim \mathrm{Be}(\omega_k). \tag{3}$$

Here, $f_{ik} = 1$ if the kth EB is active for time series i, $i = 1, \ldots N$.

2.1.2. AR-HMM

The D-dimensional observation vector $\mathbf{y}_t$ is described by an autoregressive hidden Markov model (AR-HMM), with order

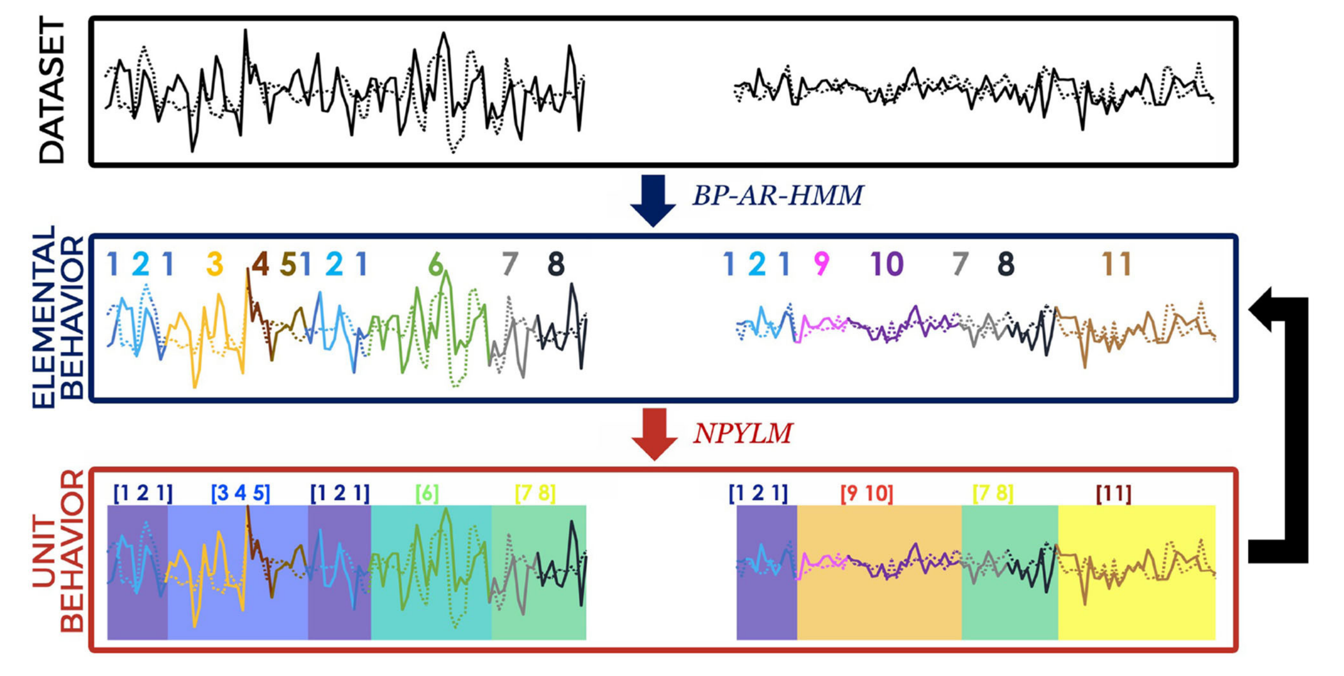


FIGURE 1 | Illustration of processing steps in our proposed method. Each time step is assigned both an EB label (line color) and UB label (background color). The summarized sequence of EB labels (shown as numbers above the lines) obtained from BP-AR-HMM are then grouped together (indicated by the square brackets) using NPYLM.

r, latent variable (*state sequences)* z_t, and transition probability matrix π_k. That is,

$$z_t \mid z_{t-1} \sim \pi_{z_{t-1}}, \tag{4}$$

$$\mathbf{y}_t \mid z_t \sim \mathcal{N}\left(\mathbf{A}_{z_t}\tilde{\mathbf{y}}_t, \Sigma_{z_t}\right), \tag{5}$$

$$\mathbf{y}_t = \sum_{l=1}^{r} A_{l,z_t}\mathbf{y}_{t-l} + \mathbf{e}_t(z_t), \quad \text{with } \mathbf{e}_t(z_t) \sim \mathcal{N}\left(0, \Sigma_{z_t}\right) \tag{6}$$

For the kth EB, the corresponding AR-HMM parameters are denoted as $\theta_k = \{\mathbf{A}_k, \Sigma_k\}$, while the transition probabilities are denoted by π_k. Since active EBs vary for each sequence, *feature-constrained* transition distributions (Fox et al., 2009) are used. That is, given $\mathbf{f}_i$,

$$\pi_{kj}^{(i)} = \begin{cases} 0 & f_{ij} = 0 \\ P\left(z_t^{(i)} = j \mid z_{t-1}^{(i)} = k\right) & f_{ij} = 1 \end{cases}, \quad \text{with } \sum_j \pi_{kj}^{(i)} = 1. \tag{7}$$

A gamma prior would be placed on the transition matrix, with

$$\eta_{jk} \mid \gamma, \kappa \sim \text{Gamma}\left(\gamma + \kappa\delta_{j,k}, 1\right), \tag{8}$$

$$\pi_j^{(i)} = \frac{\boldsymbol{\eta}_j^{(i)} \otimes \mathbf{f}_i}{\sum_{k|f_{ik}=1} \eta_{jk}^{(i)}}, \tag{9}$$

where γ, κ are the transition and transition sticky parameter, respectively. Moreover, $\delta_{j,k}$ is the Kronecker delta function, and $\otimes$ is the Hadamard (or element-wise) vector product.

Moreover, matrix normal priors would be placed on the dynamic parameters. That is,

$$\Sigma_k \mid n_0, S_0 \sim \text{IW}\left(S_0, n_0\right), \tag{10}$$

$$\mathbf{A}_k \mid \Sigma_k, M, L \sim \mathcal{MN}\left(\mathbf{A}_k; M, \Sigma_k, L\right), \tag{11}$$

where n_0 is the degrees of freedom, S_0 a scale matrix, M the mean dynamic matrix, and L, Σ_k defines covariance of $\mathbf{A}_k$.

2.1.3. Posterior Inference

Samples are generated from the posterior distribution using Markov chain Monte Carlo (MCMC) algorithm. To be specific, the samples to be produced are the EB indicator vector $\mathbf{f}$ given $\boldsymbol{\theta}, \boldsymbol{\eta}$, state sequences $\mathbf{z}$ given $\mathbf{f}$, $\boldsymbol{\theta}$, $\boldsymbol{\eta}$, and variables $\boldsymbol{\theta}$, $\boldsymbol{\eta}$ given $\mathbf{f}$ and $\mathbf{z}$. The hyperparameters $\alpha, c, \kappa, \gamma$ would also be sampled. Basically, MCMC alternates between sampling $\mathbf{f}|\mathbf{y}, \boldsymbol{\theta}$ and $\boldsymbol{\theta}|\mathbf{y}, \mathbf{f}$, with the hyperparameters sampled in between the cycles. To generate unique EB vectors, birth-death reversible jump MCMC sampling (Fox et al., 2009) and split-merge techniques (Hughes et al., 2012) would be utilized. These samples would then be used to carry out posterior inference.

2.1.4. Advantages

Using this model provides several advantages over the sticky hierarchical Dirichlet process—HMM (sticky HDP-HMM) (Fox et al., 2008a) used in DAA. First, we can segment multiple time series, and discover common and unique behaviors from these sequences, thanks to the BP prior. This would not be possible if we use sticky HDP-HMM since it would require all the time series sequences to share exactly the same behaviors (and not just a

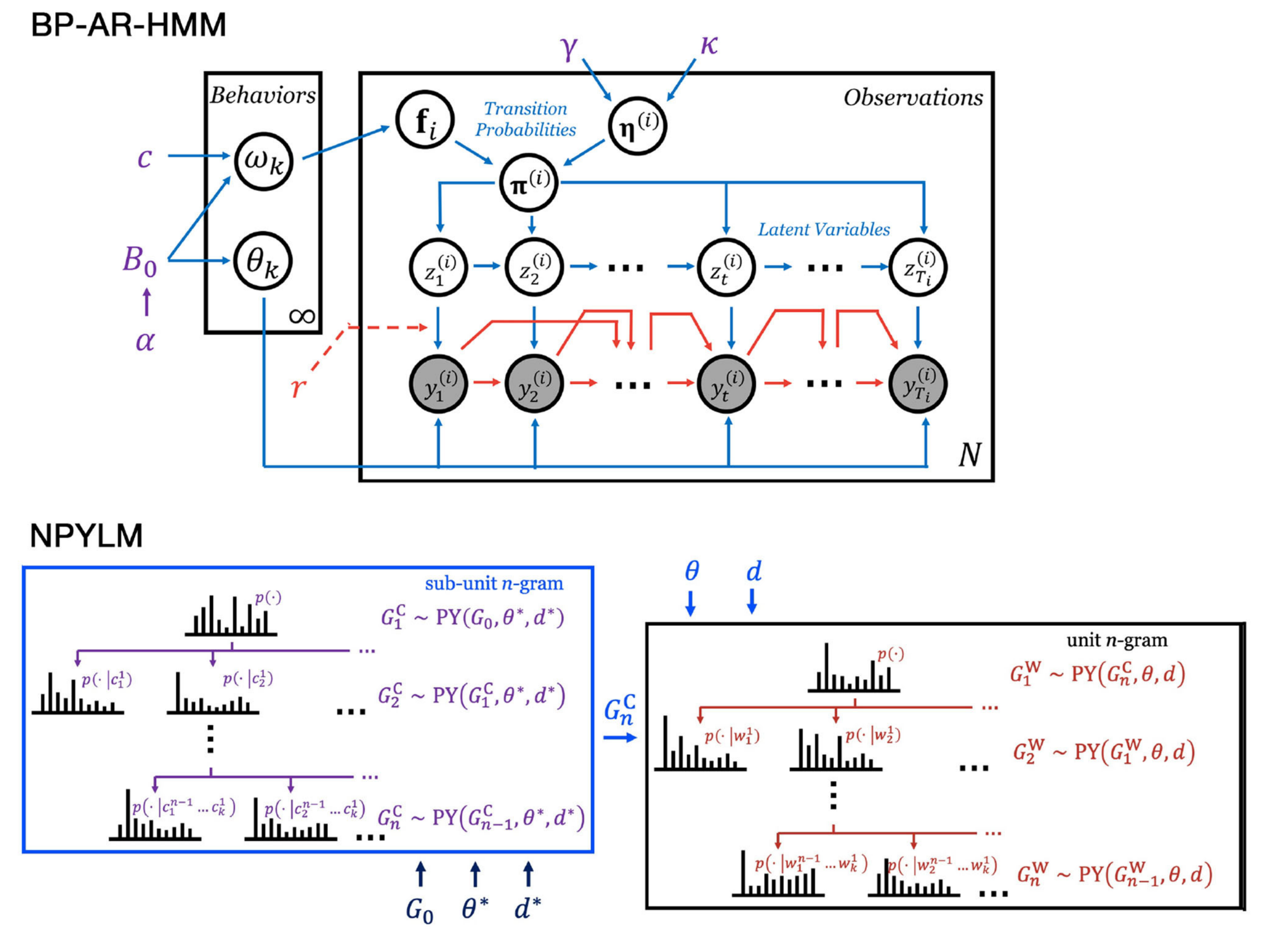


FIGURE 2 | Graphical representations of the non-parametric Bayesian models. (Top) BP-AR-HMM. (Bottom) NPYLM.

subset of it). The difference between BP and HDP is most evident on the transition probability matrices used for each sequence. HDP in HDP-HMM assigns a state to each time step according to a transition matrix shared by all time series, while BP in BP-AR-HMM assigns a state according to transition matrix specific to each sequence.

Furthermore, using an AR model also allows us to discover the dynamic properties of the data. This is again not present in DAA. Specifically, BP-AR-HMM fits AR models for the given time series $\{\mathbf{y}_t\}$. Hence, the interactions among the variables are expressed in its AR coefficient matrix $\mathbf{A}_k$ (Harrison et al., 2003; Gilson et al., 2017), making our method suitable for subsequent interaction analysis.

2.2. NPYLM

NPYLM is originally proposed as a hierarchical language model where both letters and words are modeled by hierarchical Pitman-Yor processes (Mochihashi et al., 2009; Neubig et al., 2010). In each layer of the hierarchical model, words and letters are modeled as n-grams, which are produced by Pitman-Yor processes. In general, words can be considered as high-level unit segments (UB in this study), while letters as low-level sub-unit segments (EB in this study). Similar to how words are made up of letters, these high-level unit segments are also composed of low-level sub-unit segments.

2.2.1. Pitman-Yor Process

Pitman-Yor (PY) process is a stochastic process that generates probability distribution G that is similar to a base distribution G_0. This is denoted by

$$G \mid G_0, \theta, d \sim \mathrm{PY}(G_0, \theta, d) \tag{12}$$

where G_0 is a base measure, θ the concentration parameter, and d the discount parameter.

2.2.2. Hierarchical Pitman-Yor Language Model

Given a unigram distribution G_1^{W}, we can generate a bigram distribution G_2^{W} such that this distribution will be similar G_1^{W}, especially for the high-frequency units. That is, $G_2^{\mathrm{W}} \sim \mathrm{PY}(G_1^{\mathrm{W}}, \theta, d)$. Similarly, a trigram distribution can also be

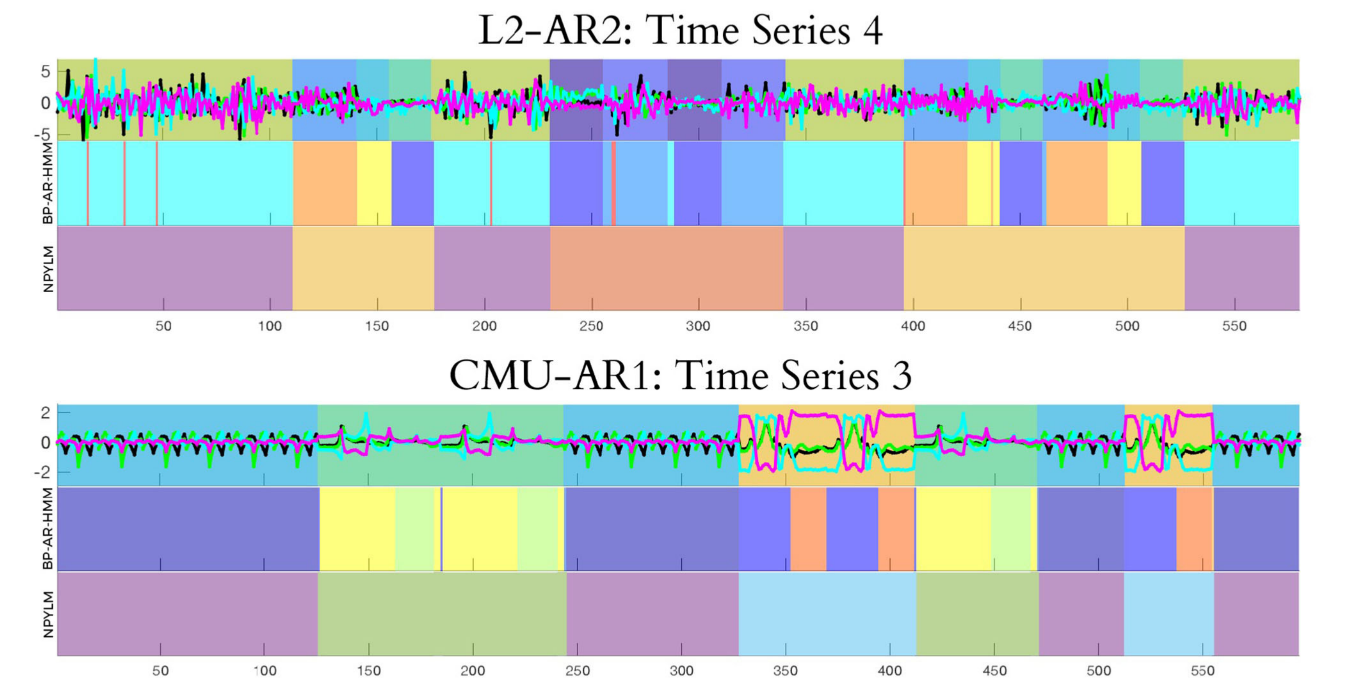


FIGURE 3 | Plot of the first four dimensions of the input data and an example of segmentation results for both EB and UB steps. (Top) Toy data. (Bottom) CMU motion data.

generated similar to the bigram distribution, such that $G_3^{\mathrm{W}} \sim \mathrm{PY}(G_2^{\mathrm{W}}, \theta, d)$. In general, then, the n-gram model is Pitman-Yor distributed with base measure from the $(n-1)$-gram model, and the base measure of the unigram model being G_0^{W}. This hierarchical structure of n-gram models is referred to as hierarchical Pitman-Yor language model (HPYLM).

Specifically, for the *unit* n-gram model, the probability of a unit $w = w_t$ given a context $h = w_{t-n} \ldots w_{t-1}$ is calculated recursively as

$$p\left(w \mid h\right) = \frac{c\left(w \mid h\right) - d \cdot t_{hw}}{\theta + c(h)} + \frac{\theta + d \cdot t_h}{\theta + c(h)} p\left(w \mid h'\right), \quad (13)$$

where $h' = w_{t-n-1} \ldots w_{t-1}$ is the shorter $(n-1)$-gram context, $c\left(w \mid h\right)$ is a count of w under context h, and $c\left(h\right) = \sum_w c\left(w \mid h\right)$. Here, $p\left(w \mid h'\right)$ can be considered as a prior probability of w. On the other hand, t_{hw} is a count under the context h', while $t_h = \sum_w t_{hw}$ is a count under the context h. Finally, d, θ are the discount and concentration parameters, respectively.

To define the base measure G_0^{W} for the unit unigram model (and consequently define $p\left(w \mid h'\right)$ for G_1^{W}), NPYLM uses a *sub-unit* n-gram model G_n^{C} as the aforementioned base measure. This sub-unit n-gram model G_n^{C} also uses hierarchical Pitman-Yor processes, and is structured similarly to the unit n-gram model G_n^{W}. Moreover, the probability for the sub-unit n-gram is also calculated recursively using Equation (13), where G_0, d^*, θ^* are the base measure, discount parameter, and concentration parameter for sub-unit unigram model, respectively. As a result, an HPYLM (in this case, the sub-unit n-gram) is actually embedded inside another HPYLM (the unit n-gram), resulting to the "nested" part of NPYLM.

2.2.3. Posterior Inference

Samples are generated from the posterior distribution using Gibbs sampling and forward filtering-backward sampling (Mochihashi et al., 2009; Neubig et al., 2010; Taniguchi and Nagasaka, 2011). To be specific, a unit is removed from the current *unit* n-gram model, then a "new" unit is sampled by generating a new segmentation of the sequence of sub-units. The "new" unit is then added to the *unit* n-gram model, thereby updating the said model. This process of blocked Gibbs sampling is repeated several times, with forward filtering-backward sampling used to generate new segmentation.

2.2.4. Advantages

This model assumes that the input sequence has a hierarchical structure. Thanks to this hierarchical structure, NPYLM is suitable to model motion data composed of a sequence of UBs, each of which is composed of a sequence of EBs. This second step allows us to have high-level semantic, more meaningful behaviors, rather than the low-level short, simple behaviors (akin to motion primitives).

Moreover, since NPYLM is an unsupervised language model, using this model in the second step enables us to do segmentation without having an existing dictionary. In addition, using blocked Gibbs sampler significantly reduces computational time for the

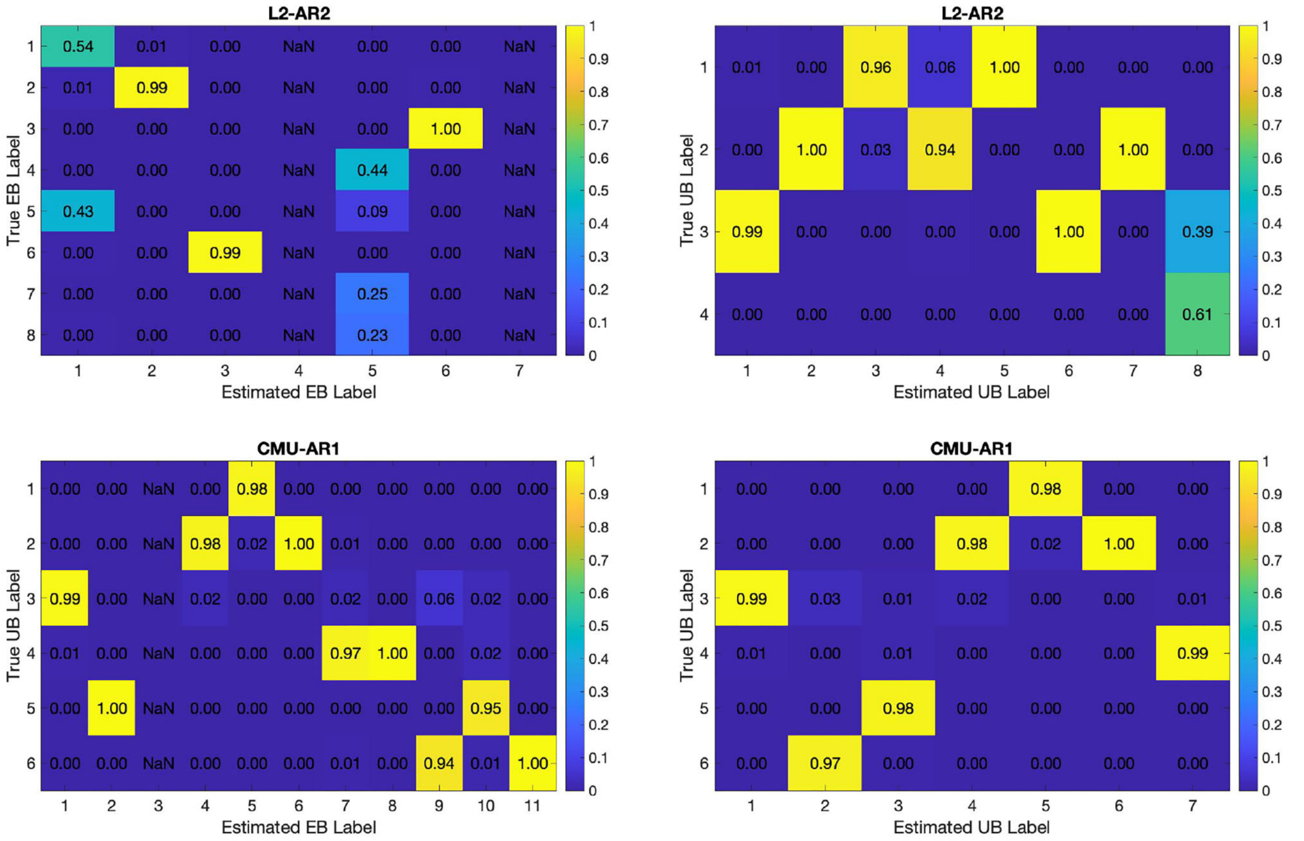


FIGURE 4 | Confusion matrices for an example of segmentation results for both EB and UB steps. Numbers normalized per column. (Top) Toy data. (Bottom) CMU motion data.

sampling procedure (Mochihashi et al., 2009; Taniguchi and Nagasaka, 2011).

3. SYNTHETIC EXPERIMENTS

We carried out experiments with two datasets to check the performance of our method and compare it with that of DAA. One was a toy dataset synthesized from known AR models to evaluate the estimation accuracy for segments using the ground truth. Using this dataset, we also investigated the effects of complexity (AR order) of the time series.

3.1. Toy Data

To evaluate the estimation accuracy, three subdatasets, $\mathbf{L}\boldsymbol{m}$ ($m = 1, 2, 3$), were generated from switching m-th lag order AR models with hierarchical structure. UBs were randomly chosen from a library of four UBs (based on predefined transition probability matrices), to form sequences of concatenated UBs. Each UB consists of several EBs, where each EB has sparse AR(m) coefficient matrices, generated independently for each subdataset. Elements of the AR coefficient matrices were set within the range $(-1, 1)$. EBs under the same UB share the same sparsity structure for their respective AR coefficient matrices. Finally, each subdataset $\mathbf{L}\boldsymbol{m}$ ($m = 1, 2, 3$) has four time series sequences of four dimensions each.

Our method with r-th AR ($r = 1, 2, 3$) was then applied to each subdataset $\mathbf{L}\boldsymbol{m}$ ($m = 1, 2, 3$). We carried out our proposed segmentation method for thirty runs, where each run had ten different chains of sampling. In each chain, the UB labels obtained from the prior iteration were used as the initial EB labels of the next one, with the first chain having all time steps (across all sequences) assigned the same initial EB label.

The hyperparameter setting of BP-AR-HMM and NPYLM were based on the values used in Fox et al. (2014) (for BP-AR-HMM) and Neubig et al. (2010) (for NPYLM), respectively. The parameters of BP-AR-HMM were set as follows: the concentration parameter $c = 3$, the mass parameter $\alpha = 2$, both with Gamma$(1, 1)$ prior, for the beta process; the transition parameter $\gamma = 1$, the transition sticky parameter $\kappa = 25$, with Gamma$(1, 1)$ and Gamma$(100, 1)$ prior, respectively, for the transition matrix. The first 5,000 samples of the MCMC algorithm were discarded as burn-in, and the following 5,000 samples were used. The state sequences were summarized in each of the thirty runs, where states with associated time shorter than 1% of the total time were discarded. These were then forwarded to NPYLM. Settings for NPYLM were as follows: the discount parameter $d = 0.5$ with Beta$(1.5, 1)$ prior; the concentration

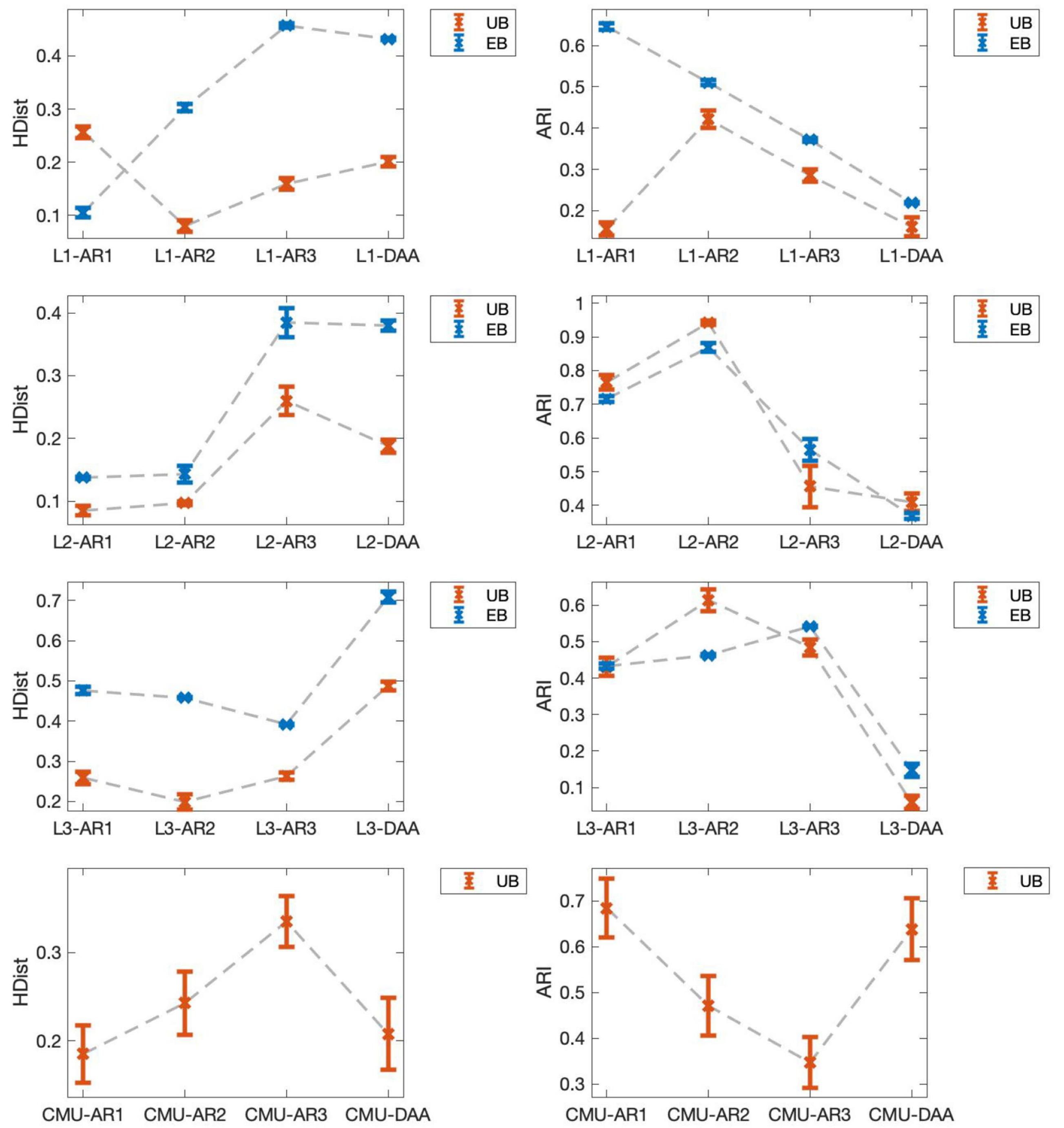


FIGURE 5 | Evaluation of model accuracy. (Left) Average normalized Hamming distances. Bars: 1 SE. Lower value is better. (Right) Average adjusted Rand index. Bars: 1 SE. Higher value is better.

parameter $\theta = 0.1$ with Gamma(10, 0.1) prior. The first 5,000 samples of the blocked Gibbs sampling were discarded as burn-in, and the following 5,000 samples were used. Posterior inference for BP-AR-HMM was carried out using the codes developed by Hughes (2016), while the codes developed by Neubig (2016) were used to carry out posterior inference for NPYLM.

Finally, our method was also compared with DAA. For DAA, state sequences were also summarized in each of the thirty runs of each subdataset $\mathbf{L}\boldsymbol{m}$ ($m = 1, 2, 3$). The parameters of DAA were set so that they were comparable to those of our method. As sticky HDP-HMM is usually for single time series sequences, the time series were concatenated into one long time series before being

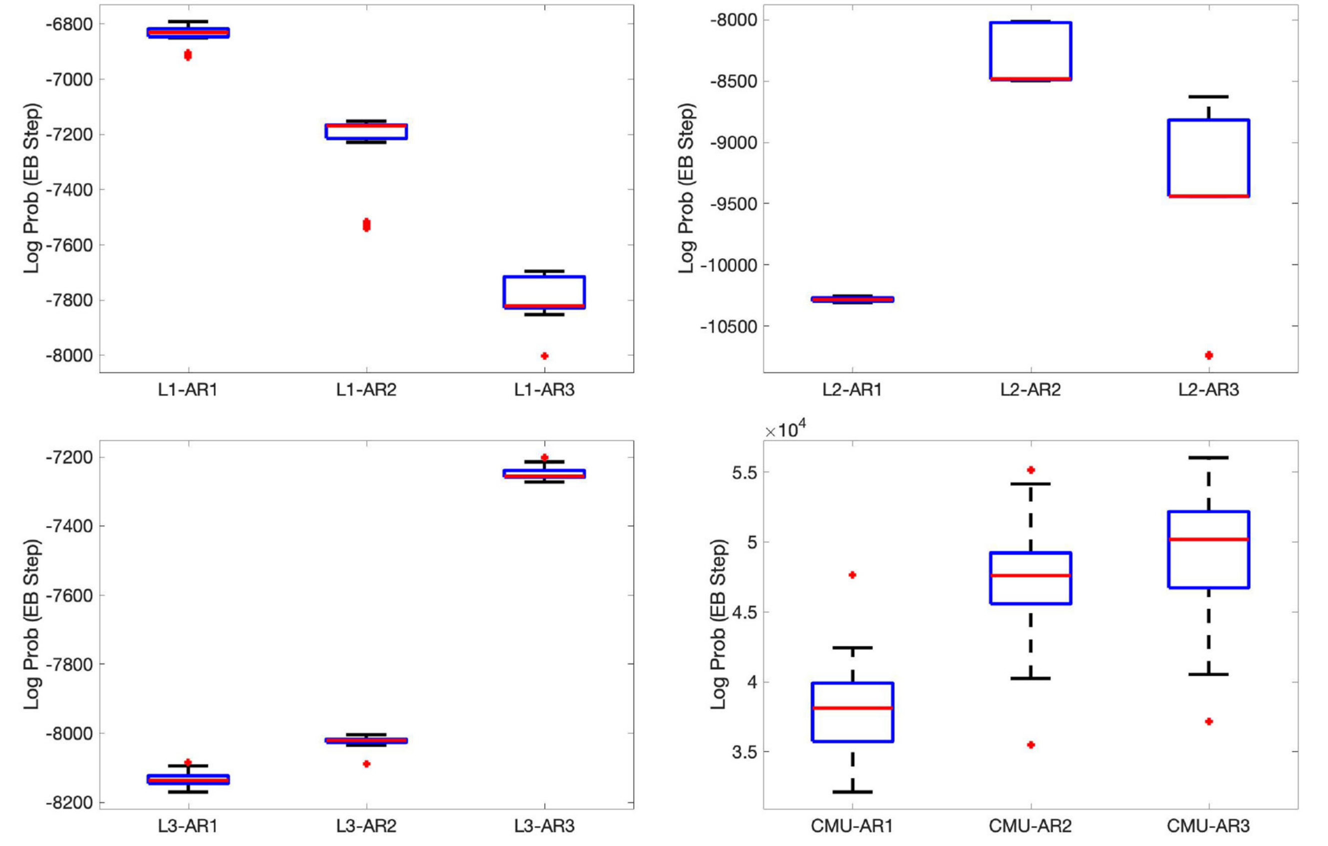


FIGURE 6 | Boxplots of the joint log probabilities $P(\mathbf{y}, \mathbf{F}, \mathbf{z})$ of the EB step. Red line: median. Edges of blue box: first and third quartiles. Whiskers: most extreme data points except outliers.

applied to the first step of DAA. The EB labels were then split back and summarized afterward, where states with associated time shorter than 1% of the total time discarded. DAA was carried out using the codes recommended in http://daa.tanichu.com/code.

3.1.1. Large-Scale Toy Data

We generated three additional subdatasets, $\mathbf{T}s$ ($s = 10, 20, 100$), to explore how our proposed method would fair on a large-scale simulation. The subdatasets were generated from switching AR(1) models, using the same parameter settings described earlier, but with s ($s = 10, 20, 100$) time series sequences instead. Our method with AR(1) was applied to subdataset $\mathbf{T}s$ ($s = 10, 20, 100$), using the same settings described above, but with ten, ten, and three runs for **T10**, **T20**, and **T100**, respectively, where each run still had ten different chains of sampling.

3.2. Evaluation With Toy Data

The result of our method with AR(r) applied to the toy subdataset **L***m* was denoted by **L***m***-AR***r*, while the result of DAA was denoted by **L***m***-DAA**. Our method segmented time series with high accuracy (with respect to both EBs and UBs), and had better accuracy than DAA. The top panel of **Figure 3** shows a visualization example of segmentation results. In the panel, the background color of the top row indicates its own EB. UBs are represented by sequential patterns consisting of sets of EBs. The second and third rows show the estimated EBs and UBs, respectively. Their boundaries show high consistency to the ground truth.

Confusion matrices for the EB labels (left) and UB labels (right) of an example segmentation results are also illustrated in **Figure 4**. Here, the correspondence between true labels and estimated labels is represented with the values normalized per column, to allow one-to-many correspondence from a true label to estimated labels. Columns with entries close to 1 indicate that corresponding estimated labels are assigned with high specificity, while rows with multiple entries close to 1 indicate multiple estimated labels correspond to one true label.

Next, we evaluated the effects of model mismatch to segmentation, using the resulting average normalized Hamming distances between the true EBs and the estimated EBs (EB HDist). Hamming distances were computed as the total number of time steps where the estimated label was different from the ground truth in a given sequence, then summed for all time series sequences in each run. Then, normalization was done by dividing the Hamming distance by the total number of time steps of the given sequence The EB HDist were smallest when the true AR order was used. For example, EB HDist of **L1-AR1** were smaller than those of **L1-AR2** and **L1-AR3**, as seen in **Figure 5**. Similar

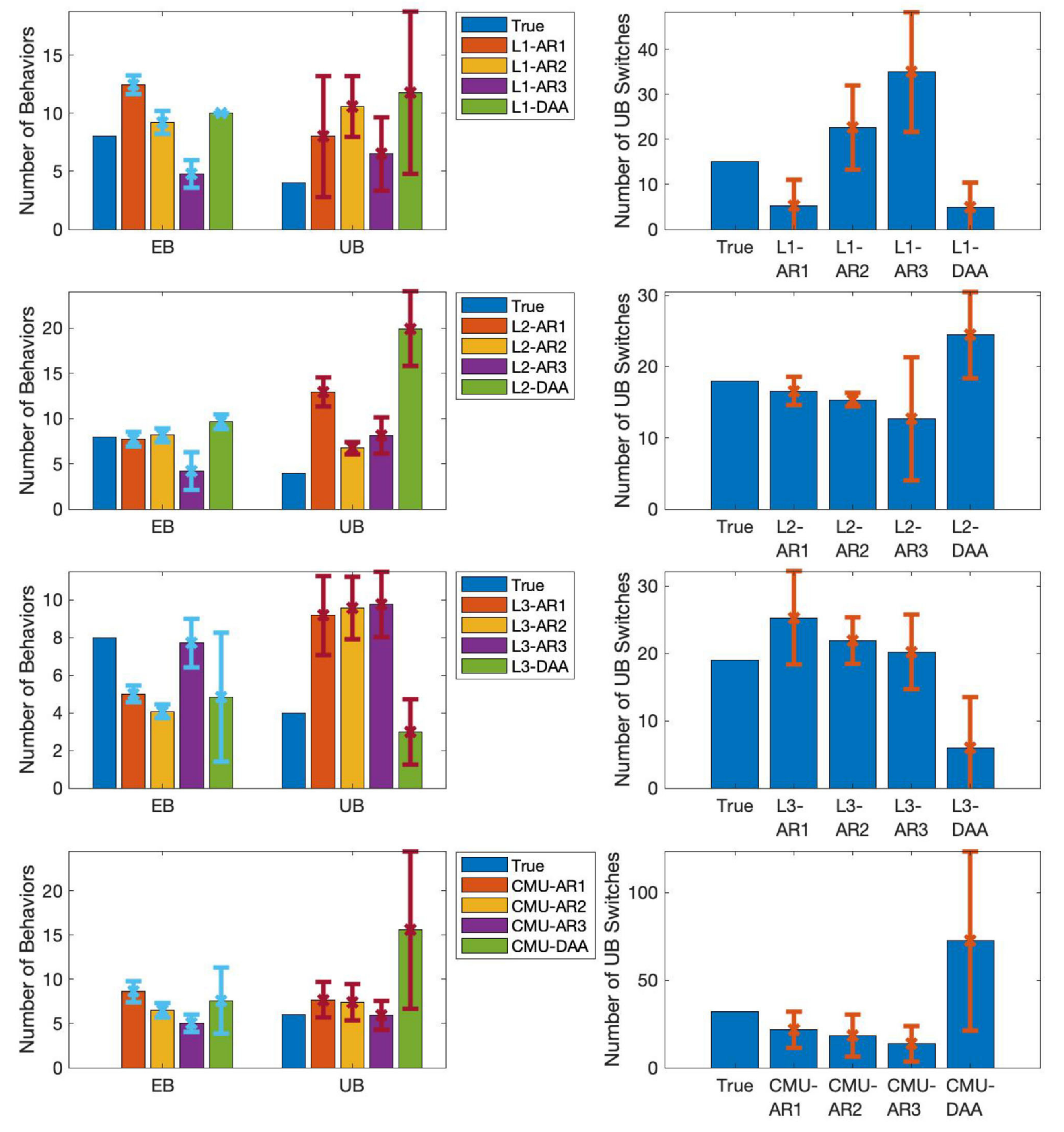


FIGURE 7 | Switching characteristics of estimated labels. (Left) Average number of discovered EBs and UBs for each experiment. Bars: 1 SD. (Right) Average number of switches in all time series sequences for each experiment. Bars: 1 SD.

results were observed in the cases of Sets **L2** and **L3**. Note that this tendency was also observed in their respective adjusted Rand index (ARI) (right figures in **Figure 5**) and joint log probabilities of data and sample variables (**Figure 6**). The joint log probability $P(\mathbf{y}, \mathbf{F}, \mathbf{z})$ is available even without the ground truth. Thus, the joint log probability can be a potential criterion for selecting the model with cross-validation.

Finally, when we compared the normalized HDist of our method (selected using the joint log probabilities) with that of DAA, our method generally had better results than DAA. **L2-AR2**

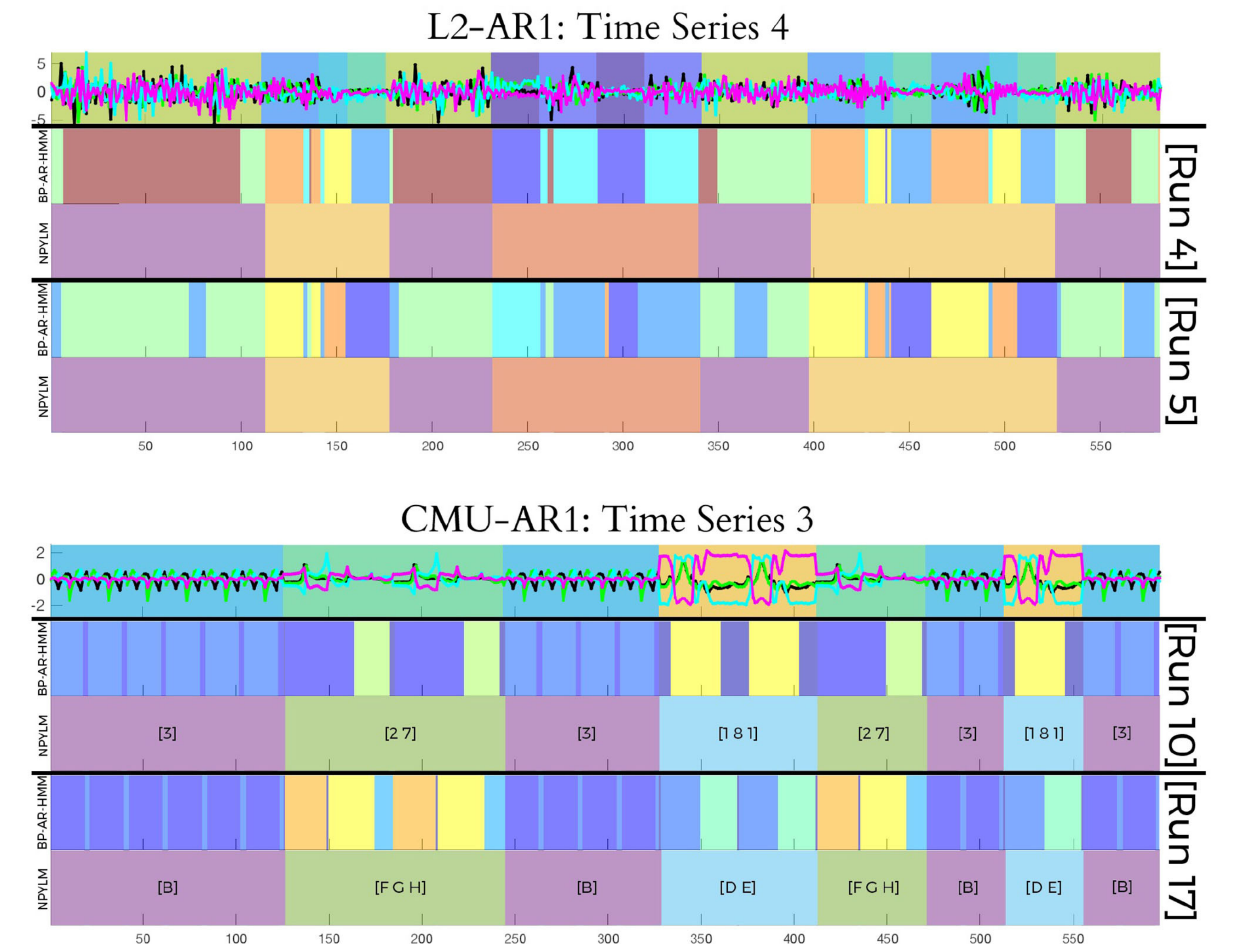


FIGURE 8 | Plot of the first four dimensions of the input data and example segmentation results using our proposed method, obtained from two different runs. (Top) Toy data, with result from **L2-AR1**. (Bottom) CMU motion data, with result from **CMU-AR1**.

and **L3-AR3**, which have the highest joint log probability among used models, have smaller EB HDist and UB HDist than **L2-DAA** and **L3-DAA**, respectively. In the case of **L1**, **L1-AR1** shows smaller EB HDist than **L1-DAA**, but larger UB HDist. Even in this case, **L1-AR2** shows better performance than **L1-DAA**. In summary, these results indicate the superiority of our method to DAA (**Figure 5**).

Aside from having better results compared to DAA, our method has other advantages. First, we note that the second step of our method reduces the error observed (left figures in **Figure 5**). Except for the results of **L1-AR1**, UB HDist are generally smaller than EB HDist. Even when segmentation in the EB level is wrong, correct segmentation in UB level is still possible, provided that the wrong pattern extraction of EB is reproduced for segments of same UBs. Second, it was also seen that the number of discovered EBs, including the EBs and the resulting segmentation, for each run varied (**Figure 8**). Despite this, the segmentation results at the UB step were more or less similar, as seen in the computed UB HDist. This observation indicates our method can identify the same UBs despite discovering different EBs. In other words, our method can absorb the difference of estimated EBs among different runs.

3.2.1. Large-Scale Toy Data

Results from subdataset **T***s* ($s = 10, 20, 100$) suggest that using our proposed method on larger datasets would yield more discovered EB labels (13.90, 17.70, 21.33) and formed UB labels (21.00, 19.70, 39.33). This then causes multiple discovered labels to correspond to the same 'true' label. Adjusting for this when computing for EB and UB HDist, the computed EB HDist are 0.6112, 0.6759, and 0.7086 for **T10**, **T20**, and **T100**, respectively, while the corresponding UB HDist are 0.1096, 0.1633, and 0.1864, respectively. These results are consistent with our earlier observation that the second step of our method reduces observed errors in the first step. That is, correct segmentation in UB level is possible despite having errors in EB level segmentation, and regardless of the number of time series sequences considered.

4. MOTION DATA EXPERIMENTS

Aside from synthetic experiments, motion data was also used in segmentation experiments. This is to see the applicability of proposed method to segment actual motion sequences.

4.1. Real Motion Data

To determine the effectiveness of our method with real motion data, one dataset was generated using the motion capture sequences of the actions of Subjects 18–23 in CMU Graphics Lab—Motion Capture Library (CMU, 2009). The dataset has four time series sequences of 16 dimensions that correspond to 8 joint angles of 2 individuals. The time series $\left\{\mathbf{y}_t^{(i)}\right\}$ $(i = 1, 2, 3, 4)$ were generated by concatenating UBs randomly chosen from six fixed types. The six UBs were the following actions: (1) walk toward each other than shake hands, (2) linked arms while walking, (3) synchronized walking, (4) alternating squats, (5) alternating jumping jacks, and (6) synchronized jumping jacks.

To evaluate the applicability of our method to real motion data, our method with AR orders $(r = 1, 2, 3)$ was applied to the dataset. The parameters in our method were set in the same way as the previous section 3.1, but with $\kappa = 200$. Similar to toy data, our method was also compared with DAA using CMU dataset. State sequences were processed similar to the toy data, but with no states being discarded, as the states switched frequently in the first step of DAA.

4.2. Real Motion Data Applicability

Similar to the experiments in the previous section, the result of our method with AR(r) applied to the CMU dataset was denoted by **CMU-AR*r***, while **CMU-DAA** was used to denote results from DAA. In terms of the average normalized UB HDist, **CMU-AR1** had the smallest error when compared with **CMU-AR2** and **CMU-AR3** (**Figure 5**). However, **CMU-AR2** and **CMU-AR3** had higher log probability than **CMU-AR1** (**Figure 5**). The optimal AR order could not be determined from the joint log probabilities in this case. Another criterion is needed to choose an optimal AR order. There is no existing available criterion, because our method is a highly complex singular model.

Comparing with the results of DAA, our method again had better performance than DAA. **CMU-AR1** had smaller UB HDist (0.1815) compared to **CMU-DAA** (0.2080) (**Figure 5**). Similarly, **CMU-AR1** had higher UB ARI (0.6847) compared to **CMU-DAA** (0.6384) (**Figure 5**). Unlike the obtained results from our proposed method, DAA had more UBs and switches, due to oversegmentation (**Figure 7**).

Finally, similar to the toy dataset results, the number of discovered EBs and the EB labels still varied for each run. However, the segmentation results of the UB step were quite similar (for example, see **Figure 8**). Here, the UB [1 8 1] refers to the alternating jumping jacks motion. In another segmentation, the same UB label corresponds to [D E], with the component EBs referring to completely different behaviors. Despite the difference in component EBs, both [1 8 1] and [D E] refer to the same true UB. Our method can thus identify the same semantic behaviors even from real motion data.

5. DISCUSSION

To discover complex temporal patterns from the time series data via segmentation, we proposed a hierarchical non-parametric Bayesian approach. We combined the BP-AR-HMM and the double articulation by NPYLM to segment time series sequences under the assumption that they are generated from hierarchically-structured dynamical systems. In our results, we found that our method has better accuracy to discover temporal patterns than DAA for both the toy and real motion datasets. It may mean the necessity of dynamics to model local temporal patterns in time series data. Also, double articulation structure of our method would be suitable to extract semantic unit behaviors from unsegmented human motion sequences similar to DAA. In addition, our proposed method has another advantageous property over DAA. Our proposed method allows for asynchronous switching of segments, unlike DAA. It should be beneficial to extract temporal patterns from natural observation without any intervention, since we cannot expect consistent switching of behaviors under the natural observation. Despite these benefits, it should be noted that our proposed method is limited by its computational complexity. Furthermore, should the assumption of the sequence having a hierarchical structure not be met, our proposed method could not necessarily be appropriate to use.

Future directions are as follows: (1) using the estimated AR coefficients for interaction analysis and causality analysis, (2) a semi-supervised extension of the proposed method, and (3) automatic determination of AR order. Some methods of the causality analysis, e.g., Granger causality, are based on the AR models in their mathematical formulations. Therefore, we can use the estimated AR coefficient matrix to connect our method to causality analysis. With this combined approach, it will be possible to analyze switching causality. Next, it is usually difficult to have categorical labels for the entire dataset, but partial labels are easier to have. In this case, using semi-supervised segmentation could help improve the interpretability of results since some of the discovered components or states would correspond to the known categories. These labeled instances may also improve the identification of distribution of corresponding categories. A semi-supervised extension of our approach would thus be more effective to discover behavioral patterns. Finally, although we tried multiple settings of the AR order to select a model, automatic determination of AR order will solve this model selection problem.

We then conclude that our method can extract temporal patterns from multiple time series sequences by segmenting these sequences into low-level and high-level segments. Our method showed superior performance to a method called double articulation analyzer. Moreover, even when it discovered different low-level segments, our method can absorb such variation, and properly and consistently identify high-level segments.

DATA AVAILABILITY STATEMENT

The datasets analyzed for this study can be found in the CMU Graphics Lab-Motion Capture Library (CMU, 2009).

AUTHOR CONTRIBUTIONS

JB, TK, and KI contributed to the design and implementation of the research, to the analysis of the results, and to the writing of the manuscript. All authors contributed to the article and approved the submitted version.

FUNDING

This work was partially supported by JSPS KAKENHI 17H05863, 25118009, 15K16395, 17H05979, 16H06569, and 18K18108.

REFERENCES

Alazrai, R., Mowafi, Y., and Lee, C. S. G. (2015). Anatomical-plane-based representation for human-human interactions analysis. *Pattern Recogn*. 48, 2346–2363. doi: 10.1016/j.patcog.2015.03.002

Barbič, J., Safonova, A., Pan, J.-Y., Faloutsos, C., Hodgins, J., and Pollard, N. (2004). "Segmenting motion capture data into distinct behaviors," in *Proceedings of Graphics Interface* (London, ON), 185–194.

Briones, J., Kubo, T., and Ikeda, K. (2018). "A segmentation approach for interaction analysis using non-parametric Bayesian methods," in *Proceedings of the 62nd Annual Conference of the Institute of Systems, Control and Information Engineers (ISCIE)* (Kyoto).

Bruno, B., Mastrogiovanni, F., Sgorbissa, A., Vernazza, T., and Zaccaria, R. (2012). "Human motion modelling and recognition: a computational approach," in *2012 IEEE International Conference on Automation Science and Engineering (CASE)* (Seoul), 156–161. doi: 10.1109/CoASE.2012.6386410

CMU (2009). *Carnegie Mellon University Graphics Lab–Motion Capture Library*. Available online at: http://mocap.cs.cmu.edu/

Fine, S., Singer, Y., and Tishby, N. (1998). The hierarchical hidden markov model: analysis and applications. *Mach Learn*. 32, 41–62. doi: 10.1023/A:1007469218079

Fox, E., Sudderth, E., Jordan, M., andWillsky, A. (2009). "Sharing features among dynamical systems with beta processes," in *Advances in Neural Information Processing Systems*, Vol. 22, eds Y. Bengio, D. Schuurmans, J. D. Lafferty, C. K. I. Williams, and A. Culotta (Vancouver, BC: Curran Associates, Inc), 549–557.

Fox, E., Sudderth, E., Jordan, M., and Willsky, A. (2008a). "An HDP-HMM for systems with state persistence," in *Proceedings of the 25th International Conference on Machine Learning* (Helsinki), 312–319. doi: 10.1145/1390156.1390196

Fox, E., Sudderth, E., Jordan, M., and Willsky, A. (2008b). "Nonparametric bayesian learning of switching linear dynamical systems," in *Advances in Neural Information Processing Systems*, Vol. 21, eds D. Koller, D. Schuurmans, Y. Bengio, and L. Bottou (Vancouver, BC: Curran Associates, Inc), 457–464.

Fox, E., Sudderth, E., Jordan, M., and Willsky, A. (2014). Joint modeling of multiple time series via the beta process with application to motion capture segmentation. *Ann. Appl. Stat*. 8, 1281–1313. doi: 10.1214/14-AOAS742

Gilson, M., Tauste Campo, A., Thiele, A., and Deco, G. (2017). Nonparametric test for connectivity detection in multivariate autoregressive networks and application to multiunit activity data. *Netw. Neurosci*. 1, 357–380. doi: 10.1162/NETN_a_00019

Grigore, E. C., and Scassellati, B. (2017). "Discovering action primitive granularity from human motion for human-robot collaboration," in *Proceedings of Robotics: Science and Systems* (Cambridge, MA).

Harrison, L., Penny, W., and Friston, K. (2003). Multivariate autoregressive modeling of fMRI time series. *Neuroimage* 19, 1477–1491. doi: 10.1016/S1053-8119(03)00160-5

Heller, K., Teh, Y.W., and Gorur, D. (2009). "Infinite hierarchical hidden Markov models," in *Artificial Intelligence and Statistics*, eds D. van Dyk and M. Welling (Clearwater, FL: PMLR), 224–231.

Hughes, M. (2016). *NPBayesHMM: Nonparametric Bayesian HMM Toolbox, for Matlab*. Available online at: https://github.com/michaelchughes/NPBayesHMM

Hughes, M., Fox, E., and Sudderth, E. (2012). "Effective split-merge Monte Carlo methods for nonparametric models of sequential data," in *Advances in Neural Information Processing Systems*, Vol. 25, eds F. Pereira, C. J. C. Burges, L. Bottou, and K. Q. Weinberger (Lake Tahoe, CA: Curran Associates, Inc), 1–9.

Keogh, E., Chu, S., Hart, D., and Pazzani, M. (2004). "Segmenting time series: a survey and novel approach," in *Data Mining in Time Series Databases*, eds M. Last, A. Kandel, and H. Bunke (World Scientific), 1–21. doi: 10.1142/9789812565402_0001

Mochihashi, D., Yamada, T., and Ueda, N. (2009). "Bayesian unsupervised word segmentation with nested Pitman-Yor language modeling," in *Proceedings of the Joint Conference of the 47th Annual Meeting of the ACL and the 4th International Joint Conference on Natural Language Processing of the AFNLP: Volume 1–ACL-IJCNLP '09* (Singapore), Vol. 1, 100–108. doi: 10.3115/1687878.1687894

Neubig, G. (2016). *latticelm*. Available online at: https://github.com/neubig/latticelm

Neubig, G., Mimura, M., Mori, S., and Kawahara, T. (2010). "Learning a language model from continuous speech," in *11th Annual Conference of the International Speech Communication Association (InterSpeech 2010)* (Makuhari), 1053–1056.

Ryoo, M., and Aggarwal, J. (2009). Semantic representation and recognition of continued and recursive human activities. *Int. J. Comput. Vis*. 82, 1–24. doi: 10.1007/s11263-008-0181-1

Saeedi, A., Hoffman, M., Johnson, M., and Adams, R. (2016). "The segmented iHMM: a simple, efficient hierarchical infinite HMM," in *International Conference on Machine Learning* (New York, NY), 2682–2691.

Taniguchi, T., and Nagasaka, S. (2011). "Double articulation analyzer for unsegmented human motion using Pitman-Yor language model and infinite hidden Markov model," in *2011 IEEE/SICE International Symposium on System Integration, SII 2011* (Kyoto), 250–255. doi: 10.1109/SII.2011.6147455

Taniguchi, T., Nagasaka, S., and Nakashima, R. (2016). Nonparametric Bayesian double articulation analyzer for direct language acquisition from continuous speech signals. *IEEE Trans. Cogn. Dev. Syst*. 8, 171–185. doi: 10.1109/TCDS.2016.2550591

Viviani, P., and Cenzato, M. (1985). Segmentation and coupling in complex movements. *J. Exp. Psychol. Hum. Percept. Perform*. 11:828. doi: 10.1037/0096-1523.11.6.828

Williams, B., Toussaint, M., and Storkey, A. (2007). "Modelling motion primitives and their timing in biologically executed movements," in *Advances in Neural Information Processing Systems*, Vol. 20, eds J. C. Platt, D. Koller, Y. Singer, and S. T. Roweis (Vancouver, BC: Curran Associates, Inc), 1609–1616.

Zeger, S., Irizarry, R., and Peng, R. (2006). On time series analysis of public health and biomedical data. *Annu. Rev. Public Health* 27, 57–79. doi: 10.1146/annurev.publhealth.26.021304.144517

Zhou, F., Torre, F. D. L., and Hodgins, J. K. (2013). Hierarchical aligned cluster analysis for temporal clustering of human motion. *IEEE Trans. Pattern Anal. Mach. Intell*. 35, 582–596. doi: 10.1109/TPAMI.2012.137

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 Neuroinformatics

TECHNOLOGY AND CODE
published: 16 November 2020
doi: 10.3389/fninf.2020.594799

MP3: Medical Software for Processing Multi-Parametric Images Pipelines

Clément Brossard[1,2], *Olivier Montigon*[1], *Fabien Boux*[1,3], *Aurélien Delphin*[1], *Thomas Christen*[1], *Emmanuel L. Barbier*[1] *and Benjamin Lemasson*[1,2]*

[1] *University of Grenoble Alpes, INSERM, U1216, Grenoble Institut des Neurosciences (GIN), Grenoble, France,* [2] *MoGlimaging Network, HTE Program of the French Cancer Plan, Toulouse, France,* [3] *University of Grenoble Alpes, Inria, CNRS, G-INP, Grenoble, France*

This article presents an open source software able to convert, display, and process medical images. It differentiates itself from the existing software by its ability to design complex processing pipelines and to wisely execute them on a large databases. An MP3 pipeline can contain unlimited homemade or ready-made processes and can be carried out with a parallel execution system. As a viewer, MP3 allows display of up to four images together and to draw Regions Of Interest (ROI). Two applications showing the strengths of the software are presented as examples: a preclinical study involving Magnetic Resonance Imaging (MRI) data and a clinical one involving Computed Tomography (CT) images. MP3 is downloadable at https://github.com/nifm-gin/MP3.

Keywords: software, medical images, image processing, pipelines, database, MRI, CT

OPEN ACCESS

Edited by:
Koichi Fujiwara,
Nagoya University, Japan

Reviewed by:
Michael McElwain,
National Aeronautics and Space Administration, United States
Edwin Valentijn,
University of Groningen, Netherlands

***Correspondence:**
Benjamin Lemasson
benjamin.lemasson@univ-grenoble-alpes.fr

Received: *14 August 2020*
Accepted: *14 October 2020*
Published: *16 November 2020*

Citation:
Brossard C, Montigon O, Boux F, Delphin A, Christen T, Barbier EL and Lemasson B (2020) MP3: Medical Software for Processing Multi-Parametric Images Pipelines. Front. Neuroinform. 14:594799. doi: 10.3389/fninf.2020.594799

1. INTRODUCTION

Researchers in medical imaging now have access to large amounts of data via open source databases [The Cancer Imaging Archive (TCIA; Dataset, 2020b), Center Traumatic Brain Injury (CTBI; Dataset, 2020a)]. In parallel of the quantity of images available, the complexity of the medical images post-processing is increasing. Where data scientists applied a single process to their images to obtain valuable results (Lemasson et al., 2016), state of the art analysis requires the execution of very complex interdependent processes called *pipelines* involving the execution of many operations, referred to as *modules*, on large databases (Funck et al., 2018). Among the modules used (e.g., bias removal, brain extraction, or image registration) some of them are part of toolboxes well-recognized by the community [such as SPM (Software, 2020) or FSL (Jenkinson et al., 2012) for the neuroimaging community] whereas others are home-made. How to reconcile flexibility, adaptive, speed, performance, and reproducibility of a post-processing? Several software have been recently developed to solve this issue. One can divide them in two classes: specific and generic software. Specific one, are built to process a specific type of data, for instance a modality [e.g., Positron Emission Tomography (PET), Magnetic Resonance Imaging (MRI),...], an organ imaged (e.g., liver, brain,...), or even a format [e.g., Digital imaging and communications in medicine (DICOM), Neuroimaging Informatics Technology Initiative (NIfTI),...] using a predefined pipeline. On the other hand, generic software, aimed to apply different processes on several types of data. Most of the available software can be gathered in the first category. This includes BrainCAT (Marques et al., 2013) which generates diffusion tensor images (DTI) from MRI diffusion weighted images, MRtrix (Tournier et al., 2019), which allows visualization and specific processes on diffusion MRI data, Pydpiper (Friedel et al., 2014) which offers registration algorithms, or FuNP (Park et al., 2019) which gathers some of the most widely used functions available for MRI data in a single post-processing pipeline. In those software, either the type of input data or the processing applied is unalterable. The second category contains software that can apply many different processes on a large type of data. To our knowledge, only few software available in free access belong to this category. One can quote Vaa3D (Peng et al., 2010), able to handle large dataset of images converted in Tagged Image File Format (TIFF) in a few seconds, to display them in 3D, and to define and execute homemade pipelines. One can also refer to GraphMIC (Zehner et al., 2015), which provides a node based interface for famous image processing libraries and allows to link the modules of these libraries to create complex image processing pipelines without programming, simply by connecting existing modules. This powerful software makes the design of complex pipeline user-friendly as well as the integration of home-made modules. Those powerful software lack of functionalities that we believe will become essential to develop new imaging biomarkers in order to handle and process large heterogeneous databases. Instead of processing one file or subject at a time, one will need a tool able to handle large cohorts, containing heterogeneous (multi-format) and multi-parametric data, and to process them with specific complex pipelines and with respect to the intra-subject time dependencies. To answer these limitations, we propose an open source software called "*Medical software for Processing multi-Parametric images Pipelines (MP3)*" (Software, 2019). This software therefore aims to facilitate the design of complex image analysis pipelines using existing or home-made modules as well as their execution on large heterogeneous databases.

2. METHOD

2.1. Overview

MP3, a MATLAB (Software, 2017) toolbox, intends to assist an end-to-end research study, from the loading and the converting of raw images to the statistical analysis through the creation of a database containing metadata and the design and execution of complex analysis pipelines. MP3 is composed of three linked graphical user interfaces (GUI) that stands as its backbone: the converter, the viewer, and the pipeline manager (**Figure 1A**). Briefly, those GUI enable the conversion, display, and processing of different medical image formats and architectures (Bruker, DICOM, PAR/REC, NIfTI, BIDS; Gorgolewski et al., 2016). The imported data is summarized in a database able to be homogenized, filtered, or improved with metadata such as the name of the patient or the day of the acquisition. The viewer can display up to four 5D images simultaneously and to draw ROI. Eventually, a graphical interface called Pipeline Manager creates, manages, and executes complex pipelines using editable modules. A pipeline execution system named PSOM (v2.2.2) (Bellec et al., 2012) allows to judiciously parallelize the execution of the modules. In this section, we detail the aims and functionalities specific to each of these GUI and present the concept of an MP3 project and its architecture. Detailed information as well as videos presenting latest developments of MP3 are available online (Github: https://github.com/nifm-gin/MP3, YouTube: https://www.youtube.com/playlist?list=PL-Tj6Wc9aE9x7i6s-RLetvNE0isnEsFm7).

2.2. Data Import and Conversion

MP3 is built on the NIfTI format (Cox et al., 2004), which contains the imaged volumes as a matrix (up to seven dimensions), and a short header that stores information about the image as its type or its size. This header also contains a transformation matrix that describes the position of the volume in a conventional space which is very useful to co-register two NIfTI volumes without modifying the image. In order to store metadata of each image acquisition that may be essential for some post-processing (echo time, repetition time, etc.), a consortium led by France Life Imaging defined conventional metadata and a way to write them in a JSON file (Kain et al., 2020). A pair of NIfTI/JSON files entirely describes a medical image. To allow the use of classical formats of medical images, we developed a converter able to transform Bruker, PAR/REC, DICOM, and NIfTI files to a NIfTI plus JSON format. It is also possible to import data organized in the Brain Imaging Data Structure (BIDS) convention. Each user can also customize the metadata to be collected from the raw data (Bruker, DICOM, etc.) that will be stored in the JSON file by editing the provided YAML file. Full details about the customization of the image conversion is available online (https://populse.github.io/mri_conv/Home/index.html). The converter interface (**Figure 1B**) can be launched from the main MP3 window. To start a new project, one defines a folder to store the project. Then, thanks to the different parts of the converter (data browser, parameters, images list, etc...), one can easily navigate through complex and often human unreadable medical image architecture. When the desired images are found, they will be automatically converted to NIfTI/JSON and metadata such as timepoints and subjects names may be modified. Then, back to the viewer GUI, a database has been created and will be used to manage the project data (Video 1: https://www.youtube.com/watch?v=ebofxMSquFs&list=PL-Tj6Wc9aE9x7i6s-RLetvNE0isnEsFm7).

2.3. Project Architecture

On the hard drive, the main directory of each project, sorted in multiple sub-folders, contains all the data needed to open the project via MP3. We defined several types of files, among which *Scans*, classical medical images, described as a NIfTI and a JSON file, and *ROI*, described as a NIfTI file (containing a binary matrix). Each imported scan file is stored in the "Raw_data" sub-folder, while the processed files, written after a pipeline execution, are saved in the "Derived_data" sub-folder and the ROI in "ROI_data." Other sub-folders can be part of the project, such as for instance "Tmp," where lie temporary files, or "Saved_Pipelines," that stores the .mat files that summarize the designed pipelines (see section 2.5). To manage these sub-folders, files, and associated metadata, MP3 relies on a database containing, for each entry, information referred to as *tags*, as its Type-Tag (Scan or ROI), Subject-Tag, Timepoint-Tag, or Group-Tag associated, as well as its Path-Tag or Filename-Tag. Since MP3 is developed on MATLAB, we decided not

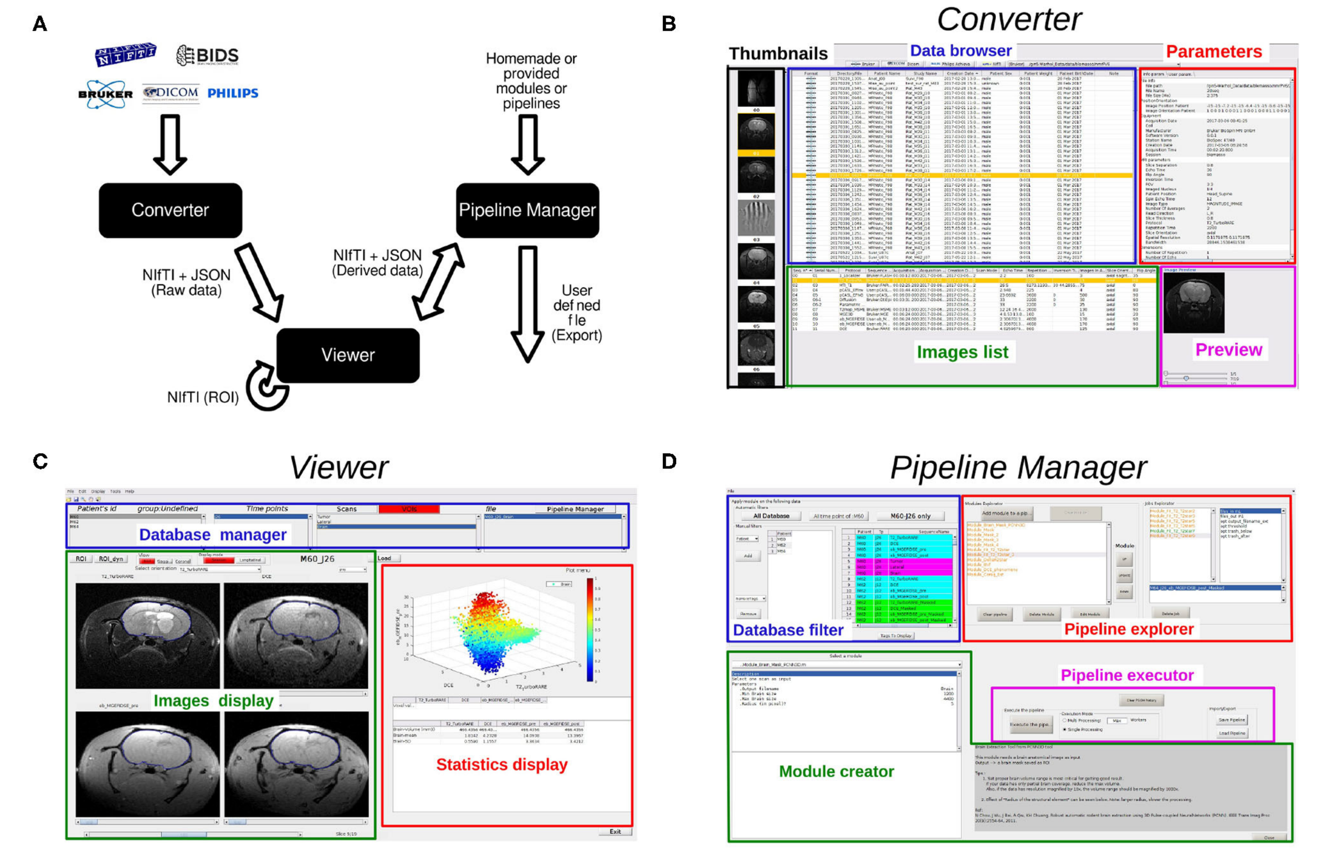


FIGURE 1 | (A): Architecture overview and data flow, **(B)**: Converter, **(C)**: Viewer, **(D)**: Pipeline Manager.

to use a real database, as we could have developed in SQL language, but to use a MATLAB *table* object. The power of these variables relies on their ability to be quickly and easily filtered, which is a key operation in MP3. Each entry is linked to the corresponding files through this table. It is therefore this database that undergoes operations as renaming, sorting, or filtering to select part of our available data or to homogenize the database to make it more consistent. Since a project is completely described by its database, MP3 offers the possibility of easily transferring a whole project to another user. To save space and to easily share any project, MP3 handles the compressed NIfTI format *.nii.gz*.

2.4. Viewer

The Viewer is a MATLAB GUI that is divided into three parts (**Figure 1C**): a *database manager*, to display and manage the database, the *image display*, that displays up to four 5D images from the database and supports the drawing of ROIs, and the *statistics display*, which displays scatter plots, curves, histograms, or first order statistics, like mean or standard deviation within a ROI (Video 2: https://www.youtube.com/watch?v=X26RV7VmXTA&list=PL-Tj6Wc9aE9x7i6s-RLetvNE0isnEsFm7).

2.4.1. Database Manager

The database (section 2.3) is presented as four listboxes each displaying a tag (Subject-Tag, Timepoint-Tag, Sequence_Name-Tag, and Filename-Tag) and working as filters that reduce the database. Thereby, when the Subject-Tag list displays all the different subjects of the database, selecting one of them filters the database and reduces the files displayed in the other lists. This way of presenting the database is inspired by the BIDS architecture and is a simple and efficient way to quickly access a specific scan. One can switch from the Scans list to the ROIs list by just a click on the specific button above the Sequence_Name-Tag list. These lists allow renaming the value of a tag, copying, or deleting an entry in the database thanks to specific right-click menus associated with each listbox. To compute these operations on a large number of entries, a sub-menu of the Edit menu offers some "Delete from database" or "Rename from database" features.

2.4.2. Image Display

One can load and display up to four 5D scans simultaneously. One of the main advantages of the NIfTI files is their transformation matrix stored in the header (section 2.2). This matrix locates the center of the scanner and its geometry. It is therefore possible, thanks to a light interpolation provided by functions of the SPM toolbox (Software, 2020), to display different orientations and different resolutions. This ability is used to open the selected scans in a selected referential. One can change the selected referential thanks to a popup menu displayed on the top of the figures. Open different scans in the same referential is particularly useful when displaying two scans acquired with a different field of view or a different voxel size. Three push buttons above the images can be used to change the referential orientation among axial, sagittal, or coronal. On a displayed image, one can use some classic features such as zoom and pan the image or vary its contrast to see hidden patterns (right-click on the mouse). One can also set the same contrast to all displayed images, and then quantitatively compare several images values (middle-click on the mouse). The viewer also offers a graphical tool able to create a contour that defines a ROI. This contour can be manually or automatically drawn thanks to an algorithm based on an active contour growth (Wang et al., 2009). ROIs are automatically stored as NIfTI files and can then be displayed on any other scan, regardless of its orientation or geometry.

2.4.3. Statistics Display

This part of the viewer is aimed at displaying quantitative information about the scans or ROIs loaded. It incorporates a sub-window to display graphics and a table for numeric values. The first metric is the values of the voxels. When the mouse pointer flies over an image, the value of the current voxel on each loaded scan is displayed on the table, which gives a quick overview of the quantitative values of each image. Another MP3 feature is its ability to study the temporal evolution of a four dimensional scan. Indeed, a click on a 4D voxel displays its values along the fourth dimension, which is particularly useful to study MRI perfusion or functional MRI. When an ROI is loaded, mean and standard deviation within the ROI for each loaded scan are displayed in a table, in addition to its volume in mm^3. When only one scan is loaded, the figure above that table hosts a histogram and when several scans are displayed, this histogram is replaced by a scatter plot, which represents the pixel values of one scan vs. the others. Since the MP3 project is open-source, one can easily display any simple feature or statistics by editing some of the main functions.

2.4.4. Longitudinal Follow-Up View Mode

Another way to analyze images from the database is to switch the display mode of the viewer from *session* to *longitudinal*. This mode allows comparison of scans taken at different times for a single patient. One can then easily check a co-registration, and conduct volumetric analyses.

2.5. Pipeline Manager

Since the basic features (section 2.4.3) are not enough to process complex analyses and are limited to the analysis of one subject we developed a third graphical user interface called the *Pipeline Manager* (**Figure 1D**). The pipeline manager allows the creation, editing, saving, and execution of complex pipelines on any file of a project. Moreover, it integrates fast and reproducible computation processes (mutli CPU) and history handling. This section exposes the philosophy of the pipeline manager and the main functionalities of this GUI (Video 3: https://www.youtube.com/watch?v=QOULzRsrzzg&list=PL-Tj6Wc9aE9x7i6s-RLetvNE0isnEsFm7).

2.5.1. Principal of the Pipeline Manager

Thanks to our structured project—data conversion (section 2.2) and database (section 2.3), the pipeline manager is prepared to manipulate our data. The pipeline manager allows iterations of a given process over the whole database (or a sub-part of the database). For instance, we are able to create a process that computes an image *B* from one specific image *A* and to apply it on each image *A* of the database. This process is called *pipeline*, because it can be a simple computation or a complex sequence of independent steps.

2.5.2. Pipeline Definition

As illustrated on **Figure 2A**, a pipeline is composed of one or several *modules*. A module is a function, more or less complex, that is designed to compute one or more NIfTI files thanks to one or more other NIfTI files, and some parameters. For instance, a module called *Module_Smooth* takes as input a Sequence_Name-Tag and 3 parameters (the dimension of the filter—2D or 3D Gaussian, the size of the filter—the variance of the Gaussian, and the Extension string of the output files), and generates a smoothed image. A module shall contain a basic operation, although nothing prevent a module from containing a whole process. A module applied to a database creates a certain number of *jobs*, i.e., a certain number of occurrences of that module. Therefore, since a pipeline is simply a sequence of modules linked together, applying a pipeline to a database creates a certain number of jobs for each module. This is potentially a large number of jobs.

2.5.3. Procedure to Create a Pipeline

Now that the concept of pipeline is introduced, we share the procedure to create one (**Figure 2B**). First, one needs to select the part of the data on which to execute the pipeline. Indeed, MP3 allows the user to apply a pipeline to the entire database or only on a sub-part of it (i.e., the 10 first subjects, or on a specific timepoint). This operation is managed by the "database filter" on the upper left corner (**Figure 1D**). This part of the GUI contains manual or predefined filters as well as a table that shows the filtered database. After the data selection, one can go to the module creator (**Figure 1D**) and choose a module among all the ones currently available (section 2.6.1). Selecting a module displays all its required parameters. For instance, selecting the *Module_Smooth* shows a line aimed to select a Sequence_Name-Tag value, another aimed to set a filter size, etc. After the modification of all parameters (which can also be left to their default value), one can add the module to the pipeline by clicking on the button *add module to a pipeline*. When one adds a module to a pipeline, the module automatically creates as many jobs as necessary (depending on the database). Each output Scan or ROI of each job is then added to the filtered database displayed on the upper left. Since the pipeline has not been executed yet, all those files do not exist yet but are already accessible for the user in the displayed database. This offers the possibility to parameterize another module with the tags of those *virtual entries*, and therefore link some modules together. The operation of filtering the database, selecting a module, parameterizing it, and adding it to the pipeline, which creates jobs (**Figure 2B**), can be reproduced an unlimited number of times. The *Pipeline explorer*, on the upper right corner (**Figure 1D**), allows an explanation of any part of the pipeline, from the different modules to each related job, to each job parameter in the job. Any operation (e.g., editing a module, deleting a module or job...) is entirely under the control of the user.

2.5.4. Saving and Sharing

In order to share a pipeline between several users, computers or projects, MP3 offers the option to save and load pipelines. Since a pipeline is a list of modules, we just store in a .mat file the sequence of modules with their parameters, but we do not store the jobs, since they are specific to a database. Loading a pipeline consists of applying each module to the new database. If the tag values of the pipeline to load are not consistent with the database, a module cannot generate the jobs. The module name is then displayed in red in the pipeline explorer and one just needs to edit the affected module, select the adapted tag value and save it to make it compliant, which turns the entry from red to green. This color convention intends to make clear the content of a pipeline as well as the consequences of its execution. If some modules are displayed in orange, it means that some of their jobs (in orange too) have already been executed (during a previous execution of the pipeline). In that case, one can decide to overwrite these files or to delete the related jobs when executing the pipeline.

2.5.5. Execution of a Pipeline

Once a pipeline is well-designed, there are two execution ways in MP3. The first way is the *Single Processing*. Each job is executed in the module order: all the jobs of the first module, then all the jobs of the second, etc. This quick execution, without strong dependencies between the modules and a low computing time, is especially useful when testing or developing a module. The second way is to use a pipeline execution system called "PSOM" (v2.2.2) (Bellec et al., 2012) by selecting the option *Multi Processing*. This powerful system enables the execution of a pipeline on several cores. The dependencies between jobs are taken into account and the jobs are then distributed upon each core. This system also offers a garbage collector in order to save Random Access Memory and a way to monitor the number of jobs launched in parallel. As it has an influence on the computer's behavior during the execution of the pipeline, one can set this number of *Workers* before the execution. During the execution, all output files are written in a temporary folder. At the end, each output file of a successful job is saved in its data folder according to its type (Derived_data or ROI_data). In order to help navigating in the database (cf. *Database filter* section), we defined a color code based on each scan type. The blue represents raw data, derived data are in green, while the pink means ROI data, and the yellow colorizes the virtual files (output files of not executed pipeline yet).

2.6. Modules

The power of the pipeline manager lies in its modules. They are the basic operations needed to design complex pipelines. We adapted the way to define modules exposed in PSOM (v2.2.2) (Bellec et al., 2012). A module is then a MATLAB function, stored in a .m file whose name begins by the string "*Module_*." All those functions lie in a folder called *Modules* of the MP3 source code and are sorted by area, with all the modules of an area in a folder. On launch, the pipeline manager reads the *Module* folder and its sub folders to list every available module. Thereby, the addition or deletion of a file in the subfolders of this repository updates the module list when launching the pipeline manager.

2.6.1. Provided Modules

MP3 is provided along with 12 modules, performing basic operations, such as smooth, threshold, or mask an image with a ROI, compute a brain extraction module (Chou et al., 2011), export a .csv file containing first order statistics or an HTML report of images and ROI. We also provide modules used in the preclinical pipeline exposed section 3.1, which are MRI oriented, and a module able to delete files of the project's database. Finally, a module interfacing the famous toolbox SPM (Software, 2020) allows reslicing images to match the referential of a reference image.

2.6.2. Template

Other users must be able to create their own modules. To facilitate the development of a new module, a module named *Module_Template* is available. It is extensively commented and explains how a module works. One then has to complete and adapt it to the wanted behavior and to save the new .m file in the *Module* folder to rapidly apply a new module on a database. To create a new module, even if it is simply an interface with another language, one need to know the basics of MATLAB programming.

2.6.3. Traceability

At each job execution, the module parameters and all the input/output filenames are stored in the JSON file associated with each output scan. Thereby, a scan obtained through the pipeline manager is linked to the raw data through the history of each module applied to the raw scan, and each module parameters. A basic GUI, called "File History," launchable from the viewer when a scan is loaded, is able to display all of this history. For each scan, we can go back to the past and know exactly how this scan was obtained and when each module has been executed.

2.6.4. Module Repository

The homemade modules are aimed to gather the laboratory knowledge and know how. It is also a way to avoid loss of skill due to the departure of a team member. For example, the development team wrote more than 60 modules that match our needs of image processing. All those modules are available at https://github.com/nifm-gin/MP3_User_Modules_Repository. One can find MRI oriented modules, interfaces to others toolboxes functions, such as the co-registration modules from Software (2020) or Jenkinson et al. (2012), or more advanced

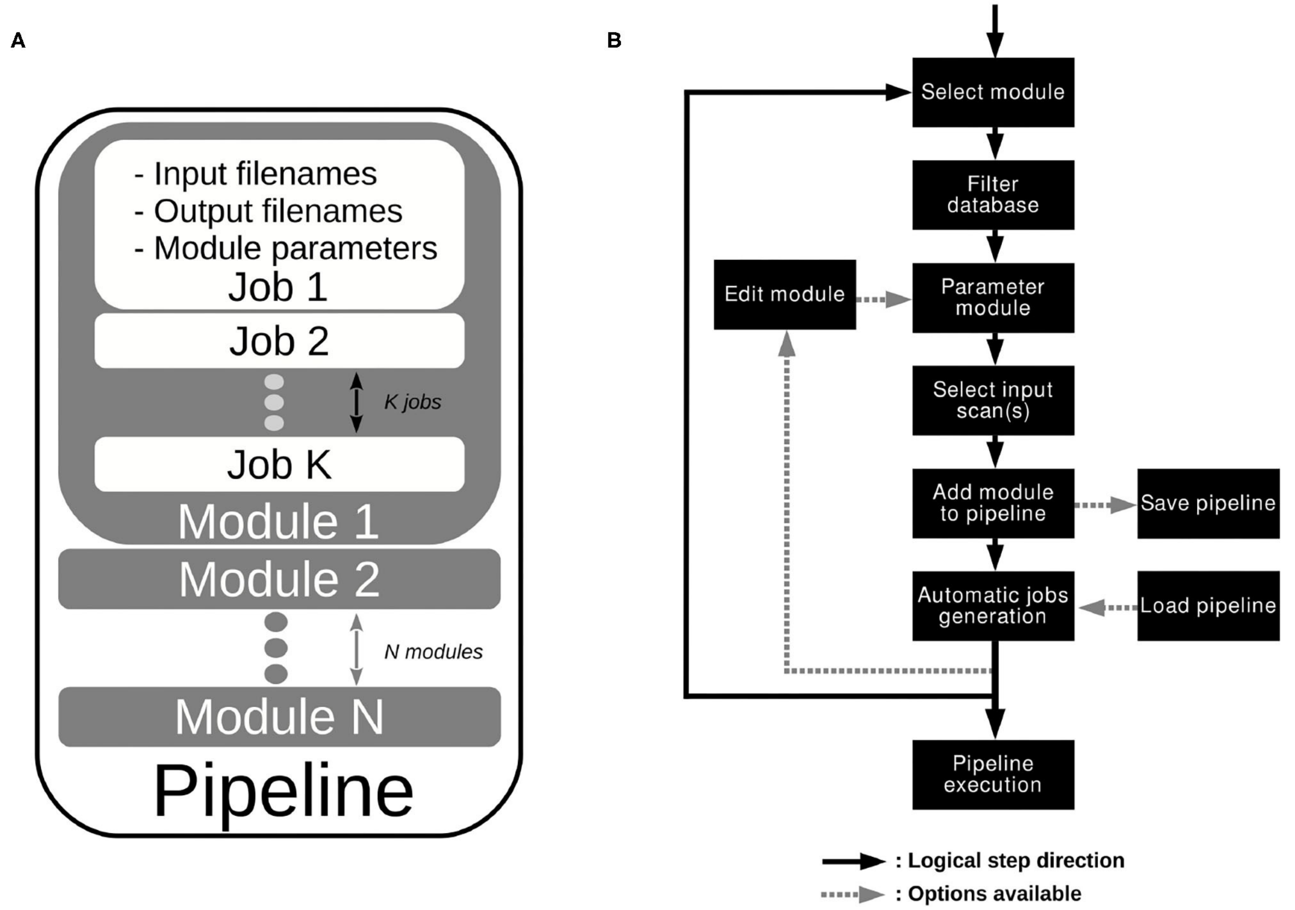


FIGURE 2 | (A) Pipeline architecture. **(B)** Work flow editor.

processing, such as clustering, texture analysis, MR fingerprinting analysis, quantitative MR relaxometry, MR perfusion analysis (cerebral blood flow, cerebral blood volume, vessels size, vessels permeability, etc.).

3. RESULTS

MP3 has been verified and validated on several MRI and CT data studies, but nothing prevents its use on any file that can be converted in NIfTI.

3.1. Preclinical Application

A research study published in 2017 compared the blood-barrier permeability changes induced by synchrotron microbeam or uniform radiation therapy. Eighteen rats bearing intracranial tumors were treated and imaged by multi-modal MRI using the Grenoble MRI facility IRMaGE (Bouchet et al., 2017). All procedures related to animal care conformed to the Guidelines of the French Government with licenses 380325 and 380321 (authorized lab A3818510002 and A3851610004). Part of these data are available online as an example of a MP3 project (Dataset, 2019). This project contains data of two rats among the 18 ones each imaged at three timepoints. Each timepoint contains 10 scans corresponding to the data described in Bouchet et al. (2017), a ROI delineating the brain, and some masked scans obtained thanks to the former ROI. The data processing, composed of 16 modules, was designed as an MP3 pipeline and applied on the database of the study. As shown on **Figure 3**, it created several occurrences of the pipeline aimed to process the data of each timepoint of each subject in the same way.

3.2. Clinical Application

As previously said, any image able to be converted in Bruker, DICOM, PAR/REC, or NIfTI format can be processed in MP3, no matter its nature. For instance, MP3 is used in a non-published study aimed to predict the evolution of brain injuries following a traumatic brain injury (TBI) using CT-scans acquired at the admission in the hospital, and at respectively 1 and 3 days after. A pre-processing pipeline for CT-scans, defined and described by Muschelli et al. (2015) was integrated as part of our own post-processing pipeline. This process resizes, clips, filters and extracts the brain from CT-scans. For our study, we added a module of coregistration (between the timepoints of a patient) from the SPM toolbox (Software, 2020) and a

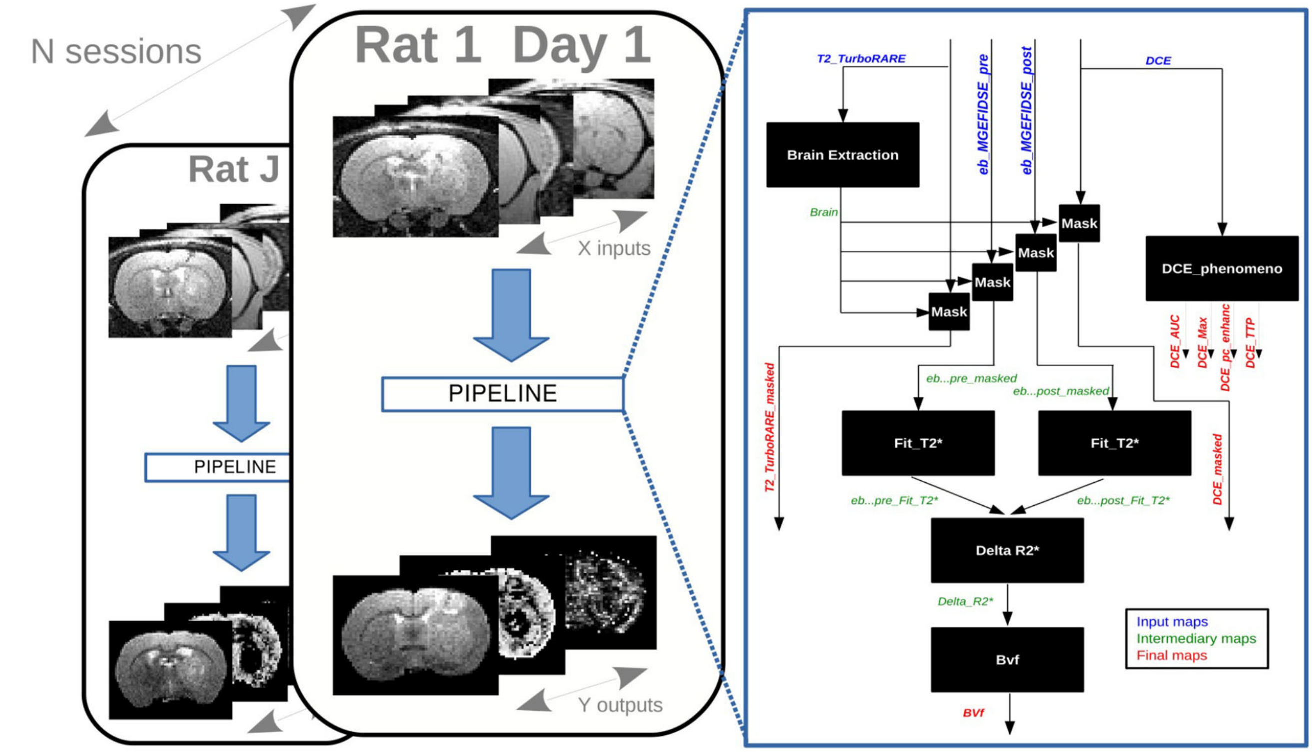


FIGURE 3 | Preclinical study: occurrences of a pipeline on a database.

homemade module able to process a local entropy map, as well as some other texture images. All those modules are available at https://github.com/nifm-gin/MP3_User_Modules_Repository.

The pipeline applied to a patient and its three timepoints is described on **Figure 4A**. This patient is part of a cohort whose data acquisition was allowed by the French institution "Comité de protection des personnes" and respects the patients written inform consent obligation. The CT-scans and entropy maps, respectively inputs and outputs of the pipeline, are displayed on **Figure 4B**. One can see that entropy scans are co-registered and skulls were well-removed. Such pipeline may be easily applied to a large sample of patients in order to follow the evolution of the brain entropy, used as a potential prognostic biomarker of neurological outcome.

4. LIMITATIONS

Despite all the work done to make it accessible and useful to the needs of the greatest number of people, MP3 does not yet offer all the performance that can be found in the best software in research fields such as physics or astronomy, which are more advanced on these aspects of interoperability and data reuse. MP3 does not meet all the FAIR (Findable, Accessible, Interoperable, Reusable) principles, which aim at improving the sharing of digital resources as described by Wilkinson et al. (2016). Indeed, MP3 is not based on a real database that is easily adaptable, nor on a standardized data architecture. Since standards definition in the medical imaging field was only at its beginning when we started the development of MP3 we did not integrate them deeply in the software. However, we decided to make MP3 compatible with recently defined standards, such as the BIDS convention (through our import/export functions). Another limitation concerns the execution of jobs on an external cluster that is not available today. Although this is conceptually possible, it would require additional development and testing. However, a multi-core execution is easy to use from a local server or computer. For example, we have tested MP3 on a project containing 450 subjects, 1,100 sessions for a total of 10,000 scans (3D, 4D, or even 5D). In this project, we successfully executed (on 32 CPU in parallel), onto the entire database a complex processing pipeline of 18 jobs per session which represented around 20,000 jobs.

5. CONCLUSION

This article exposed a new open source software able to support an end-to-end research study on a large amount of data. Thanks to three graphical interfaces, MP3 offers to convert and visualize medical images and to interact with them by comparing or analyzing them. Based on a pipeline execution system called PSOM, MP3 also enables the creation, management, and execution of complex pipelines on heterogeneous cohorts described in databases, which handle time dependencies and multiparametric data. MP3 can be used either by end-used (non-developers) or by developers which can improve it or develop

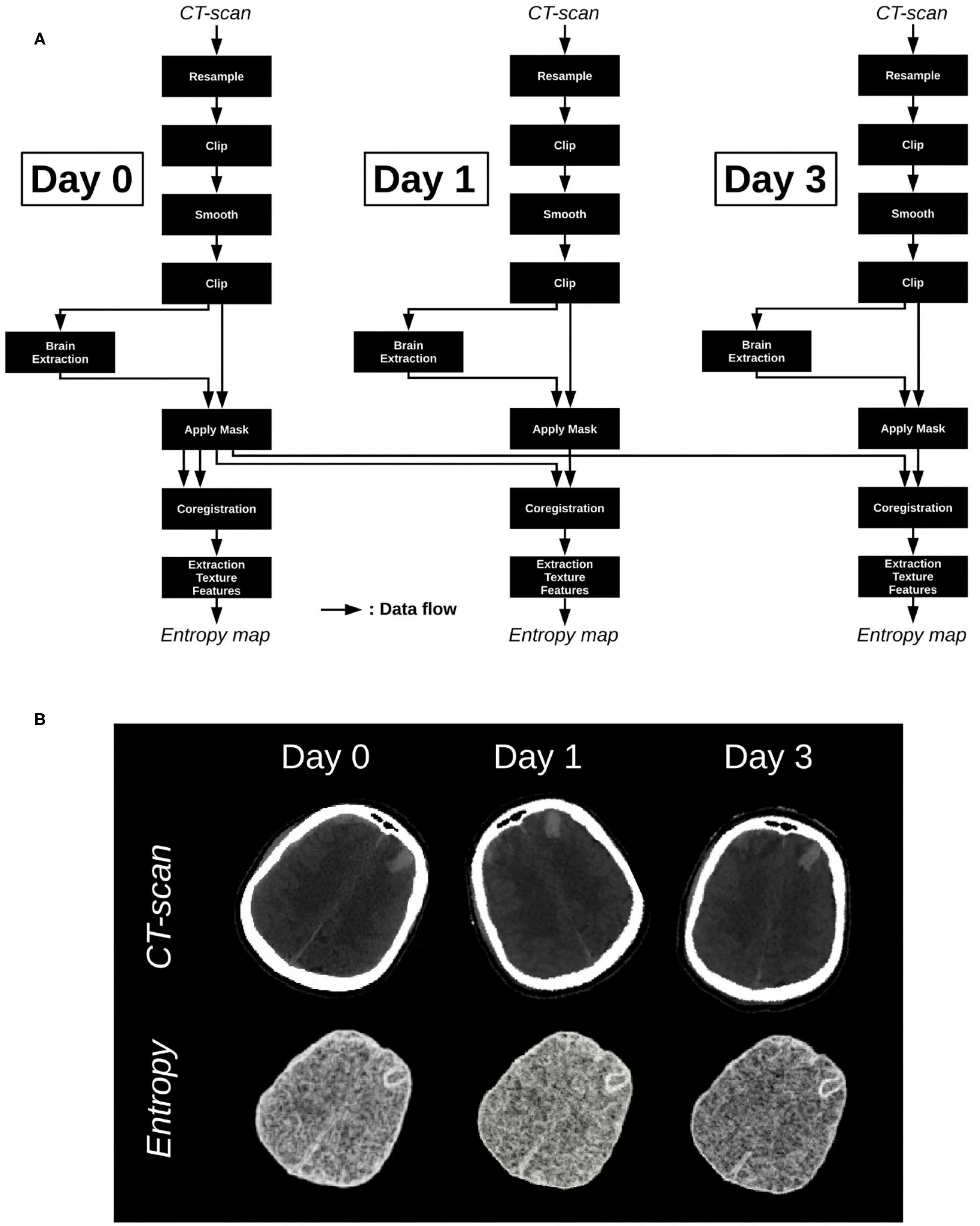


FIGURE 4 | (A) Processing pipeline applied on the images acquired at day 0, day 1, and day 3. **(B)** Input maps: CT-scans, and output maps: Entropy maps.

their own modules. MP3 has been tested on several studies, from preclinical to clinical, from MRI to CT data, and allowed us to process a preclinical cohort of more than 450 animals. MP3 can be downloaded on github: https://github.com/nifm-gin/MP3 and modules developed by our lab (more than 60) are available on: https://github.com/nifm-gin/MP3_User_Modules_Repository.

To conclude, we believe that, despite its limitations, MP3 makes it possible to facilitate large cohort analysis (preclinical/clinical) while improving the robustness and reproducibility of medical imaging studies. Indeed, one can think that in a near future every scientific publication will have to make both the raw data but also the processing pipeline used available and software such as MP3 will be able to facilitate this step.

6. REQUIREMENTS

MP3 can be run on MATLAB 2017b and higher and needs the Image Processing Toolbox. In order to fully enjoy the software, we recommend installing the Statistics and Machine Learning Toolbox and the Parallel Computing Toolbox. Since the converter GUI is developed in Java, Java 8 is also required. Any of the three main Operating System (OS) can handle this software. MP3 is open source, open development and available on github (Software, 2019). To develop new modules, one needs to know MATLAB programming.

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/nifm-gin/MP3; https://github.com/nifm-gin/MP3_User_Modules_Repository.

ETHICS STATEMENT

The study involving human participants were reviewed and approved by the French institution Comité de protection des personnes. The patients/participants provided their written informed consent to participate in this study. The animal study was reviewed and approved by the Guidelines of the French Government with licenses 380325 and 380321 (authorized lab A3818510002 and A3851610004). Written informed consent was obtained from the individual(s) for the publication of any potentially identifiable images or data included in this article.

AUTHOR CONTRIBUTIONS

All authors listed have made a substantial, direct and intellectual contribution to the work, approved it for publication, proofread, and corrected the final manuscript. CB and BL co-developed the main software. OM developed the converter. CB, FB, AD, and BL co-developed modules for MP3. CB and BL wrote the first draft of the paper.

FUNDING

This work has been supported by (i) the MoGlimaging network funded by the Heterogeneity of Tumors and Ecosystem program within the framework of the french Cancer Plan and (ii) the Fondation des Gueules Cassées (Paris, France), and the hospital CHUGA (Grenoble, France) through the project Radiomic-TBI.

ACKNOWLEDGMENTS

Grenoble MRI facility IRMaGe was partly funded by the French program Investissement d'Avenir run by the Agence Nationale pour la Recherche; grant Infrastructure d'avenir en Biologie Santé—ANR-11-INBS-0006.

REFERENCES

Bellec, P., Lavoie-Courchesne, S., Dickinson, P., Lerch, J., Zijdenbos, A., and Evans, A. C. (2012). The pipeline system for Octave and Matlab (PSOM): a lightweight scripting framework and execution engine for scientific workflows. *Front. Neuroinform.* 6:7. doi: 10.3389/fninf.2012.00007

Bouchet, A., Potez, M., Coquery, N., Rome, C., Lemasson, B., Br?uer-Krisch, E., et al. (2017). Permeability of brain tumor vessels induced by uniform or spatially microfractionated synchrotron radiation therapies. *Int. J. Radiat. Oncol.* 98, 1174–1182. doi: 10.1016/j.ijrobp.2017.03.025

Chou, N., Wu, J., Bai Bingren, J., Qiu, A., and Chuang, K.-H. (2011). Robust automatic rodent brain extraction using 3-D pulse-coupled neural networks (PCNN). *IEEE Trans. Image Process* 20, 2554–2564. doi: 10.1109/TIP.2011.2126587

Cox, R., Ashburner, J., Breman, H., Fissell, K., Haselgrove, C., Holmes, C., et al. (2004). "A (Sort of) new image data format standard: NIfTI-1, Vol. 22," in *10th Annual Meeting of the Organization for Human Brain Mapping* (Budapest).

Dataset (2019). *nifm-gin/Data_test_mp3. original-date: 2019-06-28T14:46:06Z.*

Dataset (2020a). Home//CENTER-TBI.eu

Dataset (2020b). *Welcome to The Cancer Imaging Archive.*

Friedel, M., van Eede, M. C., Pipitone, J., Chakravarty, M. M., and Lerch, J. P. (2014). Pydpiper: a flexible toolkit for constructing novel registration pipelines. *Front. Neuroinform.* 8:67. doi: 10.3389/fninf.2014.00067

Funck, T., Larcher, K., Toussaint, P.-J., Evans, A. C., and Thiel, A. (2018). APPIAN: automated pipeline for PET image analysis. *Front. Neuroinform.* 12:64. doi: 10.3389/fninf.2018.00064

Gorgolewski, K. J., Auer, T., Calhoun, V. D., Craddock, R. C., Das, S., Duff, E. P., et al. (2016). The brain imaging data structure, a format for organizing and describing outputs of neuroimaging experiments. *Sci. Data* 3, 1–9. doi: 10.1038/sdata.2016.44

Jenkinson, M., Beckmann, C. F., Behrens, T. E. J., Woolrich, M. W., and Smith, S. M. (2012). FSL. *NeuroImage* 62, 782–790. doi: 10.1016/j.neuroimage.2011.09.015

Kain, M., Bodin, M., Loury, S., Chi, Y., Louis, J., Simon, M., et al. (2020). SHANOIR-sharing neuroimaging resources. *Front. Neuroinform.* 14:20. doi: 10.3389/fninf.2020.00020

Lemasson, B., Wang, H., Galb?n, S., Li, Y., Zhu, Y., Heist, K. A., et al. (2016). Evaluation of concurrent radiation, temozolomide and abt-888 treatment followed by maintenance therapy with temozolomide and ABT-888 in a genetically engineered glioblastoma mouse model. *Neoplasia* 18, 82–89. doi: 10.1016/j.neo.2015.11.014

Marques, P. C. G., Soares, J. M., Alves, V., and Sousa, N. (2013). BrainCAT - a tool for automated and combined functional Magnetic Resonance Imaging and Diffusion Tensor Imaging brain connectivity analysis. *Front. Hum. Neurosci.* 7:794. doi: 10.3389/fnhum.2013.00794

Muschelli, J., Ullman, N. L., Mould, W. A., Vespa, P., Hanley, D. F., and Crainiceanu, C. M. (2015). Validated automatic brain extraction of head CT images. *NeuroImage* 114, 379–385. doi: 10.1016/j.neuroimage.2015.03.074

Park, B.-y., Byeon, K., and Park, H. (2019). FuNP (Fusion of Neuroimaging Preprocessing) pipelines: a fully automated preprocessing software for functional magnetic resonance imaging. *Front. Neuroinform.* 13:5 doi: 10.3389/fninf.2019.00005

Peng, H., Ruan, Z., Long, F., Simpson, J. H., and Myers, E. W. (2010). V3D enables real-time 3D visualization and quantitative analysis of large-scale biological image data sets. *Nat. Biotechnol.* 28, 348–353. doi: 10.1038/nbt.1612

Software (2017). *MATLAB version 9.3.0.713579 (R2017b)*. Natick, MA: The Mathworks, Inc.

Software (2019). *nifm-gin/MP3*. Available online at: https://github.com/nifm-gin/MP3

Software (2020). *SPM -Statistical Parametric Mapping*. Available online at: https://www.fil.ion.ucl.ac.uk/spm-statistical-parametric-mapping/

Tournier, J.-D., Smith, R., Raffelt, D., Tabbara, R., Dhollander, T., Pietsch, M., et al. (2019). MRtrix3: A fast, flexible and open software framework for medical image processing and visualisation. *NeuroImage* 202:116137. doi: 10.1016/j.neuroimage.2019.116137

Wang, L., He, L., Mishra, A., and Li, C. (2009). Active contours driven by local Gaussian distribution fitting energy. *Signal Process*. 89, 2435–2447. doi: 10.1016/j.sigpro.2009.03.014

Wilkinson, M. D., Dumontier, M., Aalbersberg, I. J., Appleton, G., Axton, M., Baak, A., et al. (2016). The FAIR Guiding Principles for scientific data management and stewardship. *Sci. Data* 3:160018. doi: 10.1038/sdata.2016.18

Zehner, A., Szalo, A. E., and Palm, C. (2015). "GraphMIC: easy prototyping of medical image," in *Computing Applications, Interactive Medical Image Computing (IMIC) Workshop, MICCAI 2015* (Munich).

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.

ORIGINAL RESEARCH
published: 09 February 2021
doi: 10.3389/fdgth.2020.567656

Disclosing Pharmacogenetic Feedback of Caffeine via eHealth Channels, Assessment of the Methods and Effects to Behavior Change: A Pilot Study

Kerti Alev[1]*, *Andres Kütt*[2] *and Margus Viigimaa*[3]

[1] *Digital Health, Tallinn University of Technology (TalTech), Tallinn, Estonia,* [2] *Information Technology and Communication Technologies, Information Technology Department, Tallinn University of Technology (TalTech), Tallinn, Estonia,* [3] *North Estonia Medical Center, Tallinn University of Technology, Tallinn, Estonia*

Background: The integration of genetic testing into eHealth applications holds great promise for the personalization of disease prevention guidelines. However, relatively little is known about the impact of eHealth applications on an individual's behavior.

Aim: The aim of the pilot study was to investigate the effect of the personalized eHealth application approach to behavior change in a 1-month follow-up period on groups with previously known and unknown caffeine impacts.

Method: We created a direct-to-consumer approach that includes providing relevant information and personalized reminders and goals on the digital device regarding the caffeine intake for two groups of individuals: the intervention group (IG) with the genetic raw data available and the control group (CG) to test the impact of the same content (article about caffeine metabolism) on participants without the genetic test. Study participants were all Estonians ($n = 160$).

Results: The study suggests that eHealth applications work for short-term behavior change. Participants in the genetic IG tended to increase caffeine intake if they were informed about caffeine not being harmful. They reported feeling better physically and/or mentally after their behavioral change decision during the period of the study.

Conclusions: Our pilot study revealed that eHealth applications may have a positive effect for short-term behavior change, regardless of a prior genetic test. Further studies among larger study groups are required to achieve a better understanding about behavior change of individuals in the field of personalized medicine and eHealth interventions.

Keywords: mHealth, pharmacogenetics, behavior change, Caffeine, eHealth, decision support, digital health, mobile

OPEN ACCESS

Edited by:
Chester Drum,
National University Health System, Singapore

Reviewed by:
Aishwarya Bandla,
National University of Singapore, Singapore
Tyler John Loftus,
University of Florida, United States

****Correspondence:***
Kerti Alev
kerti@medikeep.eu

Specialty section:
This article was submitted to Personalized Medicine, a section of the journal Frontiers in Digital Health

Received: *30 May 2020*
Accepted: *15 October 2020*
Published: *09 February 2021*

Citation:
Alev K, Kütt A and Viigimaa M (2021) Disclosing Pharmacogenetic Feedback of Caffeine via eHealth Channels, Assessment of the Methods and Effects to Behavior Change: A Pilot Study. Front. Digit. Health 2:567656. doi: 10.3389/fdgth.2020.567656

INTRODUCTION

With the development of interactive devices and extensive connectivity to the web, the online channel has become popular to deliver tailored healthy lifestyle-promoting interventions. In general, online behavioral health change programs (eHealth channels) seem to work very well-according to those studies and meta-analyses (1–5). eHealth channels act as an important factor of self-efficacy and patient education, and are essential to the improvement of patient–clinician communication, along with improvement in trust, adherence, and social support (6).

The Mayo Clinic observed understanding and perspectives on pharmacogenomics for patients in 2017: how their patients understood the effect of their personalized medicine. It was discovered that 30% of the participants did not understand the results, and further work is needed to establish a better understanding of the results (7). It was found that a chunking strategy could help genetic feedback participants better obtain information that they did not relate to before, even better if the participants obtained it in an extended period (8). Also, performance feedback and self-monitoring techniques have given good results in behavior change programs (9). Focused goal setting, when it is combined with personal feedback messages, is also considered a promising approach (9).

Many studies noted that more frequent communication could improve behavioral health outcomes. Minimal contact will not help to boost smoking cessation (10). Orleans et al. suggested that more frequent follow-up phone calls increased the cessation by up to 50%. According to Hietaranta-Luoma et al. the positive effect of lifestyle change seemed to fade during the "silent" period of the trial. According to Brouwer et al., the most used behavior-changing techniques were feedback, interactive elements, and email/phone contact (add reference). Furthermore, the most effective techniques were listed as peer support, counselor support, email/phone contact, and updates of the intervention website. Schmidlen et al. suggest that genetic counselors offer health topics for the participants to choose from based on their interest and offer visual aids to explain the risk scores. Another systematic review revealed that without lifestyle counseling, the reaction to behavioral change can be modest (11). After online genetic counseling in the breast cancer study, the willingness to adapt to a healthier lifestyle after genetic testing results was high among women who participated in the study for breast cancer risk (12). A study trial plan has been made to define the steps in behavioral change after the direct-to-consumer genetic testing (DTCGT). The results could lead to a better understanding of behavior change and DTCGT in the future (13).

We studied individuals with and without previous genetic test and offered them a tool or method for behavior change in a commonly used medicine and nutrition ingredient—caffeine. Caffeine belongs to our everyday products, but some tend to drink it more than others. Caffeine consumption is partially genetically regulated by metabolism of the gene CYP1A2 [we used the single-nucleotide polymorphism (SNP) rs762551].

While the study participants were organized into a group with their genetic information [intervention group (IG) and a group without genetic test [control group (CG)], both groups were offered the same information about caffeine: how it is related to genetics, metabolism, and caffeine health risks. Both groups were offered digital tools (mobile app MediKeep[1] with a notification system and plain email-based reminders), set goals, and reminders for behavior change (online web form to choose preferences).

This leads to the topic of the current study. This study assesses if genetic test makes DTCGT participants change their behavior in the short term (1-month follow-up) when using eHealth services with strategic methods: information chunking, creating personalized reminders and goals on their digital device.

This study has the following sub-aims:

1. Whether people with genetic test results (when reading the article about caffeine genetic health traits and risks) tend to decide to change their behavior compared to the people reading the same article without a genetic test.
2. Does the genetic test result help people to decide to increase specific nutrition if they find it genetically suitable for them?
3. Does using goals and reminders help people using eHealth service to change their behavior in the short term (1-month follow-up)?

MATERIALS AND METHODS

This pilot study consists of the assessment of the channels and methods based on the previous studies (**Table 2**) and an empirical study among $n = 160$ participants in Estonia. The authors conducted a study, based on the previous findings (**Table 2**), to test the hypothesis among DTCGT participants (IG) and the same content impact on participants (CG) without any prior knowledge of their genetic metabolism for caffeine.

The study has been approved by the Research Ethics Committee of the University of Tartu (approval no.: 290/T-10 since 8.03.2019 until 30.06.2019).

Methods and Channels for eHealth and Genetic Feedback

With the development of interactive devices and extensive connectivity to the web, the online channel has become a popular way to deliver tailored healthy lifestyle-promoting interventions. In general, online behavior health change programs seem to work very well according to studies and meta-analyses (1–5). The importance of self-efficacy and patient education is essential to improve patient–clinician communication, along with improvement in trust, adherence, and social support (6).

However, a study among Food4Me participants found no effect in physical activity change between groups with genetic test and fat mass- and obesity-associated (FTO) genetic risk. The personalized feedback led to improved self-reported physical activity outcomes in general, suggesting that personalized advice rules out genetic risk score (1).

According to Brouwer et al. (4), the most used behavior-changing techniques were feedback, interactive elements, and email/phone contact. Furthermore, the most effective techniques

Abbreviations: AI, Artificial Intelligence; DNA, Deoxyribonucleic Acid; DTC, Direct to Consumer; DTCGT, Direct-to-Consumer Genetic Testing; FTO, Fat Mass and Obesity Associated; REVEAL, The Risk Evaluation and Education for Alzheimer's Disease; SNP, Single-Nucleotide Polymorphisms.

[1] MediKeep (https://www.medikeep.eu) is a mobile app for home pharmacy management, and during the study, the company offered a research platform based on active ingredients in caffeine. The functionality included creating an account and receiving reminders within the app and mobile and email notifications.

were listed as peer support, counselor support, email/phone contact, and updates of the intervention website.

In the REVEAL trials, the study group followed the people for a whole year after learning their genomic risks (14). What they learned, in general, was that giving out graphics or illustrating the risk scores on the timelines along with the genetic counseling in their practice was the best way of doing it. They were not claiming it to be the absolute best methodology. Artificial intelligence (AI) and smartphone-based motivation and action support system can improve and maintain the physical activity among adult populations while they got actively engaged and reminded about the program (15).

However, in a personalized medicine setting, problems were found in understanding the results. The Mayo Clinic observed understanding and perspectives on pharmacogenomics for patients in 2017: how their patients understood the effect of their personalized medicine. It was discovered from the results that one third did not understand the results, and further work is needed to establish a better understanding of results (7). A chunking strategy could help genetic feedback participants better obtain the information that they did not relate to before, even better if they obtained it over a more extended period (8).

Performance feedback and self-monitoring techniques have given good results in behavior change programs (9). Focused goal setting, when combined with personal feedback messages, is also considered a promising approach (9).

Many studies noted that more frequent communication could improve behavior health outcomes. Minimal contact will not help boost smoking cessation (10). Orleans et al. (16) suggested that more frequent follow-up phone calls increased the cessation by up to 50%. According to Hietaranta-Luoma et al. (17), the positive effect of lifestyle change seemed to fade during the “silent” period of the trial.

Schmidlen et al. (18) suggest that genetic counselors offer health topics for the participants to choose from based on their interest and offer visual aids to explain the risk scores. Another systematic review revealed that without lifestyle counseling, the reaction to behavior change can be modest (11). After online genetic counseling in the breast cancer study, the willingness to adapt to a healthier lifestyle after genetic testing results was high among women who participated in the study for breast cancer risk (12). A study trial plan has been made to define the steps in behavior change after the DTCGT. The results could lead to a better understanding of behavior change and DTCGT in the future (13).

Behavioral Change and eHealth

While there are guides for health services or tools for best practices like Cochrane reviews and NICE guidance, the behavior change factors were more effectively described in “The behavior change wheel: A new method for characterizing and designing behavior change interventions” by Michie et al. (19). Those functions could be easily translated as methods to the eHealth application format.

According to this “behavior change wheel,” the change starts with three components: capability, opportunity, and motivation—the COMB-B system (20).

- **Intervention** functions surround the base components. To improve change, the deficit among intervention functions should be decreased as suggested by the article.
- **Education** stands for increasing the knowledge or understanding.
- **Persuasion** means using communication.
- **Incentivization** is creating an expectation of a reward.
- **Coercion** indicates punishment.
- Other functions are listed as training, restriction, environmental restructuring, and enablement.

Edwards et al. (21) found that among mobile applications which were using gamification in mobile health, the most popular techniques were self-regulatory. Those included several proven health behavior change methods like goal setting, self-monitoring, and feedback (21).

Literature Overview: Assessment of the Channels and Methods Based on Previous Studies

In November 2018, the words “DNA test” and “behavior change” were searched on the Google.com search platform in Estonia by the authors of this study. An article came into interest, “Genetic testing does not change how most people behave, study finds”[2], leading to the original systematic review and meta-analysis in BMJ.com, published in 2016 (22). Rather than searching for individual studies for the current research, the BMJ meta-analysis was used to assess if other factors are contributing to DNA test and behavior change.

The trials were eligible for the BMJ 2016;352:i1102 study if they were randomized controlled trials or quasi-randomized controlled trials among adults and included one group that got personalized DNA risk estimates for risks where the behavior could change the health risk outcome. The analysis includes 18 clinical trials around the world: the United States (8), the United Kingdom (5), Japan (3), Finland (1), and Canada (1). Behaviors included smoking, alcohol consumption, diet, and physical activity.

Table 1 lists all included studies ($n = 19$), with close to 6,000 participants in total, while **Table 2** is a checklist and shows what kind of channels and methods were used in the study. The authors were not able to fill all the blocks because of missing or unclear information in the study, and such cells were marked with a question mark (?).

The major finding from **Table 2** indicates that while the BMJ 2016;352:i1102 study focused on the change between CG and IG and mostly no change was found, it did reveal that there was a positive change in most of the groups. It confirmed the assumption that while people’s education and knowledge were raised about the topic, positive behavioral change was also found after participating in the trial/program.

It was also suggested that some topics might be more motivational and easier to act on for the participants to take behavioral actions. For example, while willingness to increase

[2] https://www.theverge.com/2016/3/15/11241334/genetic-testing-disease-risk-dna-behavior-changes.

physical activity was higher among the FTO gene risk group, actual results in physical activity did not rise (32). However, taking supplements for Alzheimer's prevention is considered an easier task than regular gym visits or changes in diet (14).

Only Hietaranta-Luoma et al. (17) had three or more follow-ups. Clear "reminder" techniques were detected in four of the studies (25, 26, 33, 36), while it was uncertain in one (14). Physical or beneficial rewards were present in four of the studies (25, 29, 34, 36), while it was not available in one (33). Most of the trials increased knowledge of the participants, while it was not done in two (23, 30), resulting in no effect on health behavior in any of those groups. More than half of the trials offered printed materials, 11 offered face-to-face counseling or sessions with doctors, seven had communications via email at least once, six trials included a phone call, and only three had some kind of online interaction or communication. Increased motivation for behavioral change was detected in seven of the trials.

The studies had several limitations, from small study groups to not well-targeted participants. For example, college students (mean age 22) might not be interested in reducing body fat or they might not drink alcohol as much to reduce their drinking in general (32).

TABLE 1 | BMJ 2016;352:i1102 included clinical trials, their number, topic, and number of participants.

No.	Name/Year	Topic	No. of participants
1	Audrain et al. (10)	Smoking	426
2	Hishida et al. (23)	Smoking	562
3	Ito et al. (24)	Smoking	617
4	McBride et al. (25)	Smoking	557
5	Sanderson et al. (26)	Smoking	61
6	Chao et al. (14)	Medicine or supplements	162
7	Hendershot et al. (27)	Reduced alcohol	200
8	Hietaranta-Luoma et al. (17)	Reduced alcohol	107
9	Komiya et al. (28)	Reduced alcohol	329
10	Glanz et al. (29)	Sun protection	73
11	Chao et al. (14)	Diet	162
12	Godino et al. (30)	Diet (Diabetes II)	557
13	Hietaranta-Luoma et al. (17)	Diet	107
14	Marteau et al. (31)	Diet	341
15	Meisel et al. (32)	Diet	279
16	Nielsen et al. (33)	Diet	138
17	Voils et al. (34)	Diet	601
18	Weinberg et al. (35)	Screening and behavior programs	783
19	Grant et al. (36)	Screening and behavior programs	177

[Note that (17) and (14) are represented in two categories].

Empirical Study Design

Subjects

The subjects were all native Estonians, and all spoke the Estonian language. They were all adults (≥18 years) recruited online via social media and forum advertisements with targeted keywords: caffeine, genetic testing, digital technology, and personalized medicine. The previous genetic test was not a prerequisite to participate in the study. If the participant had a previous DTCGT raw data available, they were asked to provide the information.

Study Design

Invitation to the study was sent out via social media channels, local forums, and Facebook groups in Estonia (#Tervis, #uhkegeenidoonor, and #MediKeep). The questionnaires were published on the Tyeform.com platform, which allows fluid usability and logic jumps in online questionnaire forms. The questionnaires and study design graphics are provided as **Supplementary Material** for this article.

The questionnaires were conducted based on Nielsen et al.'s (37) initial research questions for "A randomized trial of genetic information for personalized nutrition," The questionnaire was slightly modified for the empirical study, including the option to set a goal and reminder (9) besides the third questionnaire for self-assessment of the behavior health results after 1-month follow-up.

TABLE 2 | BMJ 2016;352:i1102 included clinical trials in meta-analysis, their interactions, and their communication channels.

	Study no./Topics	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19
	Positive change in intervention group				x		x			x		x		x			x			
	Positive change in any group	x	-	x	?	x	x	x	x	x	x	x	?	x	-	-	x	-	?	x
Interactions	Three or more follow-ups	-	-	-	-	-	-	-	x	-	-	-	-	x	-	-	-	-	-	-
	Reminders	-	-	-	x	x	-	-	-	-	-	?	-	-	-	-	x	-	-	x
	Rewards	-	-	-	x	-	-	-	-	-	x	-	-	-	-	-	?	x	-	x
	Increased knowledge	x	-	x	x	x	x	x	x	-	x	x	-	x	?	x	x	x	x	x
	Printed materials	-	x	-	x	x	x	-	-	-	x	x	x	-	-	x	-	x	-	x
	Increased motivation	x	-	-	x	x	-	x	x	x	-	?	?	x	-	?	?	-	-	x
Channels	Face to face	x	-	x	-	x	x	-	x	x	x	?	-	x	x	-	-	x	x	x
	Email	-	-	x	-	-	-	x	x	-	-	?	-	x	-		x	-	x	-
	Telephone	x	-	-	x	x	-	-	-	-	x	?	-	-	x	-	-	-	x	-
	Webpage/online interaction	-	-	-	-	-	-	x	-	-	-	?	-	-	-	-	x	-	x	-

Participants were organized into groups based on their first questionnaire. Participants with previous genetic testing results and in whom CYP1A2 gene SNP rs762551 was available were organized into an IG "with genetic test results," while those "without a genetic test" conducted comprised the CG. They were also organized by the preference of the next communication channel: mobile app MediKeep, email, or both. If the subjects were not sure about having their DNA raw data, additional email was sent to clarify the answer. If the answer confirmed the genotype results, the participant was added to the IG. If the answer was negative and they did not know their caffeine metabolism from any of the testing companies, the participant was assigned to the CG. For file transfer, additional service was provided via the MASV[3] portal for fast file delivery solution, especially when handling large files. For security reasons, raw data files were signed with id.ee[4] services (encryption to be sent only for the author's national ID or digital signature).

After the successful first questionnaire, the groups received an email or mobile app message to read an article about genetics and caffeine metabolism in the Estonian language. The IG received their results with information about their genotype, while the CG was offered to read the article without their genetic profile.

The second questionnaire was linked just below the article. The second online form asked the participants if they understood the article, if it had any new information for them in regard to caffeine, and how they related to the results. Participants were asked if they would like to change anything in their health behaviors and if they would like to set a goal and reminders. Again, two channels were offered: by email or via a MediKeep mobile application push-notification system.

If no goals or reminders were set, participants were directed to the final question informing about the third questionnaire to be sent in the 1-month follow-up. Participants who opted-in for goals and reminders were also asked if they would like to increase or decrease their caffeine intake and if they had any specific favorites (coffee, green tea, cacao, chocolate, or anything else). They were informed that the reminders would be sent no more than five times a week for 1 month. The reminders were set manually for every participant based on their answer to get a more personalized meaning, while also trying to mimic AI. They were set to be sent out automatically in a 3-day interval but no more than 10 times per participant for 1 month.

The reminders were also set by time, in the morning, or at lunch, depending on personal preferences. The message texts were changed slightly once a week to prevent reminder exhaustion and spam, while the goal remained the same. Participants were reminded at least three times by email if they had not answered their second questionnaire in 3 weeks. The data management was done in Typeform.com[5] and imported to Google Spreadsheet in March–May 2019, and all duplicate responses were deleted after new response information was confirmed.

TABLE 3 | Questionnaire participation rates and devices in Typeform.com.

	Responses	Total visits	Unique visits	Completion rate%	Average time to complete
First questionnaire	213	1,364	1,111	19.3	10:33
PCs and laptops	121	689	549	22.20	10:40
Smartphones	83	635	526	15.80	10:51
Tablets	9	40	36	25	06:15
Second questionnaire	172	359	314	55.1	03:14
PCs and laptops	107	182	143	75.5	02:56
Smartphones	62	169	154	40.3	03:52
Tablets	3	8	7	24.9	03:06
Completion rate compared to the first questionnaire				80.75	
Third questionnaire	160	372	180	88.9	10:44
PCs and laptops	85	181	91	93.4	17:54
Smartphones	68	176	82	82.9	02:39
Tablets	7	15	7	100	02:32
Completion rate compared to the second questionnaire				93	
Completion rate compared to the first questionnaire				75	

The third questionnaire was sent out to every participant who had finished the second questionnaire with or without the goals/reminders set. The last questionnaire asked if they had decided to change something in their behavior and when. Questions also included decision making, whether genetic test results had any impact in deciding what to change, physical and mental well-being, whether reminders and goals had any impact (with or without genetic test), and whether they would have liked to get similar topics about other genetic traits. The questionnaire also had one open-ended question for additional comments.

The results were collected in the Typeform.com platform and data collected in Google Spreadsheet. The calculations were made using the Chi-Square Calculator for a 2 × 2 contingency table[6]; take note that every cell should have a number of 1 and 20% of the cells should not have data <5 (38).

RESULTS

Questionnaires

Compared to the first questionnaire, the actual completion rate was 80.75% for the second one (**Table 3**). The third questionnaire received a completion rate of 93% compared to the second, and from all the participants who started the questionnaire, 75% completed the third and final questionnaire.

The first questionnaire received 1,111 unique visits, resulting in 213 responses by the 1st of May 2019 (one duplication removed

[3] https://massive.io

[4] https://www.id.ee

[5] https://www.typeform.com

[6] https://www.socscistatistics.com/tests

TABLE 4 | Sample demographics according to the first questionnaire.

Age groups	Intervention group	Control group	*p*
18–29	8	38	0.43
30–39	23	62	0.11
40–49	10	48	0.34
50–59	4	14	0.94
60–69	1	4	
70 and more	0	1	
Education			
Primary	1	5	
Secondary	12	53	0.46
Bachelor's degree	18	73	0.57
Master's degree	13	35	0.22
Doctoral degree	2	1	
Gender			
Female, $n = 167$	29	138	0.0012
Male, $n = 42$	17	25	

from 214 total responses). The completion rate was 19.3%, and it took an average of 10:33 min to finish the questionnaire.

All participants were native Estonian speakers. Majority of the participants were female 79.9%, $n = 167$ and aged between 30 and 39 (39.9%, $n = 85$), and only one belonged to the 70–79 age group (**Table 4**). Males were 20.1%, $n = 42$. There was a demographic difference in the IG where males were more represented (*p*-value = 0.0012).

Most of the respondents had at least a bachelor degree (42.7%, $n = 91$). Only 19.7% of the participants had not heard anything about DTCGT company tests before, 57.3% ($n = 122$) had heard somewhat, and 23% ($n = 49$) had heard a lot. Only 5% ($n = 10$) had heard a lot about the term nutrigenomics, and 25.8% ($n = 55$) had not discovered the term or field of science previously.

Majority of the participants (63.5%, $n = 134$) were genuinely interested in the relations between diet and genetics, while one participant was definitely not (0.5%). Even more of them agreed that there is a benefit in learning about how genetic makeup affects diet (73.8%, $n = 157$), 22.5% ($n = 48$) somewhat agreed, 2.8% ($n = 6$) neither agreed nor disagreed, and 0.9% ($n = 2$) would rather not disagree.

Most of the participants (43.2%, $n = 92$) were somewhat assured that learning about genetic makeup will affect what they eat, 38% ($n = 81$) agreed very much, 16.4% ($n = 35$) neither agreed nor disagreed, and 2.3% ($n = 5$) disagreed.

Participants somewhat disagreed (54.7%, $n = 116$) that the results would make them uncomfortable and anxious to learn about the genetic findings, 22.2% ($n = 47$) neither agreed nor disagreed, 20.8% ($n = 44$) somewhat agreed, and 2.4% ($n = 5$) were sure they would.

Most of them (82.5%, $n = 175$) were sure a genetic test would make them learn more about themselves, 9.4% ($n = 20$) somewhat agreed, 5.2% ($n = 119$) neither agreed nor disagreed, and 2.8% ($n = 6$) would rather not agree. However, a little bit less were ready to change anything in their behavior to be healthier (66.7%, $n = 142$), 23% ($n = 49$) somewhat agreed, 7% ($n = 15$) neither agreed or disagreed, and 3.3% ($n = 7$) would rather disagree.

They agreed (66.7%, $n = 140$), somewhat agreed (23%, $n = 49$), neither agreed nor disagreed (7%, $n = 15$), and disagreed (4.2%, $n = 9$) to take the genetic test for the doctor to monitor their health more closely.

Participants preferred the next communication channel to be via email (80.6%, $n = 170$), while 13.3% ($n = 28$) preferred both email and mobile and 6.2% ($n = 13$) preferred mobile only.

Over half of the respondents (58.7%, $n = 125$) did not have their DNA raw data, 20.7% ($n = 44$) were not sure they had the data, and 20.7% ($n = 44$) confirmed they had the data. The high percentage of participants who were "not sure" was related to the next question asking for the testing company name, where a high percentage who answered the "Other" option were Estonian Biobank participants, and they were not sure if they had their results or DNA raw data. Those people were organized into the CG instead, and other obscure answers were corrected, with additional information requested *via* email. After correction, 21.5% ($n = 46$) in total had confirmed to have their DNA data, and everyone else was organized into the "no DTCGT group" 78.5% ($n = 167$). After correction of the testing company name, $n = 38$ of the people with raw genomic data were previous clients of MediKeep OÜ, and about $n = 8$ were willing to share their data from other companies (23andme.com, FTDNA, and Geni.com).

Those participants decided to send their files through email or just looked up the marker information and copied the required rows from the DNA raw data file as plain text via email. No one from the eight participants (outside from the MediKeep client base with their raw data) used the Massive.io portal or signed/encrypted their data as described in the study introduction and suggested in the email. An additional comment from one of the emails was "I hope this file comes through. I do not worry much about my data security—so there you are!"

It was impossible to assess whether those who did not respond to the email asking about their DNA data received the email in the first place, as this email was sent from a private email address with no tracking capabilities. All MediKeep clients had their information already within the company and agreed with the study consent to look at their genotypes—resulting in a high percentage of participation in the study after the second questionnaire until the third (100%).

The second questionnaire received 172 results (314 unique visits) with a completion rate of 55.1%, and it took an average of 03:14 min to finish, not including the time to read the article and for decision making (**Table 3**). Compared to the first questionnaire, the second one had a completion rate of 80.75%. It was impossible to measure how much time the participants took for decision making. However, the majority completed the questionnaire on the same day when they received it.

Generally, the participants (75%, $n = 129$) did not feel uneasy after reading the article, 16.9% ($n = 29$; $n = 2$ with genetic test

TABLE 5 | Sub-aim 1: differentiation not found between the IG and CG regarding the decision to change current caffeine consumption with the help of set goals and reminders ($p = 0.80$).

	Intervention	Control	Marginal row totals
Reminder	14 (13.7) [0.01]	48 (48.3) [0]	62
No reminder	24 (24.3) [0]	86 (85.7) [0]	110
Marginal column totals	38	134	172 (grand total)

Table information is provided as follows: observed cell totals (expected cell totals) [chi-square statistic for each cell].

TABLE 6 | Sub-aim 2: participants with DTCGT may increase the caffeine intake, based on the knowledge about their caffeine metabolism (the $p = 0.047$. This result is significant at $p < 0.05$); however, the sample size is quite limited.

	Intervention	Control	Marginal Row Totals
Increase caffeine	4 (1.81) [2.66]	4 (6.19) [0.78]	8
Decrease caffeine	10 (12.19) [0.39]	44 (41.81) [0.12]	54
Marginal column totals	14	48	62 (grand total)

The table information is provided as follows: observed cell totals (expected cell totals) [chi-square statistic for each cell].

results) did feel uneasy, and 8.1% ($n = 14$) did not know how to answer. Almost all (98.3%, $n = 169$) understood what the caffeine article was explaining, while 1.7% ($n = 3$) did not fully understand the content; they also did not have a previous genetic test. Over half of the respondents (69.2%, $n = 119$) found the article information to be something new, while 30.8% ($n = 53$) did not find it new to their knowledge. From $n = 172$ participants who completed the second survey, 36% ($n = 62$) opted in for goals and reminders, and 64% ($n = 110$) did not. From the CG, $n = 48$ opted in, and from the IG, $n = 14$ opted in (**Table 5**).

From the reminder group, 72.6% ($n = 44$) preferred the reminder channel to be email, and 27.4% ($n = 18$) chose the mobile application for the channel ($n = 8$ Android and $n = 10$ iOS operating platforms). Android users did not receive their messages within the first week because of technical problems, and their participation in setting reminders was postponed for 1 week.

While 71% ($n = 44$) wished to decrease their coffee consumption, nobody wished to decrease or increase cacao consumption (**Table 6**). However, the other option included reduction of cola products. The products to be increased were coffee ($n = 4$), (green) tea ($n = 3$), and chocolate ($n = 1$). For the one participant who wished to increase chocolate consumption, the authors of this study sent a personalized message indicating that the participant may increase chocolate intake as they wished; however, it was suggested to choose at least 70% cocoa/dark chocolate for better health behavior.

The third questionnaire received $n = 160$ results. From 180 unique visitors, 88.9% completed the questionnaire compared to the second and third questionnaires, which had completion rates of 93 and 75%, respectively (**Table 3**). The questionnaire received $n = 38$ answers in IG and $n = 122$ in CG. Twenty-four participants in IG received their caffeine test results five or more months ago (63% from the 38 of the total in IG). Four participants did not receive their reminder messages due to technical reasons.

TABLE 7 | Sample demographics according to the third questionnaire.

	Intervention	Control	p
Female	23 (28.02) [0.9]	95 (89.98) [0.28]	0.033
Male	15 (9.98) [2.53]	27 (32.02) [0.79]	
Age			
18–29	8 (9.38) [0.2]	31 (29.62) [0.06]	0.55
30–39	21 (14.91) [2.49]	41 (47.09) [0.79]	0.02
40–49	6 (9.38) [1.22]	33 (29.62) [0.39]	0.14
50–59	3 (3.37) [0.04]	11 (10.63) [0.01]	0.8
60–69	0	3	
70–79	0	1	
Education			
Primary	1	2	
Secondary	11 (12.19) [0.12]	40 (38.81) [0.04]	0.63
Bachelor's degree	16 (16.01) [0]	51 (50.99) [0]	0.99
Master's degree	8 (8.36) [0.02]	27 (26.64) [0]	0.86
Doctoral degree	2	1	

Table information is provided as follows: observed cell totals (expected cell totals) [chi-square statistic for each cell].

In general, 40 participants out of 160 answered how they felt after 1-month follow-up: one felt better physically, 10% ($n = 4$) felt better mentally, 25% ($n = 10$) felt better physically and mentally, 25% ($n = 10$) did not feel better and 37.5% ($n = 15$) did not know how to answer.

For the control question asking if they had set the goals and reminders, 59.1% ($n = 95$) answered no, 27.7% ($n = 44$) confirmed they did, and 13.2% ($n = 21$) did not recall whether they did or not.

Sixty-five people answered how long they had followed their reminder messages on a 1–5 scale. Over a quarter (38.5%, $n = 25$) "did not follow at all," equivalent to "1" on the scale. Seven (10.8%) answered "2," while nine (13.8%) answered "3." Reminders were well followed by 15.4% ($n = 10$), equivalent to "4" on the scale; and 21.5% ($n = 14$) answered "5."

Twenty-two people (33.8%) did not open the reminder messages at all (mobile or email), while 24.6% ($n = 16$) opened all messages. The middle groups scoring "2–4" were, respectively 15.4%, $n = 10$; 13.8%, $n = 9$; and 12.3%, $n = 8$.

Half of the people (50%, $n = 66$) thought reminders have been or would have been helpful to support behavior change, while 35.2% ($n = 56$) did not know how to answer and 23.9% ($n = 38$) thought they would not have been helpful.

Whether genetic test would have helped toward behavior change was supported by 71.8% ($n = 115$) in general, 22.6% ($n = 36$) did not know whether it would have helped, and 5.7% thought it would not have helped. People were also positively minded to receive other genetic trait information on a similar basis, where 80.6% ($n = 129$) agreed, 13.8% ($n = 22$) did not know how to answer, and 5.7% ($n = 9$) disagreed.

Some demographic differences were detected (**Table 7**). The age group 30–39 was generally more represented ($p = 0.02$) compared to the other age groups in the third questionnaire. In the IG, there were statistical differences, where male participants were more represented ($p = 0.033$).

TABLE 8 | Test of the study hypothesis that genetic tests make DTCGT participants change their behavior in the short term (1-month follow-up) when using eHealth services with strategic methods: information chunking and creating personalized reminders and goals on their digital device.

	IG (with reminders)	CG (including IG without reminders)	Marginal row totals
Changed behavior	8 (6.03) [0.64]	26 (27.97) [0.14]	34
Did not change	3 (4.97) [0.78]	25 (23.03) [0.17]	28
Marginal column totals	11	51	62 (grand total)

There is no association found with the p = 0.19. Table information is provided as follows: observed cell totals (expected cell totals) [chi-square statistic for each cell].

Setting of personalized goals and reminders was managed by special software, integrated into the MediKeep mobile application; additionally, it had the functionality to send automated emails. In some cases, the emails sent via this service ended up in junk email, and according to the third questionnaire and the "Technical question: did the reminders reach your email box or smartphone messages," two participants from both IG and CG did not receive the "reminder" messages (they were also removed from the statistical analysis). All participants were receiving the reminders within 1 month and at least 10 times and 3 days apart. Reminder exhaustion was reduced by manually changing the message text but not the context. Some participants received their messages in the morning while some did in the afternoon, depending on their personal preferences (example reminder at 1 PM: "Do not take another coffee cup in the afternoon and dinner! You will get better sleep at night!").

Statistical Calculations

The aim of the study was to assess if genetic tests made people change their behavior in the short term (1-month follow-up) when using eHealth services with strategic methods: information chunking and creating personalized reminders and goals. To calculate the results, a chi-square test is performed.

Participants who decided not to change their behavior or did not see the need to change were removed from both groups. Additionally, two participants who wished to change their behavior were removed from both groups for technical reasons, as they did not receive their reminders. The total of participants in the final calculations were $n = 18$ (47.3% of 38) in the IG and $n = 49$ (40.1% of 122) in the CG.

Testing the Aim of the Study

We did not detect differences between the IG group with set reminders changing their behavior in the short term ($p = 0.19$) and the other groups (**Table 8**). There were no differences in gender for those groups ($p = 0.32$). According to the qualitative data—a test of association—no cell should have data <1, and 20% of the cells should have data of 5 (38).

Other Findings

No differences were found between the groups who decided to change their behavior related to caffeine in the first place and those who did not (based on the second questionnaire, the *p*-value is 0.91). Participants in the IG felt better mentally or physically or both (*p*-value = 0.024) after adjusting their caffeine behavior after 1-month follow-up (**Table 9**).

TABLE 9 | Incidental finding: participants in the intervention group felt better (mentally or physically) after adjusting their caffeine intake with set goals and reminders after 1 month ($p = 0.024$).

	IG	CG	Marginal row totals
Felt better	8 (4.8) [2.13]	8 (11.2) [0.91]	16
Did not feel better or did not know how to answer	4 (7.2) [1.42]	20 (16.8) [0.61]	24
Marginal column totals	12	28	40 (grand total)

Table information is provided as follows: observed cell totals (expected cell totals) [chi-square statistic for each cell].

TABLE 10 | Sub-aim 3: does using goals and reminders help people using eHealth services to change their behavior in the short term (1-month follow-up)? Association found with the *p*-value = 0.013.

	Reminders set	Reminders not set	Marginal row totals
Changed behavior	29 (24.68) [0.76]	5 (9.32) [2]	34
Did not change	16 (20.32) [0.92]	12 (7.68) [2.43]	28
Marginal column totals	45	17	62 (grand total)

Table information is provided as follows: observed cell totals (expected cell totals) [chi-square statistic for each cell].

There was no difference detected between IG and CG concerning the frequency of opening the reminder messages (*p*-value = 0.59). Values 1–3 indicated "not opening much," and 4–5 indicated "opened a lot."

It was found that all the participants who wanted to change their behavior related to caffeine and had a goal/reminder set were more likely to change their behavior successfully after 1-month follow-up ($p = 0.013$), compared to the group who wanted to change but did not have reminders set (**Table 10**).

DISCUSSION

It is essential to start with the fact that the empirical research in the current study does not measure caffeine consumption or people's behavior to drink coffee; instead it explores their actions based on personalized nutrition information and whether creating goals and reminders, with the help of their digital devices, increases behavioral health toward positive actions in the short term. While the previous studies have shown contrasting results, some claiming that learning about genetic traits would benefit the person's health behavior (14, 17, 25, 28, 33), while others not finding evidence of said benefit (10, 17, 23, 24, 26, 27, 29–32, 34–36).

The empirical part of this study did not find that DTCGT will provide an extra benefit for such a personalized medicine eHealth service, which could be due to the limited sample size in the genetic testing group. In general, people's beliefs were strongly related to the genetic test and behavior change, where 71.8% of all respondents thought that the genetic test helped or would have helped them toward better behavior change. However, previous studies have found that the evidence of genetic testing in behavior change is questionable; there are other factors that could lead to behavior change: capability, opportunity, and motivation (20). The authors of this pilot study find that among all empirical study participants ($n = 160$), the base needs for motivation change were present. The people were genuinely interested in caffeine because they responded to the call in social media, where the main keywords were "genetic testing," "personalized medicine," and "caffeine." The participants also had the capability and opportunity present; since they were using their own digital devices, the tools and software were present already or were offered for free (mobile app MediKeep). The "act" of setting caffeine consumption goals and reminders and sticking to their plan was a relatively easy task when compared to the more substantial diet adjustments or regular visits to facilities (gym or clinic). As suggested by "The Behavior Change Wheel," the COMB-B system (20), the behavior change can be affected by increasing training, education, and enablement. The current study also affected those areas in participants by offering a piece of potential new information in personalized medicine—a detailed article about caffeine (in Estonian), including pharmacogenetic information and general health benefits or risks. The article was specially tailored for the general audience while maintaining the scientific value: list of references, terms explained, and primary focus on genetic results. While 98.3% of the participants understood the article content, only 69.2% confirmed the information to be new to them.

The relatively high percentage of participants feeling uneasy about reading the article (16.9% of $n = 172$ total with $n = 27$ participants without genetic test results) might be related to the fact that the article created a bit of confusion as it was expected to be read with caffeine genetic test results and genotype. Creating confusion was intentional and expected (because of the hope for educational factors). However, the number of participants who skipped the goal setting and decision making was not measured because it was hard to distinguish between people who thought they needed to change their behavior and participants who did not dare to change their behavior as they did not have their genetic test results. There was no association between IG and CT, i.e., those who set the goals and those who did not.

Most of the participants in the study were women (79.9%). In both study groups, participants with genetic testing (63% women) and people without genetic testing (84.7% women), there were more women, but there was a significantly higher men-to-women ratio in the CG.

Participants who opted in for goals and reminders 36% ($n = 62/172$) were more likely to stick to their goals after 1-month follow-up ($p = 0.013$). The sufficient amount of data confirms the findings of previous studies where eHealth behavior solutions seem to work very well (1–5). There was no statistically significant association found on whether participants with the genetic test would be more likely to opt in for goals and reminders ($p = 0.79$) or would more likely open the reminder messages on their digital devices. However, there is a small indication that DTCGT participants are more likely also to increase their nutrition (caffeine) consumption if they find it beneficial ($p = 0.046$). The general trend for caffeine reduction in the CG might be indicated from the fact that without a genetic test (a personalized decision support), it is safer to reduce caffeine intake rather than increase it—to live healthier and reduce possible risks related to caffeine consumption.

While a 1-month follow-up is considered to be short for behavior change measurement, the authors also found that 63% of people with genetic test results had found out their caffeine metabolism more than 5 months ago. As it was not a measurable factor for this study, it is essential to mention that, while this behavior change eHealth study seemed to work for the short term, it could also work for the long term. As commented by one of the IG participants who changed their behavior in five or more months,

> *I can not drink coffee, and genetic test gave me an answer why. The same thing is with strong tea. I can drink liters of Coca, Pepsi, or Red Bull. I quit drinking coffee after my results 5 or more months ago, and I stuck to that before the reminders were offered.*

Once people have adjusted their habits with or without the help of genetic or eHealth service, they might adapt it to their routine without the need for goals and reminders later on. Further studies on the topic for long-term behavior change are needed.

Association was found between those feeling better mentally or physically after 1-month caffeine adjustment in IG ($p = 0.024$) and people who did not feel better or did not know how to answer in IG or CG. While they actually might feel better, they also might feel better based on their self-reported results, because of the self-assurance or self-confirmation of justification to their DTCGT, since it is usually an expensive spending.

In Nielsen et al. (37), 52% confirmed of having heard nothing about DTCGT in Toronto; however, the situation has probably improved over time, and the current status in Estonia is 19.6% of participants know nothing about DTCGT. However, Estonians have been well-educated about genetic testing possibilities recently, as in last year, the national biobank had a major campaign to "gift Estonia with 100 000 new biobank participants to its 100th birthday." There is a small indicator that some of the empirical study participants confused the DTCGT with their previous biobank donation.

The knowledge about the term "nutrigenomics" had quite similar results in 2019 in Estonia when compared with the study in Toronto in 2012−25.7 and 30% respectively had not heard about the term. Similarly, only 5% had heard about it a lot previously. Similar results were also found for the other questions. However, Estonians had a little bit less expected anxiety about learning their genetic traits.

Suggestions for similar studies would include the tracking of all emails when possible (even when sent individually) to be sure that the participants read or opened them in the first place. The current study does not include why some participants with their DNA raw data did not respond to the second email, while participation in the genetic group was high in general. The following questions remain: whether participants felt insecure, whether participation was going to be difficult due to the sharing/signing/uploading request, or whether they did not receive the email.

As the central hypothesis of this study remains unanswered, further research is needed on larger study samples. The authors of this study suggest doubling the number of participants for more accurate statistical analysis: starting from 214 to at least 400.

Study Limitations

The genotype and caffeine metabolism information for the IG in this empirical research may not have been new for the target group. Many DTCGT companies offer caffeine metabolism information in nutrition reports; it may be a reason why some participants in the IG chose the option not to change anything in their behavior because they already did so a while ago and it has been a routine ever since. On the other hand, the CG got an article focusing on genetic information, and it might have been the reason for the participants without a genetic test to not set a personal goal.

The study group was relatively small, starting with $n = 213$ participants along with $n = 46$ with DTCGT raw data and finishing with $n = 160$ responses. The 1-month follow-up is considered to be short, and no long-term results were measured; however, over half of the participants with genetic test results got their first caffeine-related results more than 5 months ago from the third questionnaire. Most of the answers were self-reported by the participants, while only technical data were obtained elsewhere.

The study faced several technical problems: at some point, some of the reminders or invitations to questionnaires sent by email ended up in the spam filter. Secondly, there was a technical problem in sending out Android reminders in the first week; however, it did not seem to interrupt the results as they did receive messages at least 10 times in 1 month afterward.

As several aims or hypotheses have been investigated and several tests have been run, the resulting statistical data should be considered with some caution. Bonferroni correction should be considered to correct for multiple testing.

Study Strengths

The current study is the first known study of its kind in Estonia, and probably globally, including DTCGT, behavior change, and eHealth applications. The participants were not aware of the study aims. However, they were genuinely interested in using eHealth interventions.

CONCLUSIONS

As genetic testing has become more affordable for the general public and has been accessible for everyone who wants to get tested via DTCGT companies, the following question remains: whether in the rise of eHealth apps and behavior change programs, the personalized genetic trait disclosure would add an extra benefit for the person's health behavior. While some studies prove it actionable, others remain uncertain. While the authors of this study dug deeper into the question, it was revealed based on the previous meta-analysis of 18 studies worldwide that participating in the study or eHealth program shows positive results in health behavior.

The most active association found in the empirical part of this study was related to the idea claiming that the eHealth applications work (1–5) for behavior change in the short term. However, it was not possible to find differentiation between the DTCGT group and the CG, in adding up for the behavior change, because of the insufficient amount of data. So, the following question remains: whether genetic testing for behavior health change is beneficial. The IG also represented more male participants aged 30–39.

The secondary findings created additional conclusions where people with genetic test results were more likely to increase caffeine intake if they found caffeine to be beneficial or not harmful. They were feeling better (mentally, physically, or both) after their decision to change their behavior related to caffeine in 1 month. So in light of the questions of whether one should increase the nutrition intake or whether they need to decide on a specific nutrition to feel better, genetic testing could be considered. Further studies among larger study groups are needed to have a better understanding of behavior health change in personalized medicine. eHealth applications for short-term behavior change have a positive effect, regardless of a previous genetic disclosure.

DATA AVAILABILITY STATEMENT

The raw data supporting the conclusions of this article will be made available by the authors, without undue reservation.

ETHICS STATEMENT

The studies involving human participants were reviewed and approved by The study has Research Ethics Committee of the University of Tartu approval no: 290/T-10 since 8.03.2019 until 30.06.2019. The patients/participants provided their written informed consent to participate in this study.

AUTHOR CONTRIBUTIONS

KA: main work and research, empirical study. MV: supervision and reviews. AK: forming the study aim questions and mentoring for the statistical calculations. All authors contributed to the article and approved the submitted version.

ACKNOWLEDGMENTS

We are very grateful to all the participants in the study and especially to the MediKeep genetic testing group. We want to send our deepest gratitude to Reedik Mägi, Ph.D., for his time and contribution to the base of this study. Special thanks to Priit Kruus and Dr. Peeter Ross from Tallinn University of Technology (TalTech). We thank Natalia Pervjakova for the supervision and challenge.

SUPPLEMENTARY MATERIAL

The Supplementary Material for this article can be found online at: https://www.frontiersin.org/articles/10.3389/fdgth.2020.567656/full#supplementary-material

REFERENCES

1. Marsaux CF, Celis-Morales C, Livingstone KM, Fallaize R, Kolossa S, Hallmann J, et al. Changes in physical activity following a genetic-based internet-delivered personalized intervention: randomized controlled trial (Food4Me). *J Med Internet Res.* (2016) 18:e30. doi: 10.2196/jmir.5198
2. Do HP, Tran BX, Pham QL, Nguyen LH, Tran TT, Latkin CA, et al. Which eHealth interventions are most effective for smoking cessation? A systematic review. *Patient Prefer Adherence.* (2018) 12:2065–84. doi: 10.2147/PPA.S169397
3. Lo WLA, Lei D, Li L, Huang DF, Tong KF. The perceived benefits of an artificial intelligence-embedded mobile app implementing evidence-based guidelines for the self-management of chronic neck and back pain: observational study. *JMIR Mhealth Uhealth.* (2018) 6:e198. doi: 10.2196/mhealth.8127
4. Brouwer W, Kroeze W, Crutzen R, de Nooijer J, de Vries NK, Brug J, et al. Which intervention characteristics are related to more exposure to internet-delivered healthy lifestyle promotion interventions? A systematic review. *J Med Internet Res.* (2011) 13:e2.doi: 10.2196/jmir.1639
5. Cugelman B, Thelwall Mz Dawes P. Online interventions for social marketing health behavior change campaigns: a meta-analysis of psychological architectures and adherence factors. *J Med Internet Res.* (2011) 13:e17. doi: 10.2196/jmir.1367
6. Street, Jr RL, Makoul G, Arora NK, Epstein RM. How does communication heal? Pathways linking clinician-patient communication to health outcomes. *Patient Educ Couns.* (2009) 74:295–301. doi: 10.1016/j.pec.2008.11.015
7. Olson JE, Rohrer Vitek CR, Bell EJ, McGree ME, Jacobson DJ, St. Sauver JL, et al. Participant-perceived understanding and perspectives on pharmacogenomics: the Mayo Clinic RIGHT protocol (Right Drug, Right Dose, Right Time). *Genet Med.* (2017) 19:819–25. doi: 10.1038/gim.2016.192
8. Solopchuk O, Alamia A, Olivier E, Zénon A. Chunking improves symbolic sequence processing and relies on working memory gating mechanisms. *Learn Mem.* (2016) 23:108–12. doi: 10.1101/lm.041277.115
9. Dute DJ, Bemelmans WJ, Breda J. Using mobile apps to promote a healthy lifestyle among adolescents and students: a review of the theoretical basis and lessons learned. *JMIR Mhealth Uhealth.* (2016) 4:e39. doi: 10.2196/mhealth.3559
10. Audrain J, Boyd NR, Roth J, Main D, Caporaso NF, Lerman C. Genetic susceptibility testing in smoking-cessation treatment: one-year outcomes of a randomized trial. *Addict Behav.* (1997) 22:741–51. doi: 10.1016/S0306-4603(97)00060-9
11. Stewart KFJ, Wesselius A, Schreurs MAC, Schols AMWJ, Zeegers MP. Behavioural changes, sharing behaviour and psychological responses after receiving direct-to-consumer genetic test results: a systematic review and meta-analysis. *J Community Genet.* (2018) 9:1–18. doi: 10.1007/s12687-017-0310-z
12. Meisel SF, Fraser LSM, Side L, Gessler S, Hann KEJ, Wardle J, et al. Anticipated health behaviour changes and perceived control in response to disclosure of genetic risk of breast and ovarian cancer: a quantitative survey study among women in the UK. *BMJ Open.* (2017) 7:e017675. doi: 10.1136/bmjopen-2017-017675
13. Stewart KFJ, Wesselius A, Schols AMW J, Zeegers MP. Stages of behavioural change after direct-to-consumer disease risk profiling: study protocol of two integrated controlled pragmatic trials. *Trials.* (2018) 19:240. doi: 10.1186/s13063-018-2630-7
14. Chao S, Roberts JS, Marteau TM, Silliman R, Cupples LA, Green RC. Health behavior changes after genetic risk assessment for Alzheimer disease: the REVEAL Study. *Alzheimer Dis Assoc Disord.* (2008) 22:94–97. doi: 10.1097/WAD.0b013e31815a9dcc
15. Hurling R, Catt M, De Boni M, Fairley B, Hurst T, Murray P, et al. Using internet and mobile phone technology to deliver an automated physical activity program: randomized controlled trial. *J Med Internet Res.* (2007) 9:e7. doi: 10.2196/jmir.9.2.e7
16. Orleans CT, Schoenbach VJ, Wagner EH, Quade D, Salmon MA, Pearson DC, et al. Self-help quit smoking interventions: effects of self-help materials, social support instructions, telephone counseling. *J Consult Clin Psychol.* (1991) 59:439–48. doi: 10.1037/0022-006X.59.3.439
17. Hietaranta-Luoma HL, Tahvonen R, Iso-Touru T, Puolijoki H, Hopia A. An intervention study of individual, apoE genotype-based dietary and physical-activity advice: impact on health behavior. *J Nutrigenet Nutrigenomics.* (2014) 7:161–74. doi: 10.1159/000371743
18. Schmidlen T, Sturm AC, Hovick S, Scheinfeldt L, Scott RJ, Morr L, et al. Operationalizing the reciprocal engagement model of genetic counseling practice: a framework for the scalable delivery of genomic counseling and testing. *J Genet Couns.* (2018) 27:1111–29. doi: 10.1007/s10897-018-0230-z
19. Michie S, Abraham C, Whittington C, McAteer J, Gupta S. Effective techniques in healthy eating and physical activity interventions: a meta-regression. *Health Psychol.* (2009) 28:690–701. doi: 10.1037/a0016136
20. Michie S, van Stralen MM, West R. The behaviour change wheel: A new method for characterising and designing behaviour change interventions. *Implementation Sci.* (2011) 6:42. doi: 10.1186/1748-5908-6-42
21. Edwards EA, Lumsden J, Rivas C, Steed L, Edwards A, Thiyagarajan A, et al. (2016) Gamification for health promotion: systematic review of behaviour change techniques in smartphone apps*BMJ Open* 6:e012447. doi: 10.1136/bmjopen-2016-012447
22. Hollands GJ, D. P., Freench, Griffin SJ, Prevost T, Sutton S, et al. The impact of communicating genetic risks of disease on risk-reducing health behaviour: systematic review with meta-analysis. *BMJ.* (2016) 352:i1102. doi: 10.1136/bmj.i1102
23. Hishida A, Terazawa T, Mamiya T, Ito H, Matsuo K, Tajima K, et al. Efficacy of genotype notification to Japanese smokers on smoking cessation–an intervention study at workplace. *Cancer Epidemiol.* (2010) 34:96–100. doi: 10.1016/j.canep.2009.11.008
24. Ito H, Matsuo K, Wakai K, Toshiko S, Hiroshi K, Katashi O, et al. An intervention study of smoking cessation with feedback on genetic cancer susceptibility in Japan. *Prev Med.* (2006) 42:102–8. doi: 10.1016/j.ypmed.2005.10.006
25. McBride CM, Bepler G, Lipkus IM, Lyna P, Samsa G, Albright J, et al. Incorporating genetic susceptibility feedback into a smoking cessation program for African-American smokers with low income. *Cancer Epidemiol Biomarkers Prev.* (2002) 11:521-528.
26. Sanderson SC, Humphries SE, Hubbart C, Hughes E, Jarvis MJ, Wardle J. Psychological and behavioural impact of genetic testing smokers for lung cancer risk: a phase II exploratory trial. *J Health Psychol.* (2008) 13:481–94. doi: 10.1177/1359105308088519
27. Hendershot CS, Otto JM, Collins SE, Liang T, Wall TL. Evaluation of a brief web-based genetic feedback intervention for reducing alcohol-related health risks associated with ALDH2. *Ann Behav Med.* (2010) 40:77–88. doi: 10.1007/s12160-010-9207-3
28. Komiya Y, Nakao H, Kuroda Y, Arizono K, Nakahara A, Katoh T. Application of aldehyde dehydrogenase 2 (ALDH2) genetic diagnosis in support of decreasing alcohol intake. *J Occup Health.* (2006) 48:161–5. doi: 10.1539/joh.48.161
29. Glanz K, Volpicelli K, Kanetsky PA, Ming ME, Schuchter LM, Jepson C, et al. Melanoma genetic testing, counseling, and adherence to skin cancer

prevention and detection behaviors. *Cancer Epidemiol Biomarkers Prev.* (2013) 22:607–14. doi: 10.1158/1055-9965.EPI-12-1174

30. Godino JG, van Sluijs EM, Marteau TM, Sutton S, Sharp SJ, Griffin SJ. Lifestyle advice combined with personalized estimates of genetic or phenotypic risk of type 2 diabetes, and objectively measured physical activity: a randomized controlled trial. *PLoS Med.* (2016) 13:e1002185. doi: 10.1371/journal.pmed.1002185
31. Marteau T, Senior V, Humphries SE, Bobrow M, Cranston T, Crook MA, et al. Psychological impact of genetic testing for familial hypercholesterolemia within a previously aware population: a randomized controlled trial. *Am J Med Genet A*. (2004) 128A:285–93. doi: 10.1002/ajmg.a.30102
32. Meisel SF, Beeken RJ, van Jaarsveld CH, Wardle J. Genetic susceptibility testing and readiness to control weight: results from a randomized controlled trial. *Obesity.* (2015) 23:305–12. doi: 10.1002/oby.20958
33. Nielsen DE, El-Sohemy A. Disclosure of genetic information and change in dietary intake: A randomized controlled trial. *PLoS ONE.* (2014) 9:e112665. doi: 10.1371/journal.pone.0112665
34. Voils CI, Coffman CJ, Grubber JM, Edelman D, Sadeghpour A, Maciejewski ML, et al. Does type 2 diabetes genetic testing and counseling reduce modifiable risk factors? A randomized controlled trial of veterans. *J Gen Intern Med.* (2015) 30:1591–8. doi: 10.1007/s11606-015-3315-5
35. Weinberg DS, Myers RE, Keenan E, Ruth K, Sifri R, Ziring B, et al. Genetic and environmental risk assessment and colorectal cancer screening in an average-risk population: a randomized trial. *Ann Intern Med.* (2014) 161:537–45. doi: 10.7326/M14-0765
36. Grant RW, Kelsey E, O'Brien JL, Waxler JL, Delahanty LG, Bissett R, et al. Personalized genetic risk counseling to motivate diabetes prevention: a randomized trial. *Diabetes Care.* (2013) 36:13–19. doi: 10.2337/dc12-0884
37. Nielsen DE, El-Sohemy A. A randomized trial of genetic information for personalized nutrition. *Genes Nutr.* (2012) 7:559–66. doi: 10.1007/s12263-012-0290-x
38. Bewick V, Cheek L, Ball J. Statistics review 8: Qualitative data - tests of association. *Crit Care.* (2004) 8:46–53. doi: 10.1186/cc2428

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 Big Data

OPINION
published: 12 February 2021
doi: 10.3389/fdata.2020.603044

A Policy and Practice Review of Consumer Protections and Their Application to Hospital-Sourced Data Aggregation and Analytics by Third-Party Companies

Vasiliki Rahimzadeh *

Stanford Center for Biomedical Ethics, Stanford University, Stanford, CA, United States

The Office of the National Coordinator for Health Information Technology estimates that 96% of all U.S. hospitals use a basic electronic health record, but only 62% are able to exchange health information with outside providers. Barriers to information exchange across EHR systems challenge data aggregation and analysis that hospitals need to evaluate healthcare quality and safety. A growing number of hospital systems are partnering with third-party companies to provide these services. In exchange, companies reserve the rights to sell the aggregated data and analyses produced therefrom, often without the knowledge of patients from whom the data were sourced. Such partnerships fall in a regulatory grey area and raise new ethical questions about whether health, consumer, or health and consumer privacy protections apply. The current opinion probes this question in the context of consumer privacy reform in California. It analyzes protections for health information recently expanded under the California Consumer Privacy Act ("CA Privacy Act") in 2020 and compares them to protections outlined in the Health Information Portability and Accountability Act ("Federal Privacy Rule"). Four perspectives are considered in this ethical analysis: 1) standards of data deidentification; 2) rights of patients and consumers in relation to their health information; 3) entities covered by the CA Privacy Act; 4) scope and complementarity of federal and state regulations. The opinion concludes that the CCPA is limited in its application when health information is processed by a third-party data aggregation company that is contractually designated as a business associate; when health information is deidentified; and when hospital data are sourced from publicly owned and operated hospitals. Lastly, the opinion offers practical recommendations for facilitating parity between state and federal health data privacy laws and for how a more equitable distribution of informational risks and benefits from the sale of aggregated hospital data could be fostered and presents ways

OPEN ACCESS

Edited by:
Priit Kruus,
Tallinn University of Technology,
Estonia

Reviewed by:
Anastasiya Kiseleva,
Vrije University Brussel, Belgium

****Correspondence:***
Vasiliki Rahimzadeh
vrahim@stanford.edu

Specialty section:
This article was submitted to
Medicine and Public Health,
a section of the journal
Frontiers in Big Data

Received: *04 September 2020*
Accepted: *21 December 2020*
Published: *12 February 2021*

Citation:
Rahimzadeh V (2021) A Policy and Practice Review of Consumer Protections and Their Application to Hospital-Sourced Data Aggregation and Analytics by Third-Party Companies.
Front. Big Data 3:603044.
doi: 10.3389/fdata.2020.603044

Abbreviations: CCPA, California Consumer Privacy Act (referred to as the "CA Privacy Act"); CPRA, California Privacy and Enforcement Act (i.e., Proposition 24); EHR, electronic health record; HIO, health information organization; HIPAA, Health Information Portability and Accountability Act (referred to as the "Federal Privacy Rule"); TDAC, third-party data aggregation company.

both for-profit and nonprofit hospitals can sustain patient trust when negotiating partnerships with third-party data aggregation companies.

Keywords: data aggregation, EHR, privacy, HIPAA, California Consumer Privacy Act, Proposition 24

INTRODUCTION

Less is certainly not more when aggregation of quality hospital system data is concerned. Indeed, aggregation puts the "big" in big data. Aggregation refers to the semantic integration of datasets from disparate sources, sizes, and elements into a shareable format. It allows for cross-system analyses of hospital trends shown to reduce medical error, inform safer therapies, and enable timely public health reporting (Fefferman et al., 2005), to name but a few applications (Olsen et al., 2007). New machine learning and artificial intelligence applications in healthcare likewise depend on robust data aggregation for training algorithms to automate certain care delivery tasks with precision and effectiveness (Char et al., 2018). While these data are primarily aggregated through extraction from electronic health records (EHR), and follow a complex trajectory from the point of care to aggregation (Rolnick, 2013), problems with EHR network interoperability largely persist across U.S. hospitals despite regulatory reforms to improve their meaningful use in 2009 (United States Congress, 2009) and again in 2016 (21st Century Cures Act, 2016).

Hospitals are handicapped in performing aggregation in-house due, in large part, to limited availability of EHR-based rather than insurance claims-based data, exceedingly high administrative costs of producing datasets, and technological limitations involving software (4). A growing market for third-party data aggregation services is poised to fill critical infrastructural gaps that federal agencies have been thus far slow to fill (Wang et al., 2017; Groves et al., 2013; Challenge.gov, 2009). Optum One, for instance, describes their data aggregation services as "source- and vendor-agnostic," meaning the company integrates claims, clinical, sociodemographic, genetic, and care management data—herein referred to as hospital data—to identify population-level patterns irrespective of the record platform from which the data originated.

Analyses performed on the aggregate data can be subsequently fed back to the hospitals to inform quality improvement, clinical teaching, and research, among others (Fefferman et al., 2005). Third-party data aggregation companies (TDAC) reserve the right to sell the aggregate data for marketing and other commercial purposes, provided that the data are appropriately protected. Health data (e.g., from EHRs, insurance claims databases, and genetic data) are distinct from other common consumer data types (e.g., credit card numbers, geolocation, and demographic data). The use and disclosure of protected health information are governed federally by the Health Information Portability and Accountability Act (HIPAA, herein referred to as the Federal Privacy Rule), while the Federal Trade Commission has jurisdiction over consumer data. Since 2018, three states have also passed their own Internet consumer privacy legislation in California, Nevada, and Maine (National Conference of State Legislatures, 2009).

Though the legislations differ in scale and scope, they broadly aim to strengthen the rights of consumers to decide what, how, and with whom their personal information is shared. These rights and protections applied exclusively to consumer data until Californians voted to approve Proposition 24 during the latest State Elections in November 2020. Among other amendments, Proposition 24 expanded protections outlined in the existing California Consumer Privacy Act (herein referred to as the "CA Privacy Act") to include health information as a special category of sensitive personal information and "the unauthorized use or disclosure of which creates a heightened risk of harm to the consumer" (State of California, 2018).

The expanded protections blur the neat legislative distinction between personal and health information protections under the CA Privacy Act. As Price aptly notes, health information held by entities outside the Federal Privacy Rule's ambit "might seem to improve the problem of data fragmentation; these entities can gather data unhindered by HIPAA's strictures. On the other hand, fragmentation may increase because different entities, with different forms of health data, are governed by different legal regimes" (Price, 2018).

Greater involvement of third-party aggregation of hospital-sourced data prompts asking whether individuals are patients, consumers, or both under applicable privacy laws and raises new ethical questions about what rights individuals have in the emerging medical datasphere (Béranger, 2016). It is unclear, for example, if contractual relationships between data aggregation companies and hospitals or the aggregation tasks a company performs determine which privacy regimes should apply. Reflections on these questions regarding patient rights to know whether and how their data are shared could have broader implications if other states follow suit in expanding special consumer protections to health information.

This opinion probes these questions through a close reading of the expanded protections for health information in the CA Privacy Act. Specifically, it analyzes the protections afforded to hospital-sourced data aggregated by TDACs from four ethical perspectives: 1) standards of data deidentification to minimize informational risk; 2) rights of patients and consumers in relation to their health information; 3) entities covered by the CA Privacy Act; 4) and scopes of Federal and State regulations. The opinion concludes with practical recommendations for how to achieve a more equitable distribution of informational risks and benefits from the sale of aggregated hospital data and ways to sustain patient trust in private-private partnerships between hospitals and TDACs.

Deidentification Requirements: Separate but Equal?

Both the Federal Privacy Rule and the CA Privacy Act acknowledge that certain types of data merit special protection and generally agree on the inherent characteristics that make health information identifying. The Federal Privacy Rule explicitly governs the use and disclosure of identifying health information, termed protected health information, while the CA Privacy Act protects much broader categories of identifiable personal information. The main goal of the Federal Privacy Rule is to "assure that individuals' health information is properly protected while allowing the flow of health information needed to provide and promote high quality health care and to protect the public's health and well being ... Given that the health care marketplace is diverse, the Rule is designed to be flexible and comprehensive to cover the variety of uses and disclosures that need to be addressed" (Department of Health and Human Services, 2003).

The Federal Privacy Rule and the CA Privacy Act both exempt deidentified data. Additionally, the CA Privacy Act exempts protected health information that is used and disclosed by covered entities and business associates subject to the Federal Privacy Rule. Together, these exemptions allow for deidentified health data to be securely and efficiently exchanged for quality improvement purposes, approved health research and public health management, and many other uses. It is important to note that aggregate datasets can include readily identifiable, coded (i.e., personal identifiers are linked to the data by secure keys held by those processing the data), and deidentified (i.e., irreversibly delinked) information.

The Federal Privacy Rule applies prescriptive standards for determining when protected health information is appropriately deidentified, whereas the CA Privacy Act applies a reasonableness standard. The Federal Privacy Rule requires that information must be stripped of 18 unique identifiers to be deemed deidentified, termed the safe harbor rules, or otherwise verified by a field expert. The prescriptiveness of the Federal Privacy Rule leaves little room for interpretation and therefore can be more consistently applied across health systems, providers, and research institutions.

The CA Privacy Act, in contrast, applies the Federal Trade Commission's proposed reasonability standard for deidentification. This standard requires that to be deidentified, data "cannot reasonably identify, relate to, describe, be capable of being associated with, or be linked, directly or indirectly, to a particular consumer." Reasonableness is both data- and context-specific and, as a result, interpretable. That is, some types of data carry a higher likelihood of harm resulting from reidentification depending on how they are shared and with whom. The requirement for deidentification under the CA Privacy Act thus transcends specific categories and does not adopt predetermined methods to fulfill the deidentification requirement. The reasonability standard allows for deidentification to be determined in relation to actual environments in which data are exchanged and their associated risks. In this way, the reasonability standard can tailor deidentification methods to the specific data use and can be more flexible to emerging advances in privacy-preserving technologies and accountability policies, where applicable.

Could TDACs achieve comparable protections for sensitive personal information, e.g., health information under the CA Privacy Act? Several scenarios are possible. TDACs, other businesses, and data brokers subject to the CA Privacy Act could adopt the HIPAA safe harbor rules or apply the expert determination method to deidentify health information. In this case, health information used and shared by TDACs would be protected using the same deidentification standards as if it were managed by a HIPAA-covered entity. Alternatively, companies could apply stricter deidentification requirements and therefore grant patients additional protection compared to what is federally required. This could be the case if a TDAC demonstrates the health information it aggregates can still be reasonably reidentified despite applying the safe harbor or expert determination methods. Finally, there is the possibility that companies could exploit the flexibility built into the reasonability standard and adopt weaker deidentification practices, making health information less secure under the CA Privacy Act.

Data protection scholars and ethicists alike agree that deidentification is a spectrum and not a uniform standard (Stalla-Bourdillon and Wu, 2019). Indeed, some types of inherently identifying health information (e.g., genetic data (Homer et al., 2008)) pose challenges to the efficacy of both the reasonability standard and prescriptive approaches to deidentification. As reidentification becomes more "reasonable" with advanced information technologies (Kulynych and Greely, 2017), prescriptive deidentification strategies can quickly become outdated. So while the Privacy Rule applies deidentification standards consistently, those standards can underprotect particularly sensitive types of health information. The reasonability standard may better tailor data protections to the unique sensitivities and risks of disclosure, but its flexibility can mean that protections are applied inconsistently across the various entities which collect, use, and share this information. The next section explains the case when TDACs are contractually obligated to adopt the Federal Privacy Rule's more granular deidentification standard for hospital data.

The Business Association Designation: Health Insurance Portability and Accountability Act

One regulatory pathway by which TDACs can use and disclose protected health information is to serve as a "business associate" of a HIPAA-covered entity. Covered entities can include healthcare providers, health plans, or healthcare clearinghouses. TDACs could receive protected health information from hospitals prior to aggregation and subsequently deidentify it on behalf of the HIPAA-covered entity as part of a business associate agreement provided that they apply the expert determination standard or the safe harbor rules. The Department of Health and Human Services also recognizes "data aggregation" among the qualified services a TDAC could perform under a special type of business associate's agreement (Department of Health and Human Services, 2008), called a health information organization (HIO). The HIO designation permits also TDACs to, among other things, provide data aggregation services related to the healthcare operations of the covered entities for which it has agreements.

Both patients and companies have the potential to benefit from data aggregation partnerships. Hospitals can better serve patients through monitoring quality, safety, and provider performance data that TDACs make available. TDACs benefit financially from providing aggregation services and selling trend analyses not only to individual hospitals they may partner with directly but also to researchers and other companies. These revenues allow companies to invest in new information technologies that further expand the services they can provide to hospital systems within their network. The Federal Privacy Rule permits also TDACs to share deidentified data beyond the healthcare operations.

There is growing ethical concern about the emergence of new markets for aggregated hospital data and how companies may take advantage of regulatory loopholes to bypass consent from patients themselves. A TDAC that contracts with a hospital as a business associate can legally receive health information from the covered entity, deidentify it, and sell the deidentified data in the aggregate as well as any resulting trend analyses for the company's own commercial gain without patient authorization. Individuals treated at hospitals which partner with TDACs are often unaware that such partnerships exist and that their protected health information—albeit deidentified—is being sold by third-party companies for commercial purposes in many cases (Price et al., 2019).

Deidentified hospital data can be sold without a patient's authorization under a TDAC's business associate agreement; however, patients may have the option to invoke their right to an accounting of disclosures to better understand with whom their protected health information has been shared. The Federal Privacy Rule permits individuals under 45 CFR § 164.528 to obtain a record of certain disclosures of their protected health information by covered entities or their business associates, including TDACs where applicable. Covered entities and business associates are required to account for any and all disclosures of an individual's protected health information unless it was to carry out treatment, payment, and healthcare operations; it was for national security or intelligence purposes or related to correctional institutions or law enforcement officials; it was part of a limited dataset or occurred prior to the compliance date (April 2003). Requesting an accounting of disclosures could allow patients some transparency about existing partnerships between the hospital and any third-party companies it contracts with to manage protected health information if unknown to patients at the time of care (See **Supplementary Material**).

Hospitals are also not obligated to use the data TDACs aggregate for quality improvement. Hospitals also cannot condition the future sale of this data on such improvement. Importantly, neither the covered entity that contracts with the HIO or the HIO itself is liable if a violation of the Federal Privacy Rule is discovered and an appropriate business associate agreement is in place. The HIO is instead required to report any noncompliance with the agreement terms to the covered entity. A covered entity is moreover not required to oversee HIO compliance but must act to address the noncompliance when disclosed or else terminate the agreement. Accountability for patient privacy, therefore, rests on 1) elective disclosure of noncompliance by the HIO and 2) swift action on the part of hospitals to cure the noncompliance, and liability for the privacy violation remains ambiguous. While permissible under a recognized business associate agreement, there is a chance the sale and exchange of aggregate hospital data could disproportionately benefit companies. Patients, in turn, assume the informational risks associated with having their data aggregated and sold with limited ability to share directly in the benefits. Consumer data protections, in contrast, may afford greater agency in the sale of personal information that in the future could include more categories of health data. The following section illustrates how through discussing recent reforms to consumer data protections in California.

Data Brokering under the CA

Privacy Act

The CA Privacy Act was introduced in 2018 as a state-wide legislation to afford California consumers more control over personal information that businesses and data brokers collect about them (See **Supplementary Material**). Assembly Bill No. 375 effectively enacted the CA Privacy Act on January 1, 2020, and grants consumers four primary rights:

1 Right to know: A consumer may request that a business disclose 1) categories of personal information it collects about them, 2) the sources of that information, 3) the business purposes for collecting or selling the information, and 4) third parties with which the information is being shared.

2 Right to delete: A consumer may request that a business deletes personal information and requires businesses to follow through on a verified request.

3 Right to opt out: A consumer's may direct a business not to sell their personal information at any time.[1]

4 Right to nondiscrimination: A business shall not discriminate against a consumer because they exercised their rights under the Act.

The CA Privacy Act further obligates the business with revenues greater than $15M to process data from more than 50,000 individuals, [2] households, or devices, where more than 50% of revenues are derived from the sale of information to assure consumers it has no intent to reidentify the data; has implemented technical safeguards and processes to prohibit reidentification; has taken necessary steps to prevent inadvertent release of deidentified data.

While it covers all uses, disclosures, and management of protected health information, the Federal Privacy Rule is not presumed to complement state-based consumer data protections. Many hospitals are designated not-for-profit institutions or designated as a HIPAA-covered entity and therefore exempt from the CA Privacy Act. According to the California Department of Health and Human Services, 56 (7%) of the 492 registered acute care hospitals in California are governed by for-profit corporations (California Department of Health and Human Services Facilities List Data, 2021).

1 the entity meets the criteria designating them a "business" or "data broker";

[1] In addition, the CA Privacy Act includes a special opt-in clause for the sale and brokering of personal health information from minors. A business is not permitted to sell personal information from consumers younger than 16 years of age unless a consumer older than 13 but less than 16 or their guardian explicitly consents to the sale.

[2] Passage of Proposition 24 following the 2020 Elections increased this threshold to 100,000.

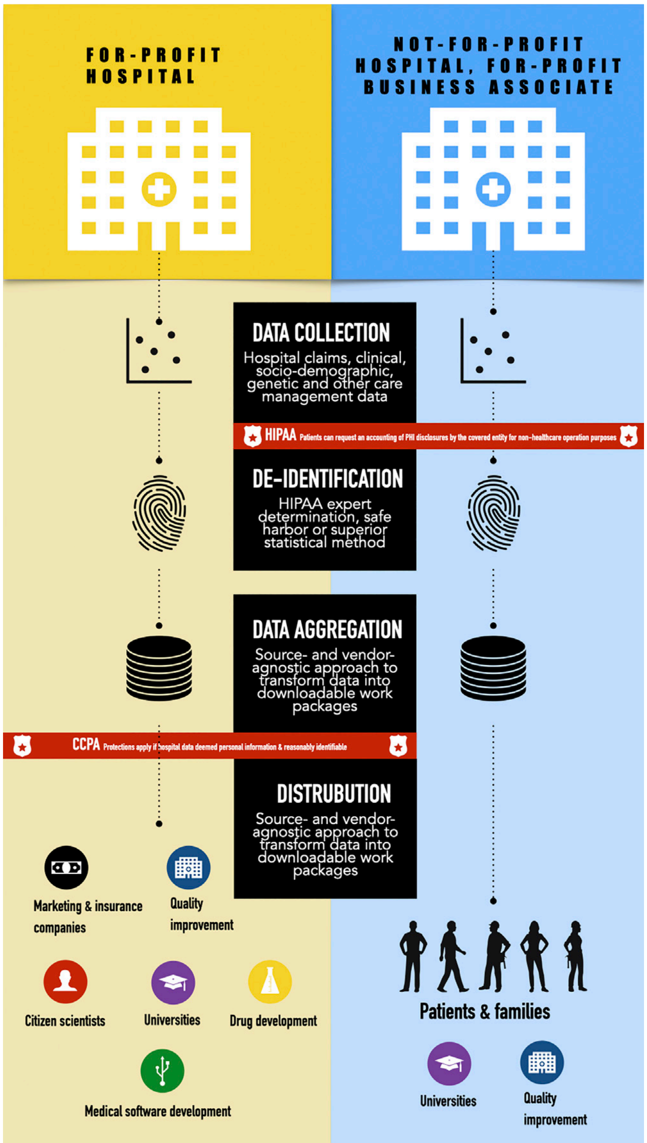

FOR-PROFIT HOSPITAL
NOT-FOR-PROFIT HOSPITAL, FOR-PROFIT BUSINESS ASSOCIATE
DATA COLLECTION
Hospital claims, clinical, socio-demographic, genetic and other care management data
HIPAA Patients can request an accounting of PHI disclosures by the covered entity for non-healthcare operation purposes
DE-IDENTIFICATION
HIPAA expert determination, safe harbor or superior statistical method
DATA AGGREGATION
Source- and vendor-agnostic approach to transform data into downloadable work packages
CCPA Protections apply if hospital data deemed personal information & reasonably identifiable
DISTRUBUTION
Source- and vendor-agnostic approach to transform data into downloadable work packages
Marketing & insurance companies
Quality improvement
Citizen scientists
Universities
Drug development
Medical software development
Patients & families
Universities
Quality improvement

FIGURE 1 | Comparison of applicable health and consumer data protections between for-profit hospitals and non-for-profit hospitals with for-profit business associates.

2 they are not a regulated entity that manages patient information according to HIPAA or California Medical Information Act regulations;
3 the data collected, used, shared, or sold are "reasonably identifiable" (Noordyke, 2020);
4 data include financial account information, racial or ethnic origin, religious beliefs, union membership, sexual orientation, genetic data, and precise geolocation data.

The consumers, i.e., patients about whom TDACs broker personal information, may be able to exercise additional consumer data privacy rights in some states where for-profit hospitals operate based in part on their federal compliance, nature, and type of data brokering activities and identifiability of the aggregate data.

A close reading of the CA Privacy Act in conversation with the Federal Privacy Rule reveals that TDACs which operate without a business associate agreement and aggregate "reasonably" identifiable health information are liable under the CA Privacy Act. Patients can also exercise the four rights to know, correct, delete, and opt out of the sale of their personal information described above (**Figure 1**). Operationalizing these rights is not without specific logistical and feasibility challenges in the ways patients are informed about how their data are used/shared with TDACs. The delivery and timing of this information could be especially fraught in an emergency or other serious clinical situations in which patients may not be fully able to appreciate the short- and long-term implications of what types of data will be aggregated and sold nor able to navigate the digital minefield that is submitting a verified opt-out request.

RECOMMENDATIONS

Complementary protections at the federal and state levels is essential for sustaining public trust with patients and consumers, particularly if more states follow California's lead. More explicit federal and state guidance is therefore needed regarding the nature and scope of data aggregation activities TDACs can perform using hospital-sourced information. First, the Office of the National Coordinator could consider narrowing permissions for how TDACs access, use, and disclose aggregate hospital data for which existing deidentification methods may be insufficient, for example, involving data that are particularly identifying or stigmatizing. Second, the National Coordinator should work more closely with state legislatures in the process of drafting consumer privacy legislation that propose to include health information to ensure complementarity. TDACs should consider, for example, applying the safe harbor, expert determination, or a superior method of deidentification to achieve complementarity with the Federal Privacy Rule.

Finally, more changes to the interplay of state and federal privacy protections for health information are expected following the approval of Proposition 24, otherwise called the California Privacy Rights and Enforcement Act. The revised CA Privacy Act in California is set to come into full force on January 1, 2023. It grants the state and California businesses new powers that have important implications for implementing expanded protections for "sensitive" personal information, specifically health and genetic information (**Table 1**). Proposition 24 carves out funding for a new agency that will oversee the amended CA Privacy Act enforcement to issue penalties and manage all consumer correction/deletion/opt-out requests. Businesses are also permitted to pass on a portion of the cost for complying with the expanded CA Privacy Act onto consumers. Indeed, the American Civil Liberties Union opposed Proposition 24 primarily for this reason. The new enforcement agency should therefore consider placing caps on how much companies can charge for stricter privacy protections, if not eliminate them outright. Capping the amount companies can pass on to consumers helps avoid

TABLE 1 | Section 10 regarding use and sale of "sensitive" information added to the California Consumer Privacy Act following vote to approve Proposition 24 in November 2020.

SEC. 10. Section 1798.121 is added to the Civil Code, to read: 1798.121. Consumers' Right to Limit Use and Disclosure of Sensitive Personal Information 1798.121

(a) **A consumer shall have the right, at any time, to direct a business that collects sensitive personal information about the consumer to limit its use of the consumer's sensitive personal information to that use which is necessary to perform the services or provide the goods** reasonably expected by an average consumer who requests those goods or services, to perform the services set forth in paragraphs (2), (4), (5), and (8) of subdivision (e) of Section 1798.140, and as authorized by regulations adopted pursuant to subparagraph (C) of paragraph (19) of subdivision (a) of Section 1798.185. A business that uses or discloses a consumer's sensitive personal information for purposes other than those specified in this subdivision shall provide notice to consumers, pursuant to subdivision (a) of Section 1798.135, that this information may be used or disclosed to a service provider or contractor, for additional, specified purposes and that consumers have the right to limit the use or disclosure of their sensitive personal information

(b) **A business that has received direction from a consumer not to use or disclose the consumer's sensitive personal information**, except as authorized by subdivision (a), shall be prohibited, pursuant to paragraph (19) of subdivision (c) of Section 1798.135, from using or disclosing the consumer's sensitive personal information for any other purpose after its receipt of the consumer's direction unless the consumer subsequently provides consent for the use or disclosure of the consumer's sensitive personal information for additional purposes

(c) **A service provider or contractor that assists a business in performing the purposes authorized by subdivision (a) may not use the sensitive personal information after it has received instructions from the business and to the extent, it has actual knowledge that the personal information is sensitive personal information for any other purpose.** A service provider or contractor is only required to limit its use of sensitive personal information received pursuant to a written contract with the business in response to instructions from the business and only with respect to its relationship with that business

(d) **Sensitive personal information that is collected or processed without the purpose of inferring characteristics about a consumer is not subject to this section**, as further defined in regulations adopted pursuant to subparagraph (C) of paragraph (19) of subdivision (a) of Section 1798.185, and shall be treated as personal information for purposes of all other sections of this act, including Section 1798.100

establishing a pay-for-privacy precedent that discriminates against lower socioeconomic groups.

CONCLUSION

Data aggregation is a necessary yet time- and technology-intensive task in making health systems safer, more effective, and less expensive by analyzing hospital data in EHRs. Third-party data aggregation companies are increasingly filling unmet needs in this regard but complicate the data protection landscape where health and consumer data protection could simultaneously apply in a growing market for hospital data. The current opinion presents an ethical comparison of these protections outlined in the CA Privacy Act and Federal Privacy Rule from four primary perspectives: 1) standards of data deidentification; 2) rights of patients and consumers in relation to their health data; 3) entities covered by the acts; 4) scopes of regulation.

The first version of the CA Privacy Act introduced landmark consumer privacy legislation in 2018. It applied to certain businesses and data brokers that met certain revenue (more than $15M) and data processing (more than 50,000 consumers, households, or devices) criteria. Yet, businesses and data brokers were able to circumvent some restrictions on the "sale" of information, for example, and imposed the same requirements on all categories of personal information irrespective of differences in sensitivity. Patients and consumers about whom health information, in particular, was systematically collected and sold were disadvantaged given the heightened sensitivity of this information and ease with which it could be readily linked with other public sources.

Consumer privacy rights can be triggered when a TDAC is not contractually designated as a business associate with covered entity aggregates health information that can be "reasonably" identifiable. Moreover, the CA Privacy Act protections could apply to hospital data sourced from privately owned and operated hospitals and sold to other businesses, entities, or data brokers subject to the CA Privacy Act. The expanded protection for health information fills a regulatory gap left open by the Federal Privacy Rule and, as a result, strengthens protection for patients treated at for-profit hospitals and consumers of health-related services such as direct-to-consumer genetic testing.

When TDACs operate as a business associate of a covered entity, patients could exercise their request for an accounting of disclosures for nonhealthcare operation purposes to better understand with whom their protected health information has been shared. Enhanced representation from patient groups in business associate negotiations is one approach to establishing a more equitable benefit-sharing structure that prioritizes patient care and financing of patient-led programs from revenues received from a partnership with TDACs. Further empirical research is needed to understand what, if any, patient privacy and other ethical interests should be factored into decisions to partner with third-party aggregation companies from the perspectives of patients and hospital administrators.

AUTHOR CONTRIBUTIONS

The author confirms being the sole contributor of this work and has approved it for publication.

FUNDING

Support for this work was provided by the Pediatric Extramural Loan Repayment Program through the National Human Genome Research Institute (NHGRI) and T32 NHGRI training program (5T32HG008953) in the Ethical, Legal, and Social Implications of Genetics and Genomics.

SUPPLEMENTARY MATERIAL

The Supplementary Material for this article can be found online at: https://www.frontiersin.org/articles/10.3389/fdata.2020.603044/full#supplementary-material.

REFERENCES

21st Century Cures Act (2016). H.R. 34, 114th Congress. 2016.

Béranger, J. (2016). *Big data and ethics: the medical datasphere*. London, United Kingdom: ISTE Press and Elsevier Ltd.

Challenge.gov. Consumer health data aggregator challenge [Internet]. Available at: https://www.challenge.gov/challenge/consumer-health-data-aggregator-challenge/ (Accessed January 26, 2021).

Char, D., Shah, N. H., and Magnus, D. (2018). Implementing machine learning in health care – addressing ethical challenges. *N. Engl. J. Med.* 378 (11), 981–983. doi:10.1056/NEJMp1714229

Department of Health and Human Services (2003). Summary of the privacy rule.

Department of Health and Human Services (2008). What may a HIPAA covered entity's business associate agreement authorize a health information organization (HIO) to do with electronic protected health information (PHI) it maintains or has access to in the network?.

Fefferman, N. H., O'Neil, E. A., and Naumova, E. N. (2005). Confidentiality and confidence: is data aggregation a means to achieve both? *J. Publ. Health Pol.* 26 (4), 430–449. doi:10.1057/palgrave.jphp.3200029

Groves, P., Kayyali, B., Knott, D., and Van Kuiken, S. (2013). The "big data" revolution in healthcare: accelerating value and innovation [Internet]. McKinsey and Company. Available at: http://www.euro.who.int/__data/assets/pdf_file/0004/287275/EHII_Booklet_EN_rev1.pdf?ua=1%5Cnhttp://www.euro.who.int/__data/assets/pdf_file/0010/96463/E93556.pdf%5Cnhttp://wma.comb.es/Upload/Documents/Mayer_MundoInternet07_39.pdf%5Cnhttp://www.images-et-re.

Homer, N., Szelinger, S., Redman, M., Duggan, D., Tembe, W., Muehling, J., et al. (2008). Resolving individuals contributing trace amounts of DNA to highly complex mixtures using high-density SNP genotyping microarrays. *PLoS Genet.* 4 (8), e1000167. doi:10.1371/journal.pgen.1000167

Kulynych, J., and Greely, H. T. (2017). Clinical genomics, big data, and electronic medical records: reconciling patient rights with research when privacy and science collide. *J. Law Biosci.* 4 (1), 94–132. doi:10.1093/jlb/lsw061

National Conference of State Legislatures. State laws related to internet privacy [Internet]. Available at: https://www.ncsl.org/research/telecommunications-and-information-technology/state-laws-related-to-internet-privacy.aspx.

Noordyke, M. (2020). "US state comprehensive privacy law comparison," in *International association of privacy professionals*.

L. A. Olsen, D. Aisner, and J. M. E. McGinnis Editors (2007). *The learning healthcare system: workshop summary*. Washington, DC: Institute of Medicine: Roundtable on Evidence-Based Medicine.

Price, W. N., Kaminski, M. E., Minssen, T., and Spector-Bagdady, K. (2019). Shadow health records meet new data privacy laws. *Science* 363 (6426), 448–450. 10.1126/science.aav5133

Price, W. N. (2018). Risk and resilience in health data infrastructure. *SSRN Electron J.* 1 (1), 65–85. doi:10.2139/ssrn.2928997

Rolnick, J. (2013). Aggregate health data in the United States: steps toward a public good. *Health Inf. J.* 19 (2), 137–151. doi:10.1177/1460458212462077

Stalla-Bourdillon, S., and Wu, D. (2019). *What we're missing in the CCPA de-identification debate*. The Hill. Available at: https://thehill.com/opinion/cybersecurity/473652-what-were-missing-in-the-ccpa-de-identification-debate (Accessed January 26, 2021)

State of California (2018). California consumer privacy act (Senate Bill No. 1121) [Internet]. Available at: https://leginfo.legislature.ca.gov/faces/codes_displayText.xhtml?lawCode=CIV&division=3.&title=1.81.5.&part=4.&chapter=&article=.

United States Congress (2009). Health information technology (HITECH Act). Index excerpts from Am recover reinvestment act 2009 [Internet], 112–164. Available at: https://www.healthit.gov/sites/default/files/hitech_act_excerpt_from_arra_with_index.pdf.

Wang, S., Jiang, X., Singh, S., Marmor, R., Bonomi, L., Fox, D., et al. (2017). Genome privacy: challenges, technical approaches to mitigate risk, and ethical considerations in the United States. *Ann. N. Y. Acad. Sci.* 1387 (1), 73–83. doi:10.1111/nyas.13259

Conflict of Interest: The author declares 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 Neuroinformatics

OPINION
published: 08 March 2021
doi: 10.3389/fninf.2021.567920

Self-Regulation in the Time of Lockdown

*Anita Pacholik-Żuromska**

Department of Cognitive Science, Institute of Information and Communication Research, Nicolaus Copernicus University in Toruń, Toruń, Poland

Keywords: outbreak, self-regulation, self, embodiment, enactivism, digital training tools

INTRODUCTION

Since the beginning of the coronavirus disease 2019 pandemic, many studies have reported the psychological impact of the lockdown (mass quarantine). According to these studies, the societies affected are exposed to increased stress, mental tension, anxiety, and depression (Lee, 2020; Ozamiz-Etxebarria et al., 2020; Pandey et al., 2020; Sinnghal and Vijayaraghavan, 2020; Sugaya et al., 2020; Yuan et al., 2020). This paper aims to consider the possibility of minimizing the psychological effects resulting from the extraordinary situation caused by the outbreaks. I will propose an answer to the question of what methods will be effective in dealing with the negative psychological effects of lockdown and whether technological progress can benefit us in any way. In my opinion, effective results can be brought about by self-regulation methods based on biofeedback because they make it possible to develop the awareness of one's own body, reduce the feeling of detachment, and thus regain self-control (cf. Goessl et al., 2017). These methods are a good example of how the body affects the mind. The digital tools providing biofeedback are easy to use, so even people who are distrustful of digitalization can be convinced of their usefulness.

LOSS OF CONTROL

The emergence of abrupt emotional reactions during the outbreak of a pandemic is a natural phenomenon. It is an outcome of evolutionary bio-behavioral development (Tops et al., 2014). To survive in a given environment, an organism must adapt by minimizing the error in the prediction of the possible state of the world (Friston, 2009, 2012). The adaptation then would be seen as a disposition toward the avoidance of an informational surprise emerging from the environment (Friston, 2009, 2012). The lower the entropy, the higher the predictability of the environment. Although the prediction error is a necessary element of learning (Joiner et al., 2017), if it occurs too often, a low, predictable environment is created, which causes reactive behavior involving strong emotions as an adaptive response (Tops et al., 2014): acting according to the direct stimuli, the involvement of exogenous attention, associative learning (Tops et al., 2014), and minimal rationality, indicating an action that is of maximal use for an agent, regardless of the consequences for others (Cherniak, 1981). On the other hand, a highly predictable environment causes proactive behavior, which is less affective and allows rational coping and self-reflection (Tops et al., 2014).

In my opinion, outbreaks and their social consequences create an abnormal situation, indicating a low, predictable environment in which, for a long time, individuals cannot function properly. Everyday reference points, such as actions and undertakings, owing to which we maintain balance and which define us and keep us in line, are fuzzy and distorted. What constitutes the ground for such negative psychological reactions is the disintegration of the self, which occurs due to the loss of the locus of control.

The evidence of this claim can be found, for example, in studies on disorders of the self, such as autism spectrum disorder and schizophrenia. The locus of control, which means the self, is here disturbed by either a too weak or too strong "bodily boundary between self and other" (Noel et al., 2017, p. 1). In these cases, the therapies changing the self-representation by altering the experience

OPEN ACCESS

Edited by:
Koichi Fujiwara,
Nagoya University, Japan

Reviewed by:
Jakub Limanowski,
Technische Universität
Dresden, Germany

****Correspondence:***
Anita Pacholik-Żuromska
apz@umk.pl

Received: *31 May 2020*
Accepted: *09 February 2021*
Published: *08 March 2021*

Citation:
Pacholik-Żuromska A (2021)
Self-Regulation in the Time of
Lockdown.
Front. Neuroinform. 15:567920.
doi: 10.3389/fninf.2021.567920

of self-location by using "synchronous administration of spatially fixed exteroceptive (i.e., visual and/or auditory) sensory signals with tactile information" were proposed (Noel et al., 2017, p. 8). Such therapies, in my opinion, fall under the category of therapies with biofeedback, which is what I mean here. Thus, not only in the time of lockdown but also in other situations, there is a need for a professional tool to help restore a sense of control and bring it back to the self through the body.

SELF-REGULATION AND BIOFEEDBACK

One of the ideas of how to deal with a low, predictable environment is to build resilience while replacing reactive behavior with proactive behavior (Dehnad, 2017) to respond less emotionally and more confrontationally to an abnormal situation. I see such potential in a digital training of self-regulation with biofeedback.

Self-regulation has many dimensions. It can be understood in a physiological sense as the maintenance of homeostasis and "compensatory responses to the discrepancy between a system's reference state and its input stage" (Jeannerod, 1993, p. 83, see also Jeannerod, 2006). In a psychological sense, self-regulation refers to emotions and suggests "initiation or alteration of ongoing emotional responses through cognitive processes" (Heatherton, 2011, p. 375). Additionally, society fulfills a regulative role when a subject compares his or her perspective with the perspective of others and modifies his or her attitudes to conform to the norms (Tomasello, 1993, 2019). In terms of biofeedback, self-regulation is understood as a result of combining the biological and psychological dimensions. In such an account, self-regulation increases the sense of control over one's own body by engaging in proprioceptive and attentional training (Cf. Blanke and Metzinger, 2009). In general, biofeedback indicates information about the state and condition of one's own body provided by various measurement tools (Schwartz and Andrasik, 2017). The purpose of biofeedback is to enhance bodily awareness to consciously and intentionally regulate physical states such as muscle tension or heart rate. To achieve this ability, the psychophysical (internal–external) coupling in self-representation should be strengthened.

In the understanding of self-regulation proposed here, it is clear that enhancing the connection with one's own body will lead to the restoration of self-control. Therefore, there is a need for innovative digital tools offering modern psychological healthcare to regulate the embodied and enactive self and thereby increase psychological resilience and sustain mind–body balance.

DIGITALIZATION OF SELF-REGULATION

There are already many studies on using digital technologies to train various social competencies (Gaggioli et al., 2019). Mobile applications dedicated to self-regulation often refer to the elements of mindfulness, such as focused attention and hand–eye coordination (Tang et al., 2007). These tools are quite interesting because they provide higher motivation for training owing to their gamification. The engagement of attention and the senses of vision and touch regulate proprioception (Gibson, 2002), although the influence on proprioception by mobile applications is far smaller than by virtual reality (VR) (Lenggenhager et al., 2007; Blanke et al., 2015). The mobile apps have other disadvantages because they do not give biofeedback and strain visual attention owing to the need to focus on the display. They are also sometimes immensely complicated for people less familiar with new technologies. More importantly, in a demanding situation, such as a pandemic, such people will not use tools that additionally stress them. Therefore, there is a need to develop devices that help in self-regulation in a low, predictable environment and offer training support to cope with stress in situations that go beyond ordinary everyday life.

The idea is then to adapt digital tools to the specific conditions governing the limitations of daily activities. Such digital devices for self-regulation training should involve movement and give biofeedback. These applications do not seem possible on one-piece hardware such as a mobile phone. Rather, there is a need for separate sensors and a device for collecting data and giving feedback on this basis. These sensors could also be the elements of training. For example, they could be interactive marbles to hold on open hands.[1] One of the training tasks would be to perform body movements to keep balance and not drop the marbles. They could also be interactive rings on fingers. There are plenty of possibilities. The point is that the sensor placed on body parts, such as arms, fingers, and legs, would collect the data from, for example, electrodermal activity, muscle tension, and pulse and send them to the control unit in a separate device. This device could be similar to the usual fitness watch, although with a voice guide (a kind of assistant) giving instructions according to the data to adjust the user movement. Important here is the calibration and personalization of the device. The advantage of such a tool would be that it requires movement and, at the same time, gives feedback without having to look at the display, and thereby it does not strain the sense of vision.

The constant need to create and develop such digital tools was recognized by Colombo et al. (2019), especially with reference to emotion regulation. The authors extensively reviewed the technologies from the internet-based interventions *via* mobile health to virtual reality, which they perceived as good support for self-regulation. Among them, they also placed biofeedback. They then diversified the biofeedback from other forms and also proposed mixed forms, for example, VR with biofeedback (Colombo et al., 2019). I can only support this idea; although I see the limits of its use by older generations, first because of their possible aversion to new technologies and second because of an increased probability of motion sickness (Lee et al., 2017). At this juncture, VR also faces some substantial limits in user experience, such as a heavy head-mounted display or wiring.

[1] In the laboratory belonging to the Department of Cognitive Science, a team of scientists, together with Neurodio LLC, has created a relaxing application *Stabilo*, in which the idea of a balancing marble was also used. https://play.google.com/store/apps/details?id=com.Neurodio.Stabilo. The aim of the application is to train mindfulness as a method of coping with stress. From my perspective, what we learn from *Stabilo* first is patience because, at the beginning, it is extremely hard to balance with the virtual marble, but once we manage it, we can easily deal with any other kinds of stress.

Nevertheless, VR offers a variety of environments where, owing to strong immersion, a subject feels like he/she is in a real situation, giving many possibilities of training by action. It is also worth mentioning another reason for the combination of gamification and biofeedback—the necessity of self-regulation during or even before playing (Seay and Kraut, 2007). My argument for this need is that sometimes subjects are not able to recognize that they need self-regulation because bodily alarming signals do not reach the field of consciousness or subjects cannot identify emotions, which are too fuzzy (Schooler and Schreiber, 2004). This happens, in my opinion, in those situations that cause reactive behavior, lacking self-reflection, and, hence, self-regulation. Under such circumstances, the use of gamified tools in the hope of reacting to stress can have even negative effects such as addiction (Seay and Kraut, 2007). In such deficits, the biofeedback delivers information about the physical state of the body—which correlates with the psychological state—before the subject realizes (if at all) that their current emotions are overwhelming them.

CONCLUSION

It can be said that the coronavirus disease 2019 pandemic situation has forced us to speed up the progress in digitalization. There is a sudden need to create tools that help in self-regulation and thus support coping with anxiety caused by the unusual stressor—the outbreak. At the same time, and unlike other digital tools in healthcare, it is difficult to define to whom such devices would be dedicated because the target is highly differentiated according to factors, including age, sex, and character—that is why such tools need to be easily calibrated and personalized. Furthermore, as has been emphasized, digital training in self-regulation cannot be an additional psychological burden, but on the contrary, it must release positive emotions; otherwise, no one will use it more than once. Finally, from a philosophical viewpoint, such tools are an example of "the extended mind" (Clark and Chalmers, 1998) as they improve self-reflection and help in self-cognition when it starts to fail in a stressful situation and when reactive behavior is triggered. Thus, they allow one to regain self-control and rebuild the connection to oneself.

AUTHOR'S NOTE

AP-Z refers in the manuscript to the mobile application Stabilo, created by Neurodio LLC and Kogni_LAB. The Lab is a unit of the Department of Cognitive Science chaired by AP-Z. AP-Z was also an initiator of the project Stabilo.

AUTHOR CONTRIBUTIONS

AP-Z confirms sole responsibility for the conception, study, and manuscript preparation.

FUNDING

This publication was supported by the Emerging Field Logic and Philosophy of Science in the frame of Excellence Initiative – Research University Programme.

REFERENCES

Blanke, O., and Metzinger, T. (2009). Full-body illusions and minimal phenomenal selfhood. *Trends Cogn. Sci.* 13, 8–13. doi: 10.1016/j.tics.2008.10.003

Blanke, O., Slater, M., and Serino, A. (2015). Behavioral, neural, and computational principles of bodily self-consciousness. *Neuron* 88, 145–166. doi: 10.1016/j.neuron.2015.09.029

Cherniak, Ch. (1981). Minimal rationality. *Mind* 90, 161–183. doi: 10.1093/mind/XC.358.161

Clark, A., and Chalmers, D. (1998). The extended mind. *Analysis* 58, 7–19. doi: 10.1093/analys/58.1.7

Colombo, D., Fernández-Álvarez, J., García Palacios, A., Cipresso, P., Botella, C., and Riva, G. (2019). New technologies for the understanding, assessment, and intervention of emotion regulation. *Front. Psychol.* 10:1261. doi: 10.3389/fpsyg.2019.01261

Dehnad, V. (2017). A proactive model to control reactive behaviors. *World J. Educ.* 7. doi: 10.5430/wje.v7n4p24

Friston, K. (2009). The free-energy principle: a rough guide to the brain?. *Trends Cogn. Sci.* 13, 293–301. doi: 10.1016/j.tics.2009.04.005

Friston, K. (2012). A free energy principle for biological systems. *Entropy* 14, 2100–2121. doi: 10.3390/e14112100

Gaggioli, A., Villani, D., Serino, S., Banos, R., and Botella, C. (eds.). (2019). Positive technology: designing E-experiences for positive change. *Front. Psychol.* 10:1571. doi: 10.3389/fpsyg.2019.01571

Gibson, J. J. (2002). "A theory of direct visual perception," in *Vision and Mind. Selected Readings in the Philosophy of Perception*, eds A. Noë and E. Thompson (London: MIT Press), 77–91.

Goessl, V. C., Curtiss, J. E., and Hofmann, S. G. (2017). The effect of heart rate variability biofeedback training on stress and anxiety: a meta-analysis. *Psychol. Med.* 47, 2578–2586. doi: 10.1017/S0033291717001003

Heatherton, T. F. (2011). Neuroscience of self and self-regulation. *Annu. Rev. Psychol.* 62, 363–390. doi: 10.1146/annurev.psych.121208.131616

Jeannerod, M. (1993). "Theory of representation-driven actions," in *The Perceived Self: Ecological and Interpersonal Sources of Self-Knowledge*, ed U. Neisser (Cambridge: Cambridge University Press), 68–88.

Jeannerod, M. (2006). *Motor Cognition: What Actions Tell the Self*. Oxford: Oxford University Press.

Joiner, J., Piva, M., Turrin, C., and Chang, S. W. C. (2017). Social learning through prediction error in the brain. *npj Sci. Learn.* 2:8. doi: 10.1038/s41539-017-0009-2

Lee, J., Kim, M., and Kim, J. (2017). A study on immersion and VR sickness in walking interaction for immersive virtual reality applications. *Symmetry* 9:78. doi: 10.3390/sym9050078

Lee, S. A. (2020). Coronavirus anxiety scale: a brief mental health screener for COVID-19 related anxiety. *Death Stud.* 44:7. doi: 10.1080/07481187.2020.1748481

Lenggenhager, B., Tadji, T., Metzinger, T., and Blanke, O. (2007). Video ergo sum: manipulating bodily self-consciousness. *Science* 317, 1096–1099. doi: 10.1126/science.1143439

Noel, J. P., Cascio, C. J., Wallace, M. T., and Park, S. (2017). The spatial self in schizophrenia and autism spectrum disorder. *Schizophr. Res.* 179, 8–12. doi: 10.1016/j.schres.2016.09.021

Ozamiz-Etxebarria, N., Idoiaga Mondragon, N., Dosil, S. M., and Picaza, G. M. (2020). Psychological symptoms during the two stages of lockdown in response to the COVID-19 outbreak: an investigation in a sample of citizens in northern spain. *Front. Psychol.* 11:1491. doi: 10.3389/fpsyg.2020.01491

Pandey, D., Bansal, S., Goyal, S., Garg, A., Sethi, N., Pothiyill, D. I., et al. (2020). Psychological impact of mass quarantine on population during pandemics—the COVID-19 Lock-Down (COLD) study. *PLoS ONE* 15:e0240501. doi: 10.1371/journal.pone.0240501

Schooler, J. W., and Schreiber, C. A. (2004). Experience, meta-consciousness and the paradox of introspection. *J. Consciousness Stud.* 11, 17–39.

Schwartz M. S., and Andrasik, F. (eds.). (2017). *Biofeedback. A Practitioner's Guide, 4th Edn*. New York, NY: Guilford Press.

Seay, F. A., and Kraut, R. E. (2007). "Project massive: self-regulation and problematic use of online gaming," in *CHI 2007 Proceedings, Games, April 28–May 3, 2007* (San Jose, CA).

Sinnghal, D., and Vijayaraghavan, P. (2020). A descriptive study of Indian general public's psychological responses during COVID-19 pandemic lockdown period in India. *PsyArXiv. [Preprint]*. doi: 10.31234/osf.io/jeksn

Sugaya, N., Yamamoto, T., Suzuki, N., and Uchiumi, C. (2020). A real-time survey on the psychological impact of mild lockdown for COVID-19 in the Japanese population. *Sci Data* 7:372. doi: 10.1038/s41597-020-00714-9

Tang, Y. Y., Ma, Y., Wang, J., Fan, Y., Feng, S., Lu, Q., et al. (2007). Short-term meditation training improves attention and self-regulation. *PNAS* 104, 17152–17156. doi: 10.1073/pnas.0707678104

Tomasello, M. (1993). "On the interpersonal origins of self-concept," in *The Perceived Self: Ecological and Interpersonal Sources of Self-Knowledge*, ed U. Neisser (Cambridge: Cambridge University Press), 174–184.

Tomasello, M. (2019). *Becoming Human: A Theory of Ontogeny*. Harvard University Press. doi: 10.4159/9780674988651

Tops, M., Luu, P., Boksem, M. A. S., and Tucker, D. M. (2014). "The roles of predictive and reactive biobehavioral programs in resilience," in *The Resilience Handbook: Approaches to Stress and Trauma*, eds M. Kent, M. C. Davis, and J. W. Reich (New York, NY: Routledge/Taylor & Francis Group), 15–32.

Yuan, S., Liao, Z., Huang, H., Jiang, B., Zhang, X., Wang, Y., et al. (2020). Comparison of the indicators of psychological stress in the population of hubei province and non-endemic provinces in china during two weeks during the coronavirus disease 2019 (COVID-19) outbreak in February 2020. *Med. Sci. Monit.* 26:e923767. doi: 10.12659/MSM.923767

Conflict of Interest: The author declares 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 Public Health

ORIGINAL RESEARCH
published: 10 March 2021
doi: 10.3389/fpubh.2021.630640

Work Habit-Related Sleep Debt; Insights From Factor Identification Analysis of Actigraphy Data

Yuki Goto[1], *Koichi Fujiwara*[2*†], *Yukiyoshi Sumi*[3†], *Masahiro Matsuo*[3†], *Manabu Kano*[1†] and *Hiroshi Kadotani*[4†]

[1] Department of Systems Science, Kyoto University, Kyoto, Japan, [2] Department of Material Process Engineering, Nagoya University, Nagoya, Japan, [3] Department of Psychiatry, Shiga University of Medical Science, Otsu, Japan, [4] Department of Sleep and Behavioural Sciences, Shiga University of Medical Science, Otsu, Japan

The present study investigates the factors of "Weekday sleep debt (WSD)" by comparing activity data collected from persons with and without WSD. Since it has been reported that the amount of sleep debt as well the difference between the social clock and the biological clock is associated with WSD, specifying the factors of WSD other than chronotype may contribute to sleep debt prevention. We recruited 324 healthy male employees working at the same company and collected their 1-week wrist actigraphy data and answers to questionnaires. Because 106 participants were excluded due to measurement failure of the actigraphy data, the remaining 218 participants were included in the analysis. All participants were classified into WSD or non-WSD groups, in which persons had WDS if the difference between their weekend sleep duration and the mean weekday sleep duration was more than 120 min. We evaluated multiple measurements derived from the collected actigraphy data and trained a classifier that predicts the presence of WSD using these measurements. A support vector machine (SVM) was adopted as the classifier. In addition, to evaluate the contribution of each indicator to WSD, permutation feature importance was calculated based on the trained classifier. Our analysis results showed significant importance of the following three out of the tested 32 factors: (1) WSD was significantly related to persons with evening tendency. (2) Daily activity rhythms and sleep were less stable in the WSD group than in the non-WSD group. (3) A specific day of the week had the highest importance in our data, suggesting that work habit contributes to WSD. These findings indicate some WSD factors: evening chronotype, instability of the daily activity rhythm, and differences in work habits on the specific day of the week. Thus, it is necessary to evaluate the rhythms of diurnal activities as well as sleep conditions to identify the WSD factors. In particular, the diurnal activity rhythm influences WSD. It is suggested that proper management of activity rhythm may contribute to the prevention of sleep debt.

Keywords: weekday sleep debt, actigraphy, machine learning, feature importance, support vector machine

OPEN ACCESS

Edited by:
Xi Long,
Eindhoven University of Technology, Netherlands

Reviewed by:
Sylvia Hach,
Unitec Institute of Technology, New Zealand
Eliana Roveda,
University of Milan, Italy

****Correspondence:***
Koichi Fujiwara
fujiwara.koichi@hps.material.nagoya-u.ac.jp

†ORCID:
Koichi Fujiwara
orcid.org/0000-0002-2929-0561
Yukiyoshi Sumi
orcid.org/0000-0001-6775-0883
Masahiro Matsuo
orcid.org/0000-0003-3283-3413
Manabu Kano
orcid.org/0000-0002-2325-1043
Hiroshi Kadotani
orcid.org/0000-0001-7474-3315

Specialty section:
This article was submitted to Digital Public Health, a section of the journal Frontiers in Public Health

Received: *18 November 2020*
Accepted: *12 February 2021*
Published: *10 March 2021*

Citation:
Goto Y, Fujiwara K, Sumi Y, Matsuo M, Kano M and Kadotani H (2021) Work Habit-Related Sleep Debt; Insights From Factor Identification Analysis of Actigraphy Data. Front. Public Health 9:630640. doi: 10.3389/fpubh.2021.630640

INTRODUCTION

Sleep debt has deleterious effects on work or academic performance and also may impair various other psychological and physical functions such as memory, learning, metabolism, and immunity (1). However, a convenient method for quantitively evaluating sleep debt has not yet been established. According to the International Classification of Sleep Disorders, Third Edition (ICSD-3), getting enough sleep duration for at least 7 days before polysomnography (PSG) should be performed for the sake of sleep debt resolution (2); however, inventories to measure the degree of sleep debt have not been developed.

We focus on "weekday sleep debt (WSD)," which refers to taking longer sleep on the weekend to compensate for lack of sleep during the weekdays (3, 4). WSD is defined as the difference between the weekend and weekday sleep durations (5).

It has been well-known that chronotype affects weekday and weekend sleep timing and duration (6). One cause of WSD may be social jetlag, which is the discrepancy between biological and social clocks (7, 8). WSD may not occur when persons live following their born chronotype determined by genetic factors. However, it is difficult for most persons to live a daily social life by following their chronotype due to work or school. The factors of WSD other than chronotype should be specified to prevent sleep debt because the chronotype are reported to be influenced by genetic factors (5) as well as age, and personality (9).

It is assumed that persons with WSD may not get enough sleep during the weekdays and take longer sleeps on the weekend than during the weekdays to compensate for lack of weekday sleep. Actigraph devices have been widely used for long-term circadian activity measurement because they are lightweight, easy-to-wear, and non-invasive (10–12). The long-term activity measurement enables the evaluation of the discrepancy between biological and social rhythms clocks.

In this study, we aim to investigate the factors of WSD by using 1-week wrist actigraphy data. The actigraphy data and answers to questionnaires of persons with and without WSD were collected from 324 male employees at a Japanese wholesale company, and 218 participants were analyzed.

Before analysis, we validated the collected actigraphy data through comparison with sleep diaries recorded by the participants. In addition, the definition of WSD was examined with the chorotype of the participants determined by the Morningness-Eveningness Questionnaire (MEQ).

We trained a classifier that predicts the presence of WSD from the collected actigraphy data by utilizing a machine learning (ML) technique and calculated feature importance based on the trained classifier to specify the factors of WSD in addition to statistical analysis of the answers to questionnaires.

MATERIALS AND METHODS

Measurements From Actigraphy Data

We calculated the following 40 parameters listed in **Table 1** from the actigraphy data, whose details are described below.

- Sleep/Awake State: Wake − up time (WU), Sleep-onset time (SO), Mid − sleep (MS), and Sleep duration (SD)

TABLE 1 | Measurements from actigraphy data.

Explanatory variable	Abbreviations
Wake-up time	WU_{Tue}, WU_{Wed}, WU_{Thu}, WU_{Fri}, WU_{mean}, WU_{std}
Sleep-onset time	SO_{Mon}, SO_{Tue}, SO_{Wed}, SO_{Thu}, SO_{mean}, SO_{std}
Mid-sleep	MS_{Mon}, MS_{Tue}, MS_{Wed}, MS_{Thu}, MS_{mean}, MS_{std}
Sleep duration	SD_{Mon}, SD_{Tue}, SD_{Wed}, SD_{Thu}, SD_{mean}, SD_{std}
Average activity during the most active 10-h period	$M10_{Tue}$, $M10_{Wed}$, $M10_{Thu}$, $M10_{Fri}$
Average activity during the least active 5-h period	$L5_{Tue}$, $L5_{Wed}$, $L5_{Thu}$, $L5_{Fri}$
Sleep regularity index	SRI
Sleep timing index	STI
Inter-daily stability	IS_{act}, IS_{light}
Intra-daily variability	IV_{act}, IV_{light}
Mid-sleep on free days	MSF
Social jetlag	SJL

The Cole-Kripke sleep/wake identification method was adopted to estimate wake-up and sleep-onset times from the actigraphy data (13). It has been reported that the agreement rate between the Cole-Kripke method and PSG was 88% (13, 14).

It is difficult to estimate sleep/awake timing using only the Cole-Kripke method because it uses 1-min activity counts data measured by means of actigraphy; therefore, we introduce a sleep/wake function $r(i)$

$$r\left(i\right) = \sum_{j=0}^{i} x_j \tag{1}$$

where x_i is a sleep/wake state at the time i estimated by means of the Cole-Kripke method: $x_i = +1$ and $x_i = -1$ denote sleep and awake, respectively. $r(i)$ becomes large when the sleep state dominates and becomes small when the wake state dominates. The extreme points of $r(i)$ may be times that the sleep/awake states change. Thus, its maximum and minimum points can be defined as the wake-up time and the sleeping time. If there are multiple maximum and minimum points within a day, we adopt the latest one as the wake-up time or the sleeping time.

Using the sleep/wake function $r(i)$, the following measurements, which represent the sleep habits, are estimated: wake-up time (WU), sleep-onset time (SO), mid-sleep (MS), and sleep duration (SD). WU, SO and MS are expressed as the minutes elapsed from midnight. Their means and standard deviations were used in addition to these values for each weekday.

- Sleep Rhythm: Sleep regularity index (SRI) and Sleep timing index (STI)

Sleep regularity index (SRI) (15) and sleep timing index (STI) (16) represent sleep rhythm. SRI is the agreement between sleep/wake states at any two-time points separated by 24 h. A large SRI means that the sleep rhythm is regular. SRI can be calculated based on the Cole-Kripke sleep/wake identification method. N and p are the numbers of measurement days and samples per day, respectively. $s_{k,j} = 1$ is satisfied if a person is determined to be in the sleep state at time j on the kth day, and $s_{k,j} = 0$ is satisfied if a person is determined to be in the awake state at time j on the kth day. SRI is defined as

$$\mathrm{SRI} = -100 + \frac{200}{p\,(N-1)} \sum_{j=1}^{p} \sum_{k=1}^{N-1} \delta_{s_{k,j},s_{k+1,j}} \tag{2}$$

where $\delta_{i,j} = 1$ if $i = j$ and $\delta_{i,j} = 0$ otherwise.

STI expresses the mean sleep midpoint during the measurement days. When the time from 0:00 to 24:00 are associated with the angles $\theta = [0, 2\pi]$, a time point j is expressed as $\theta_j = 2\pi j/p$. Using this relationship, STI is defined as

$$\Theta = \arg\left(\sum_{j=1}^{p} \sum_{k=1}^{N} s_{k,j} e^{i\theta_j}\right) \tag{3}$$

$$\mathrm{STI} = \frac{60 \times 24}{2\pi}\Theta \tag{4}$$

where i is the imaginary unit, and Θ is the argument of the sum of $e^{i\theta_j}$ over the sleep states. **Figure 1** shows a schematic diagram of STI, in which the triangles denote the times estimated as the sleep states, and $e^{i\theta_j}$ is the composition of the vectors directing these triangles. Thus, the time corresponding to the argument of $e^{i\theta_j}$ is STI.

- Activity Rhythm: Inter-daily stability (IS) and Intra-daily variability (IV)

Inter-daily stability (IS) and intra-daily variability (IV) are well-known parameters for activity rhythm evaluation based on the actigraphy data (17), which are defined as follows:

$$\mathrm{IS} = \frac{M \sum_{j=1}^{p} \left(\overline{A}_j - \overline{A}\right)^2}{p \sum_{i=1}^{M} \left(A_i - \overline{A}\right)^2} \tag{5}$$

$$\mathrm{IV} = \frac{M \sum_{i=2}^{M} \left(A_i - A_{i-1}\right)^2}{(M-1) \sum_{i=1}^{M} \left(A_i - \overline{A}\right)^2} \tag{6}$$

where A_i is the activity data at time i, $\overline{A}_j$ is the average of the activity data at time j over different days, and $\overline{A}$ is the average of all activity data. p is the number of activity data per day, and M is the total number of collected activity data. IS is the ratio of the variance of all the collected activity data to the variance of the average activity data at time j over different days. IS becomes large when the fluctuation of the diurnal activity data recorded at the same time of different days becomes large. IV is the ratio of the variance of the collected activity data to the sum of squares of the differences in activity data between a time point and the previous time point. IS and IV are measured for the synchronization of diurnal activity rhythm and the frequency of midday naps and nocturnal awakening, respectively (18).

IS and IV can be calculated from both the activity counts and illuminance in the actigraphy data, which are denoted as $\mathrm{IS}_{\mathrm{act}}$, $\mathrm{IS}_{\mathrm{light}}$, $\mathrm{IV}_{\mathrm{act}}$, $\mathrm{IV}_{\mathrm{light}}$, respectively.

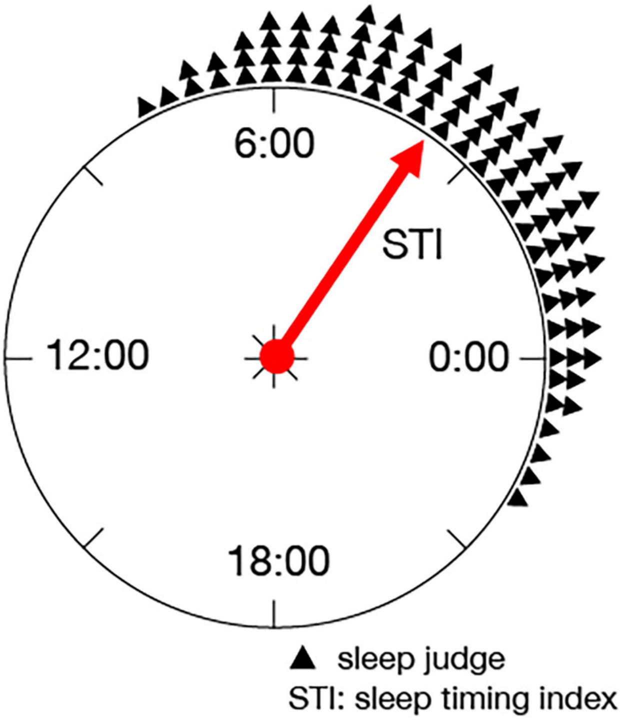


FIGURE 1 | Schematic diagram of STI: STI expresses the mean sleep midpoint during the measurement days as the argument on a unit circle.

- Rest–Activity Rhythm: Average activity during the most active 10-h period (M10) and Average activity during the least active 5-h period (L5)

Although IV represents intra-daily activity rhythm, M10 and L5 can be calculated from the actigraphy data to emphasize the rest-activity circadian rhythm. M10 and L5 are defined as the average activities during the most active 10-h period and during the least active 5-h period, respectively (17). M10 represents activity during the most active period within the day, which may be influenced by daytime napping. L5 represents movement activity during sleep.

- Biological Clock: Mid-sleep on free days (MSF) and Social jetlag (SJL)

Mid-sleep on free days (MSF) is a measurement of chronotype. Wake-up time on free days (WUF) and sleep-onset time on free days (SOF) are wake-up time and sleep onset time on the weekend, respectively. MSF is expressed as

$$\mathrm{MSF} = \frac{\mathrm{SOF} + \mathrm{WUF}}{2} \tag{7}$$

Although sleep during the weekdays does not reflect the congenital biological clock much due to the effect of the social clock, weekend sleep significantly reflects the congenital biological clock. Chronotype is classified into three types based on MSF: morning type (MSF $< 3{:}00$), intermediate type ($3{:}00 \leq \text{MSF} \leq 4:00$), and evening type (MSF $> 4{:}00$) (19).

In addition, social jetlag (SJL) is the difference between the biological clock and the social clock, which is defined as

$$\text{SJL} = \text{MSF} - \text{MSW} \tag{8}$$

where MSW is the meantime of mid-sleep on the weekdays.

Although the Morningness Eveningness Questionnaire (MEQ) (20) or the Munich Chronotype Questionnaire (MCTQ) (21) is usually used for calculating MSF and SJL, we used the actigraphy data for calculating them instead of the questionnaires in this study.

Measurements From Questionnaires

In addition to the measurements derived from the actigraphy data, we collected answers to the following, modified for the Japanese population.

- The Zung Self-rating Depression Scale (SDS) (22, 23): a 20-item quantitative measurement of symptoms of depression. The subjects rate each item regarding how they felt during the week preceding. Scores of $\leq$39, 40–49, and $\geq$ 50 on the SDS indicate no, mild, and moderate-to-severe depressive symptoms, respectively.
- The Epworth Sleepiness Scale (ESS) (24, 25): an eight-item questionnaire that is widely used for assessment of daytime sleepiness in adults. ESS is used to assess the severity of insomnia with a possible score range of 0–24 points. An ESS score of 11 points or higher indicates excessive daytime sleepiness.
- The Pittsburgh Sleep Quality Index (PSQI) (26, 27): a 19-item self-rated questionnaire assesses sleep quality. Nineteen individual items generate seven component scores: subjective sleep quality, sleep latency, sleep duration, habitual sleep efficiency, sleep disturbances, use of sleeping medication, and daytime dysfunction. The sum of scores for these seven components yields one global score. PSQI global score of > 5 indicates poor sleep quality.
- The Social Rhythm Metric (SRM) (28): a diary-like questionnaire that quantifies the extent to which a person's life is regular on a daily basis with respect to event timing. It gives a score based on the timing of five activities that are thought to constitute an individual's social rhythm (1: Get out of bed, 2: First contact with another person, 3: Start work, housework, or volunteer activities, 4: Have dinner, and 5: Go to bed), The SRM-score lies on a continuum between 0 and 7, with 0 representing lowest regularity and seven highest regularity.
- The 36-item Short-Form Health Survey (SF-36) (26, 29): a 36-item questionnaire that evaluates the health-related quality of life (QOL) from eight components: physical functioning, role limitations–physical, bodily pain, general health, vitality, social functioning, role limitations–emotional, and mental health. For each subscale, a score ranging from 0 (worst) to 100 (best) is calculated and standardized to have a mean of 50 and a standard deviation of 10.
- Morningness-Eveningness Questionnaire (MEQ) (20): a 19-item self-rated questionnaire, each having four or five response options. The last question, item 19, asks whether a participant estimates oneself as definitely a morning type, rather more a morning type than an evening type, rather more an evening type than a morning type, or definitely an evening type. The sum gives a score ranging from 16 to 86. Scores of 70–86, 59–69, 42–58, 31–41, and 16–30 are classified as definitely morning type, moderately morning type, neither type, moderately evening type, and definitely evening type, respectively.

Definition of Weekday Sleep Debt

Sleep duration (SD) of each day can be calculated using the sleep/wake function $r(i)$. By using the means of SD on the weekend (SD on a free day; SDF) (min), and SD on the weekday (SDW) (min), sleep rebound on the weekend (SRW) (min) (30) is defined as

$$\text{SRW} = \text{SDF} - \text{SDW} \tag{9}$$

A person may have chronic sleep deprivation during the weekdays when SRW $\geq$ 120 min (31). Based on this criterion, we judge that a person has WSD when SRW $\geq$ 120 min.

Machine Learning-based Factor Identification

In this research, we adopt machine learning (ML) technologies to find the factors of WSD. A classifier that predicts whether a person has WSD is trained from the measurements described above or not, and feature importance is calculated based on the trained classifier. Feature importance is a well-known method for identifying which input features in a classifier contribute to output (32).

Although various classifier training methods have been used in ML, we adopted a support vector machine (SVM) in this work. SVM is a classical nonlinear classification technique that was originally developed for classifying data into two classes (33). SVM has been widely used for various applications such as spam mail filtering, bioinformatics, and object recognition (34–36). Thus, SVM is a reliable ML technique.

TABLE 2 | Characteristics of subjects.

Age	43.8 ± 8.4
BMI, kg/m^2	23.7 ± 3.1
Current smoker	175
Habitual snorer, everyday or often	156
Habitual drinker, almost everyday	177
Alcohol consumption, g·d^{-1}·kg^{-1}	0.5 ± 0.5
Hypertension	50
Daytime sleepiness, ESS > 10	86
ESS score	8.1 ± 4.3
Total sleep time, h	5.9 ± 0.9
Sleep latency, min	10.8 ± 15.7
Sleeping pill use, yes	8

We used a permutation feature importance method to calculate feature importance, which is defined as the decrease in prediction performance when a single input feature is randomly shuffled (37, 38). E is the prediction performance of the trained classifier when the original feature set is input to the classifier, and E_j is the prediction performance of the classifier when a feature x_j is exchanged randomly into another feature. The feature importance of x_j is calculated as

$$\text{Pimp}\left(x_j\right) = E - E_j \tag{10}$$

We can judge that the feature x_j contributes to a prediction when $Pimp\left(x_j\right)$ is large.

Data Description

A cross-sectional survey was conducted in 324 employees at a drug wholesale company in Osaka, Japan, from January 26, 2004, to December 19, 2005.

All of the participants were male. It becomes difficult to focus on the relationship between work habits and activity rhythm when female employees are included in the survey. There are more factors affecting sleep and rhythm in females than in males, such as having small children or not, menstrual cycle, unmarried or married (39, 40). However, some studies have already reported that there were no significant differences in daytime activities between males and females (41, 42). Thus, it is expected that our analysis is applicable to females as well as males.

The workdays were from Monday to Friday, and the holidays were Saturday and Sunday, which are typical for Japanese companies and government offices. We used a wrist-type actigraph device (Actiwatch AW-Light, Mini-Mitter) for data recording, which measures the activity counts and illuminance within every one-minute bin. Illuminance is recorded by a photodiode equipped in the actigraph device. Its measurement range is 0.1 to 150,000 Lux with 0.1 Lux resolution, which is suitable for both interior and exterior illuminance measurement. The actigraph device worked for 2–3 months per one battery. Thus, data missing did not occur during data collection. In addition, we collected answers to the questionnaires and sleep diaries during the survey. The protocol of this study was approved by the Ethics Committee of the Shiga University of Medical Science (R2020-026).

Since all participants wore the actigraph device from around 5:00 pm on a Monday to around 8:00 am on the next Monday, the Monday (the first day) actigraphy data were not recorded in their entirety. The Monday data were excluded from the analysis except for wake-up and sleep-onset time estimation. On the other hand, sleep-onset time, mid-sleep, and sleep duration of Friday were not able to be calculated because sleep in the Friday night sleep is regarded as a weekend.

One hundred and six participants who removed the actigraph device for more than 30 min during the survey term were excluded from the analysis so that 218 participants were included in the analysis. The characteristics of the subjects are illustrated in **Table 2**, and a detailed description of the data used in this study is available in Kadotani et al. (3), Nakayama-Ashida et al. (43), and Gerstner et al. (44).

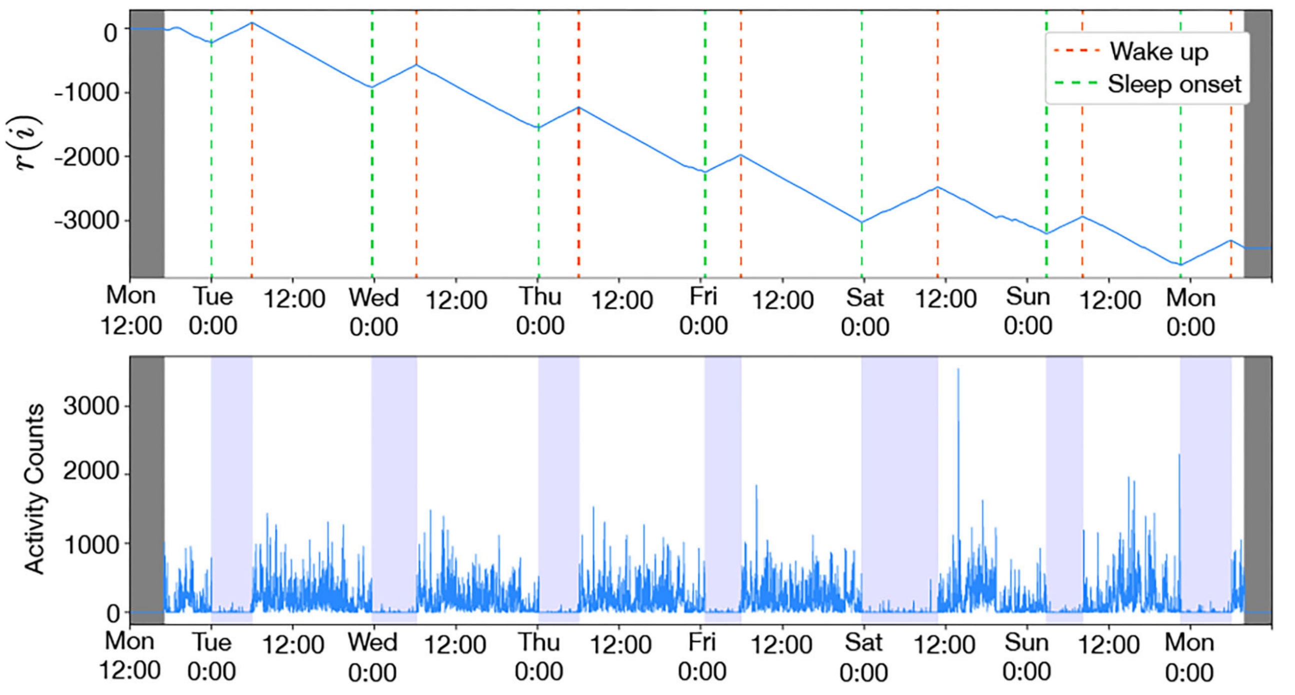
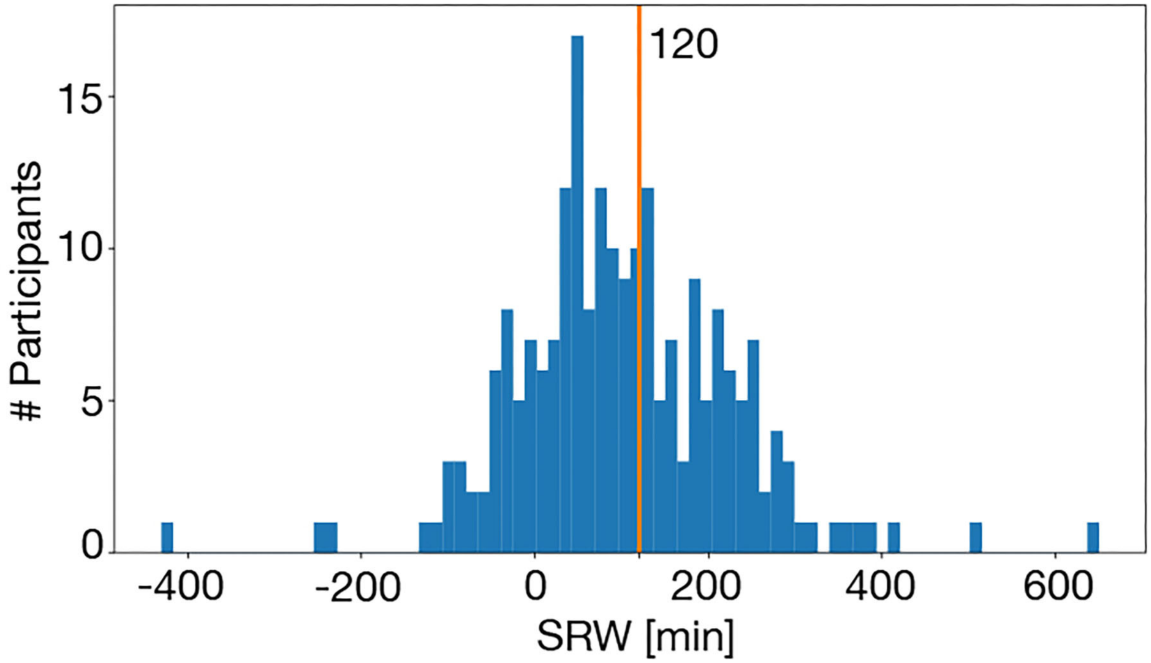


FIGURE 2 | Example of wake-up and sleep-onset times estimated by activity-counts (male, 41 years old, MSF = 3:32): the colored bands in the bottom figure denote the duration between the estimated wake-up and sleep-onset. The sleep/wake function $r(i)$ decreased during daytime and increased during the night, and the awake and sleep states repeated in the 24-h cycle. Since the awake state is longer than the sleep state, $r(i)$ had a downward trend with the up and down periodic cycle.



SRW = SDF-SDW, SRW: sleep rebound on the weekend,
SDF: sleep duration on a free day, SDW: sleep duration on the weekday

FIGURE 3 | Distribution of SRW: the red vertical line denotes the cut-off value of WSD (120 min), and the right and left sides of this line are the WSD and the non-WSD groups.

Our dataset consisted of 1-week actigraph data, which seems to be rather short for analysis. However, Tienoven et al. calculated SRM of healthy persons from their 1-week activity data (45). Fonseca et al. evaluated RM of stroke patients and healthy persons based on 1-week activity data (46). In addition, the sleep extension term required before performing PSG is also 1 week according to ICSD-3 (2). These indicate that 1-week activity data can be used for sleep rhythm evaluation.

An example of sleep/awake state estimation is shown in **Figure 2**. All participants were classified into the WSD group or the non-WSD group based on the SRW calculated from the actigraphy data. The numbers of persons with and without WSD were 89 and 129, respectively. **Figure 3** shows the distribution of SRW in this data.

Analysis Procedure

Before analysis, we validated the accuracy of the actigraphy data through a comparison between sleep durations derived from the actigraphy and the sleep diaries recorded by participants themselves. In addition, the answers of MEQ item 19, which asks self-awareness of chronotype were compared to verify whether the definition of WSD adopted in this study was supported by self-awareness of chronotype.

To specify the factors of WSD based on the feature importance, we needed to construct a classifier that predicts the presence of WSD. In general, the numbers of positive and negative examples for training should be balanced in binary classification problems to construct a good classifier (47). We randomly discarded non-WSD data so that the numbers of persons with WSD and non-WSD became even by means of random under-sampling because the number of persons without WSD was larger than that of persons with WSD.

As the input features of the WSD classifier, we used 40 features listed in **Table 1**, derived from the actigraphy data. The Gaussian kernel with a parameter γ was adopted in SVM, of which the parameter γ was tuned by means of 10-fold cross-validation. The WSD classifier training and permutation feature importance computation were repeated 100 times by changing the training and the test samples at random, in which the ratio of the number of training samples to that of test samples was fixed at 7:3. Finally, the mean of the computed permutation feature importance was calculated. In this study, we focus on features occupying the top 20% of the sum of importance, which does not necessarily mean only these factors are important (or statistically significant). We used the top 20% just for limiting the number of factors to discuss in detail, in which the cut-off value can be varied.

In addition, we examined whether the presence of WSD may relate to occupational categories or not. In this survey, employees were classified into the following categories: clerical, managers, professional and technical sales, service, transportation and communication, manufacturing, and others (43).

We compared each of the 40 input features and each of the collected answers to the questionnaires between the WSD and non-WSD groups by means of a statistical test.

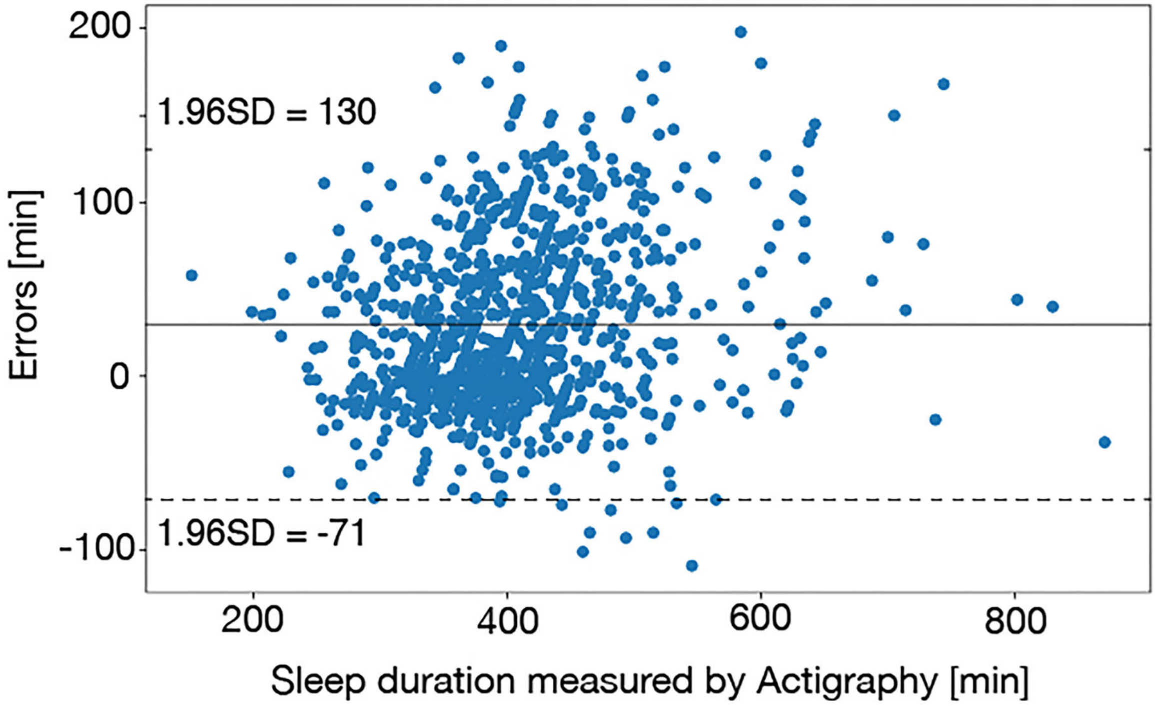


FIGURE 4 | The Bland-Altman plot between sleep durations derived from the actigraphy and the sleep diaries. Most of the sleep duration errors were scattered within 1–1.5 h, which was appropriate accuracy.

TABLE 3 | The participants distribution in item 19 of MEQ.

	Definitely a morning type	Rather more a morning type than an evening type	Rather more an evening type than a morning type	Definitely an evening type
WSD	13	31	28	17
non-WSD	11	57	35	26

Statistical Analysis

We used the t-test for comparison between the WSD and non-WSD groups. The significance level was set to $p < 0.05$. To consider the multiple comparisons of the total of 40-measurements from the actigraphy data and 14-measurements from the questionnaires, the Bonferroni correction was adopted. That is, the significance levels were corrected as $p < 0.05/40 = 0.00125$ and $p < 0.05/13 = 0.0038$. In addition, we used the χ^2 test with significance $p < 0.05$ for examining the effect of occupational categories on WSD and MEQ item 19.

Computation in this study was performed in Python 3.6.4 with SciPy 1.3.0, NumPy 1.16.2, and scikit-learn 0.20.2.

RESULTS

First, we checked the validity of the actigraphy data. **Figure 4** is the Bland-Altman plot between sleep durations derived from the actigraphy and the sleep diaries, which shows most of the sleep duration errors were scattered within −71 to 130 min. Thurman et al. (48) reported that there were errors of >1.5 h between sleep durations derived from the actigraphy and the sleep diaries in 79% of healthy adults. Short et al. (49) confirmed that the average errors between sleep durations derived from the actigraphy and the sleep diaries were 87 min in healthy adolescents. According

TABLE 4 | Comparison of questionnaires (significance level with the Bonferroni correction: $p < 0.0038$).

		WSD	non-WSD	*p*-value
PSQI		4.86 ± 1.97	4.60 ± 1.84	0.34
ESS		8.28 ± 3.99	7.56 ± 4.29	0.23
SDS		36.6 ± 5.72	36.3 ± 6.34	0.75
SRM		4.63 ± 0.860	4.97 ± 0.874	0.0056
MEQ		53.7 ± 5.38	54.6 ± 5.45	0.12
SD-36	PF	53.6 ± 8.32	53.6 ± 5.11	0.93
	RP	52.6 ± 7.89	52.2 ± 6.88	0.72
	BP	50.4 ± 9.07	51.6 ± 9.44	0.36
	GH	52.1 ± 8.69	49.7 ± 9.43	0.063
	VT	49.7 ± 8.90	49.9 ± 8.00	0.88
	SF	52.7 ± 7.61	52.8 ± 7.37	0.98
	RE	51.5 ± 8.50	51.7 ± 7.21	0.81
	MH	49.8 ± 9.86	50.8 ± 8.32	0.44

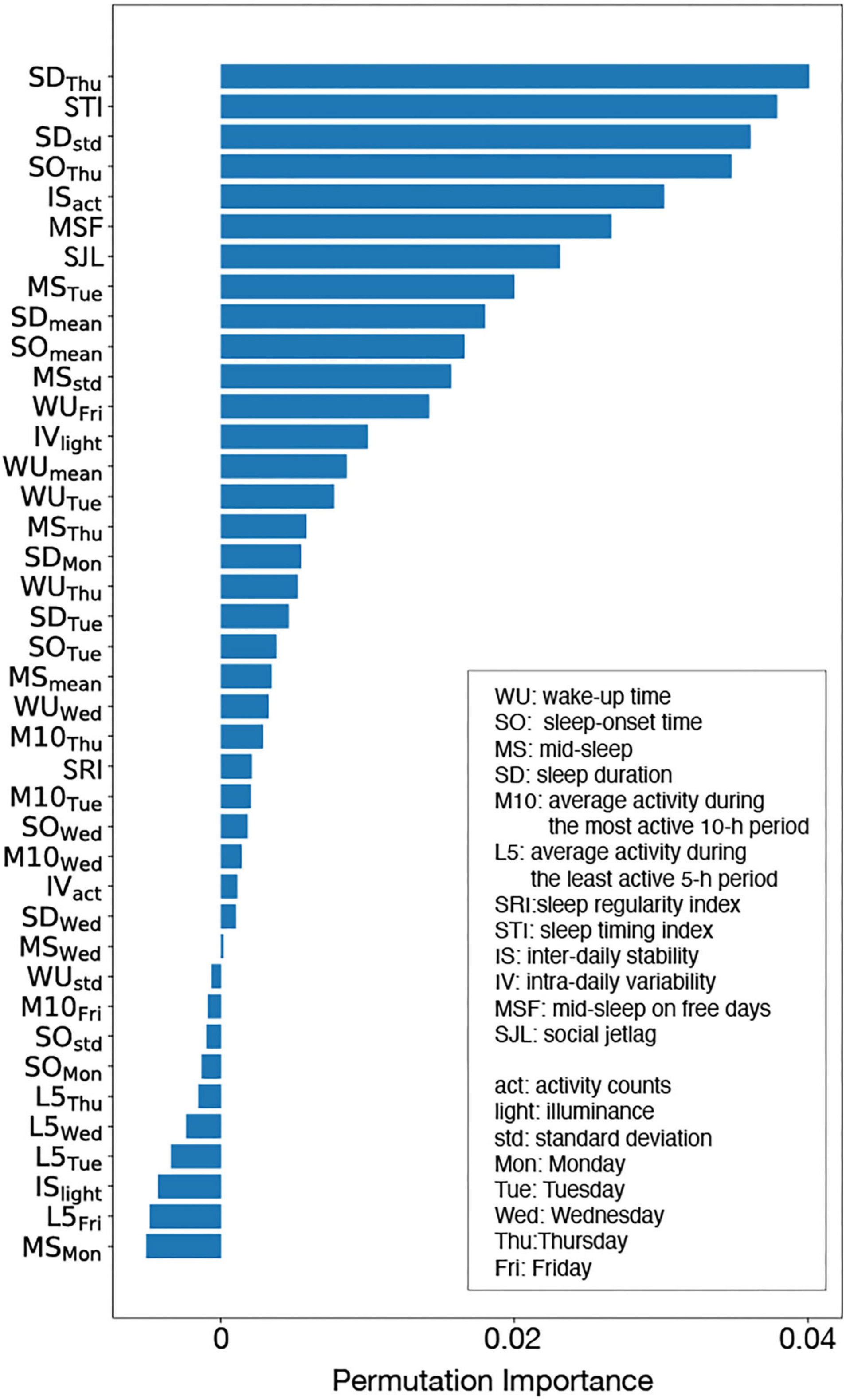


FIGURE 5 | Permutation importance: although there were some features with negative importance, such features did not contribute to WSD prediction. Sleep duration on Tuesday (SD_{Thu}), sleep timing index (STI), the standard deviation of sleep duration (SD_{std}), sleep onset on Tuesday (SO_{Thu}), inter-daily stability calculated by the activity counts (IS_{act}), mid-sleep on free days (MSF), and social jetlag (SJL) were judged as the important features for WSD.

to these studies, the sleep duration errors observed in this study was appropriate.

The participants distribution of MEQ item 19 is shown in **Table 3**, and a significant difference was confirmed in MEQ item 19 through the χ^2 test ($p = 0.12 \times 10^{-7}$). On the other hand, there was no significant difference in the total score of MEQ ($p = 0.12$). Thus, the definition of WSD adopted in this study was associated with self-awareness of chronotype.

We compared the answers to the questionnaires between the WSD and non-WSD groups using the *t*-test with the Bonferroni correction. There were no questionnaires that had a significant difference between the WSD and non-WSD groups, as shown in **Table 4**.

The calculated mean of the permutation feature importance is shown in **Figure 5**, which shows that the estimated WSD factors are as follows: sleep duration on Thursday ($\mathrm{SD}_{\mathrm{Thu}}$), sleep timing index (STI), the standard deviation of sleep duration ($\mathrm{SD}_{\mathrm{std}}$), sleep onset on Thursday ($\mathrm{SO}_{\mathrm{Thu}}$), inter-daily stability calculated by the activity counts ($\mathrm{IS}_{\mathrm{act}}$), mid-sleep on free days (MSF), and social jetlag (SJL).

Table 5 shows the comparison result of input features between the WSD and non-WSD groups by the *t*-test with the Bonferroni correction, in which features displayed by bold fonts have a significant difference between the two groups. In addition, the last column in **Table 5** indicates the estimated WSD factors by the permutation feature importance. According to the *t*-test, MSF and SJL were significantly different, as well as some features about the weekday sleep condition, including $\mathrm{SD}_{\mathrm{Thu}}$ and $\mathrm{SO}_{\mathrm{Thu}}$. The results of the *t*-test almost agreed with those of the permutation importance. Thus, these results indicate that sleep on Thursday, activity rhythm and biological clock might contribute to WSD.

Figure 6 shows the numbers of employees with/without WSD in each occupational category. The additional χ^2 test examining the effect of occupational categories on WSD showed that the occupational category was not associated with the presence of WSD in this dataset ($p = 0.87$).

We constructed additional two classifiers using the actigraphy data collected from two groups: a salesperson group ($N = 119$) and a non-salesperson group ($N = 205$). The calculated means of the permutation feature importance derived from each classifier were shown in **Figure 7**. The top five WSD factors were the same, which shows that the factors of WSD were not related to specific occupational categories.

TABLE 5 | Comparison of measurements between the two groups (significance level with the Bonferroni correction: $p < 0.00125$).

	WSD	non-WSD	*p*-value	Importance
$\mathrm{WU}_{\mathrm{Tue}}$	334 ± 49.1	348 ± 43.2	0.031	
$\mathrm{WU}_{\mathrm{Wed}}$	335 ± 48.4	346 ± 61.0	0.15	
$\mathrm{WU}_{\mathrm{Thu}}$	340 ± 55.8	349 ± 49.0	0.22	
$\mathrm{WU}_{\mathrm{Fri}}$	342 ± 39.6	348 ± 48.1	0.30	
$\mathrm{WU}_{\mathrm{mean}}$	338 ± 37.0	348 ± 41.9	0.065	
$\mathrm{WU}_{\mathrm{std}}$	19.8 ± 24.7	18.3 ± 22.0	0.66	
$\mathrm{SO}_{\mathrm{Mon}}$	1,446 ± 71.4	1,422 ± 62.6	0.012	
$\mathrm{SO}_{\mathrm{Tue}}$	1,442 ± 81.5	1,420 ± 70.6	0.040	
$\mathrm{SO}_{\mathrm{Wed}}$	1,419 ± 83.2	1,403 ± 76.4	0.17	
$\mathrm{SO}_{\mathrm{Thu}}$	**1,429 ± 71.2**	**1,395 ± 71.4**	**0.00061**	*
$\mathrm{SO}_{\mathrm{mean}}$	1,434 ± 58.8	1,410 ± 52.9	0.0025	
$\mathrm{SO}_{\mathrm{std}}$	44.0 ± 25.6	41.1 ± 24.6	0.41	
$\mathrm{MS}_{\mathrm{Mon}}$	170 ± 46.5	165 ± 42.3	0.43	
$\mathrm{MS}_{\mathrm{Tue}}$	169 ± 43.8	163 ± 53.9	0.39	
$\mathrm{MS}_{\mathrm{Wed}}$	159 ± 52.8	156 ± 47.9	0.65	
$\mathrm{MS}_{\mathrm{Thu}}$	166 ± 43.7	152 ± 48.7	0.027	
$\mathrm{MS}_{\mathrm{mean}}$	166 ± 36.7	159 ± 39.5	0.18	
$\mathrm{MS}_{\mathrm{std}}$	24.3 ± 16.6	23.9 ± 15.3	0.84	
$\mathrm{SD}_{\mathrm{Mon}}$	**328 ± 79.7**	**366 ± 66.6**	**0.00031**	
$\mathrm{SD}_{\mathrm{Tue}}$	333 ± 101.5	366 ± 76.0	0.011	
$\mathrm{SD}_{\mathrm{Wed}}$	361 ± 94.4	386 ± 85.5	0.053	
$\mathrm{SD}_{\mathrm{Thu}}$	**352 ± 75.0**	**393 ± 73.1**	**0.00011**	*
$\mathrm{SD}_{\mathrm{mean}}$	**344 ± 65.3**	**378 ± 53.5**	**0.000082**	
$\mathrm{SD}_{\mathrm{std}}$	52.6 ± 30.9	45.3 ± 30.8	0.093	*
$\mathrm{M10}_{\mathrm{Tue}}$	22,633 ± 6,991	21,362 ± 7,114	0.19	
$\mathrm{M10}_{\mathrm{Wed}}$	22,019 ± 6,524	21,320 ± 7,200	0.58	
$\mathrm{M10}_{\mathrm{Thu}}$	22,490 ± 6,784	21,134 ± 6,810	0.47	
$\mathrm{M10}_{\mathrm{Fri}}$	23,555 ± 7,677	22,109 ± 7,406	0.22	
$\mathrm{L5}_{\mathrm{Tue}}$	900 ± 655	953 ± 711	0.15	
$\mathrm{L5}_{\mathrm{Wed}}$	741 ± 573	860 ± 765	0.36	
$\mathrm{L5}_{\mathrm{Thu}}$	742 ± 654	830 ± 705	0.17	
$\mathrm{L5}_{\mathrm{Fri}}$	796 ± 660	890 ± 851	0.38	
SRI	69.8 ± 7.51	70.2 ± 7.72	0.74	
STI	151 ± 39.9	147 ± 40.9	0.53	*
$\mathrm{IS}_{\mathrm{act}}$	0.468 ± 0.0638	0.491 ± 0.072	0.015	*
$\mathrm{IV}_{\mathrm{act}}$	0.591 ± 0.0943	0.573 ± 0.0928	0.16	
$\mathrm{IS}_{\mathrm{light}}$	0.343 ± 0.0684	0.348 ± 0.0724	0.65	
$\mathrm{IV}_{\mathrm{light}}$	0.725 ± 0.239	0.693 ± 0.233	0.33	
MSF	**255 ± 108.3**	**210 ± 78.2**	**0.0010**	*
SJL	**89 ± 105.4**	**51 ± 69.1**	**0.0034**	*

**The estimated WSD factors by the permutation feature importance. Bold features indicate that have a significant difference between the two groups.*

DISCUSSION

Chronotype and WSD

There was a significant difference in MEQ item 19 between WSD and non-WSD groups ($p = 0.12 \times 10^{-7}$) although a significant difference in the total score of MEQ was not confirmed ($p = 0.12$).

Turco et al. (50) assessed the reliability of MEQ item 19 in comparison with the total score of MEQ, the time of subjective sleepiness, and real-life sleep timing variables. They found that significant differences in sleep-wake timing between the answers of MEQ item 19. In addition, such differences were still observed when sleep-wake habits were analyzed separately on work and free days. In addition, Arrona-Palacios and Díaz-Morales (51) confirmed Turco's study in the Mexican and Spanish populations. These findings suggest that MEQ item 19 is solely effective for subjective chronotype evaluation as well as objective chronotype evaluation.

Thus, it is concluded that the definition of WSD supported by MEQ item 19 was appropriate from the viewpoint of chronotype.

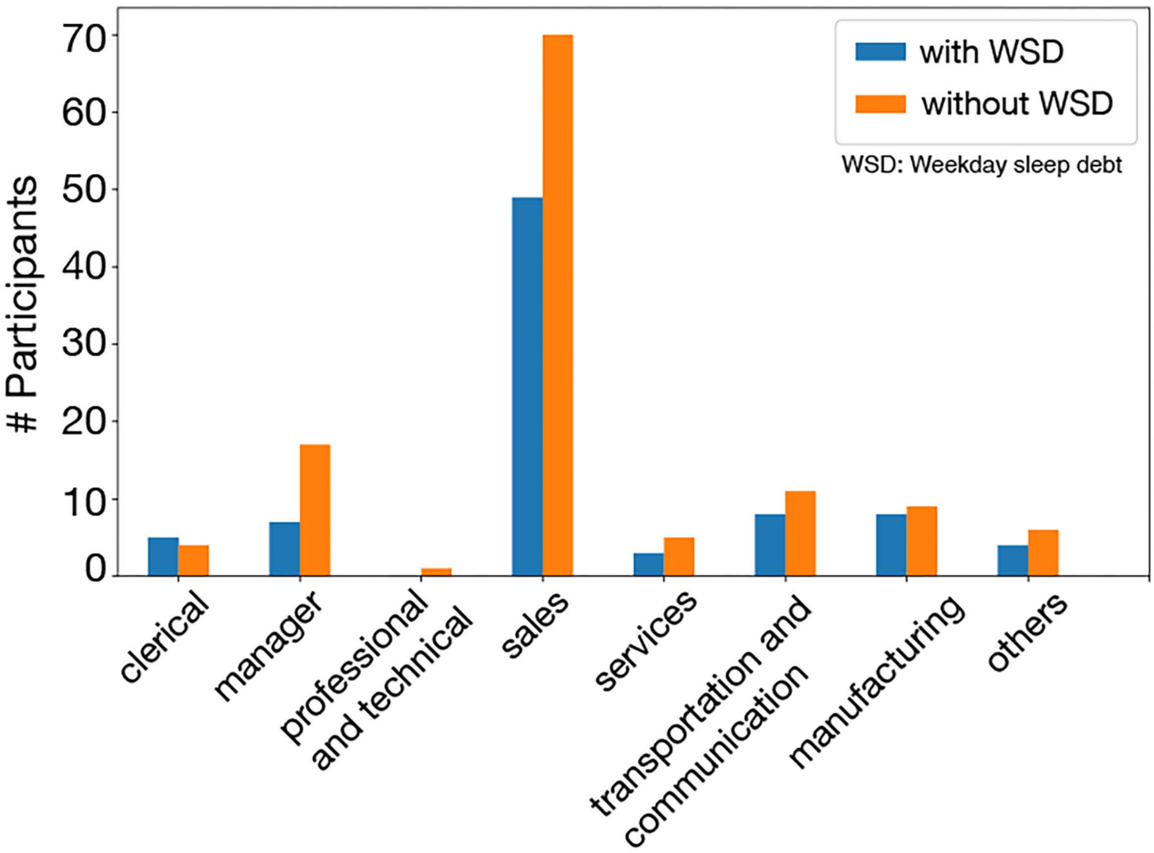


FIGURE 6 | WSD and non-WSD groups in occupational categories: in all categories, there was not much difference between the WSD and the non-WSD groups.

Day-to-Day Variance

The weekday sleep conditions were different between the WSD and non-WSD groups, according to **Table 2**. Sleep duration during the weekdays of the WSD group was significantly shorter than that of the non-WSD group except for Wednesday ($p < 0.05$), which may be plausible from the viewpoint of the assumption that WSD compensates for sleep deprivation during the weekdays. SD_{Thu} and SO_{Thu} were important features for WSD based on the analysis result of permutation importance analysis, which indicated that Thursday was associated with WSD.

Morning Evening Tendency

There were significant differences in MSF and SJL, the biological clock features ($p < 0.05$). MSF and SJL of the WSD group were later than those of the non-WSD group. That is, persons with WSD have eveningness tendencies. In addition, MSF and SJL were also important to WSD according to the permutation importance analysis. Since all of the participants worked at the same company, they might live and work following the same social clock. Thus, this result indicated that disagreement between the social clock and the individual biological clock might contribute to WSD, which is consistent with the previous study (5).

Activity Levels Variance

The inter-daily stability by the activity counts (IS_{act}) evaluates the synchronization of diurnal activity rhythm, and it was also an important feature for WSD according to the permutation importance analysis, although the significant difference of IS_{act} between the WSD and the non-WSD groups was not confirmed. **Figure 5** shows the 24-h activity counts during the weekdays of two participants with and without WSD. Inter-daily synchronization of activity rhythm, including sleep conditions, might affect WSD.

On the other hand, intra-daily variability based on the activity counts (IV_{act}) was not significantly different between the WSD and non-WSD groups, and its feature importance was low. Since IV_{act} denotes the frequency of midday naps and nocturnal awakening, its low importance suggests that sleep disorders, such as narcolepsy or sleep apnea, might not contribute to WSD. In addition, IS and IV by illuminance (IS_{light} and IV_{light}) were not significantly different and of low importance, implying that illuminance might not affect WSD.

It was of note that Thursday was more important for WSD presence than other weekdays. This may be related to the participants' working conditions. All participants worked at the same drug wholesale company, which provides various medical supplies to hospitals and clinics. Since private clinics

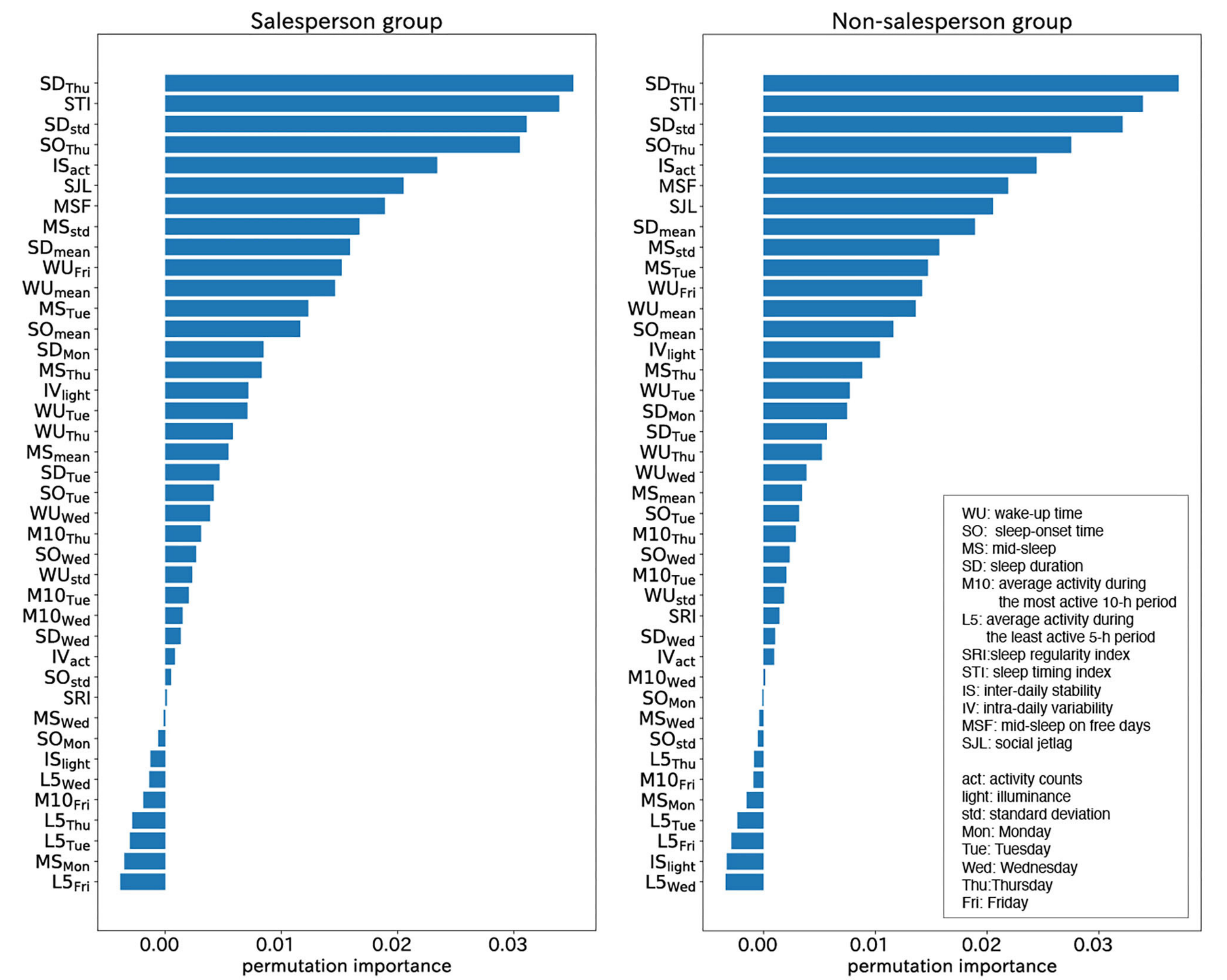


FIGURE 7 | Permutation importance calculated from salesperson group (left) and non- salesperson group (right): the top five WSD factors were the same. The factors of WSD were not related to specific occupational categories.

in Japan tend to close on Thursday evening (52), work habits on Thursdays are different from other workdays at the drug wholesale company. For example, employees might tend to do desk work rather than outside work or to go home earlier than other workdays. Such differences in work schedules might affect their social clock and disturb the synchronization of diurnal activity rhythm, which is consistent with the causes of WSD (5). **Figure 8** clearly illustrates that the activity of a participant with WSD on Thursday was significantly different from those of other days while the activity of a participant without WSD synchronized even on Thursday. This difference might indicate increased desk time and reduced outside work on Thursday.

Thus, it is reasonable that Thursday was an important factor of WSD at the company. This analysis indicates that the work style differences on a specific day of the week may cause WSD.

Occupational Category

In this survey, participants were classified into a salesperson group and a non-salesperson group, and the top five permutation feature importance derived from each group was the same, as shown in **Figure 6**. This result indicated that the factors of WSD were not related to specific occupational categories, which also suggests that our analysis may be applicable to data collected from other companies.

The limitations of this study include the collected data; all of the participants were male and Japanese. Hence, we could consider neither gender nor racial differences in this study. In addition, the participants worked at the same company; it was difficult to investigate the effects of other days of the week on WDS since we did not compare our data with persons working at companies that have different work habits on a specific day of the week. Accordingly, we need to collect data from employees

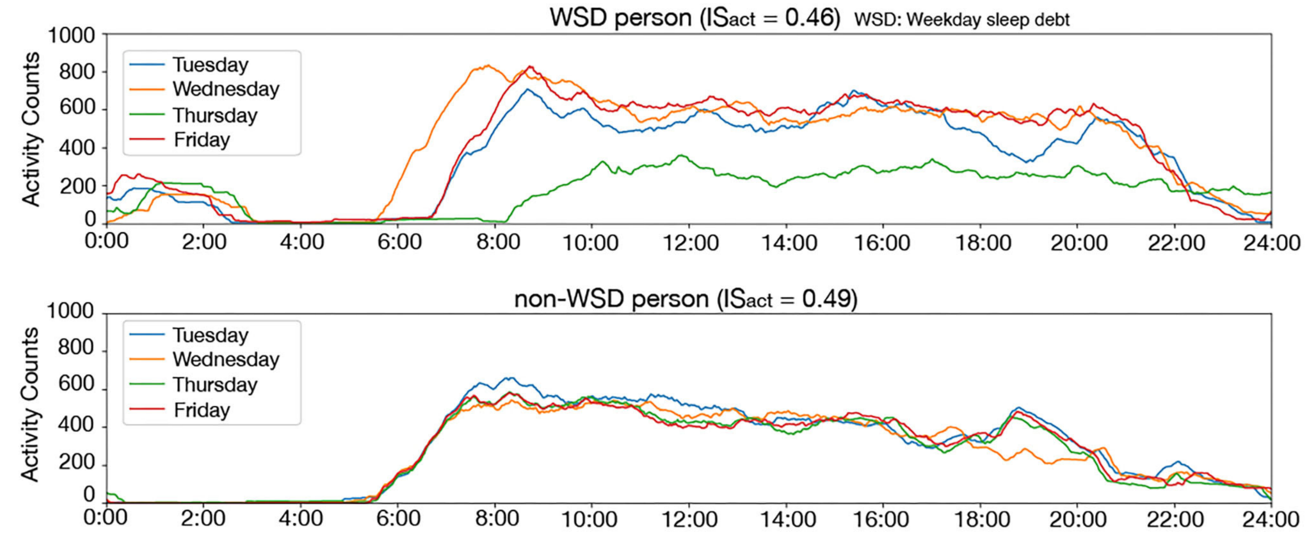


FIGURE 8 | Twenty four-hours activity counts on the weekday of a person with WSD **(top)** and without WSD **(bottom)**: the activity counts are smoothed by using the 60-min moving average for clear display. This figure clearly illustrates that the activity counts on Thursday were clearly different from other days in a person with WSD.

working in various types of industries to confirm our results. In addition, we were not able to analyze the Monday activity data due to the constraint of the data collection, which might affect the analysis results.

CONCLUSION

In this study, we collected actigraphy data from 324 healthy male employees at a drug wholesaler and calculated permutation feature importance based on SVM to identify the factors of WSD. We compared the answers to questionnaires between the WSD and the non-WSD groups.

Our analysis results indicated that sleep duration during the weekdays and the individual biological clock might affect WSD, which is consistent with previous studies. In addition, we demonstrated a new finding that turbulence of diurnal activity rhythm synchronization as well as nocturnal sleep rhythm, even for 1 day, is associated with WSD, which is a new finding of this work.

In the future, we will develop a quantitative evaluation methodology for sleep debt based on this study.

DATA AVAILABILITY STATEMENT

The data analyzed in this study is subject to the following licenses/restrictions: The actigraphy data will be made available by the corresponding author to colleagues who propose a reasonable scientific request after approval by the institutional review board of the SUMS Hospital. Requests to access these datasets should be directed to Hiroshi Kadotani, kadotanisleep@gmail.com.

ETHICS STATEMENT

The studies involving human participants were reviewed and approved by the Ethics Committee of the Shiga University of Medical Science. The patients/participants provided their written informed consent to participate in this study.

AUTHOR CONTRIBUTIONS

YG and KF analyzed the clinical data and composed the manuscript. YS, MM, and HK collected clinical data. MK and HK checked the analysis result and the manuscript. All authors contributed to the article and approved the submitted version.

FUNDING

This work was partially supported by JST PRESTO #JPMJPR1859 and JSPS KAKENHI #17H00872.

REFERENCES

1. Leproult R, Van Cauter E. Role of sleep and sleep loss in hormonal release and metabolism. *Endocr Dev*. (2010) 17:11–21. doi: 10.1159/000262524
2. AASM (2014). *The AASM Manual for the Scoring of Sleep and Associated Events: Rules, Terminology and Technical Specifications,Version 2.0.3*. Darien, CT.
3. Kadotani T, Kadotani H, Arai H, Takami M, Ito H, Matsuo M, et al. Comparison of self-reported scales and structured interviews for

the assessment of depression in an urban male working population in Japan: a cross-sectional survey. *Sleep Sci Pract*. (2017) 1:9. doi: 10.1186/s41606-017-0010-y

4. Oh YH, Kim H, Kong M, Oh B, Moon JH. Association between weekend catch-up sleep and health-related quality of life of Korean adults. *Medicine.* (2019) 98:e14966. doi: 10.1097/MD.0000000000014966
5. Kitamura S, Katayose Y, Nakazaki K, Motomura Y, Oba K, Katsunuma R, et al. Estimating individual optimal sleep duration and potential sleep debt. *Sci Rep*. (2016) 6:35812. doi: 10.1038/srep35812
6. Roepke SE, Duffy JF. Differential impact of chronotype on weekday and weekend sleep timing and duration. *Nat Sci Sleep*. (2010) 2010:213–20. doi: 10.2147/NSS.S12572
7. Wittmann M, Dinich J, Merrow M, Roenneberg T. Social jetlag: misalignment of biological and social time. *Chronobiol Int*. (2006) 23:497–509. doi: 10.1080/07420520500545979
8. Roenneberg T, Pilz LK, Zerbini G, Winnebeck EC. Chronotype and social jetlag: a (self-) critical review. *Biology.* (2019) 8:54. doi: 10.20944/preprints201905.0092.v1
9. Walker RJ, Kribs ZD, Christopher AN, Shewach OR, Wieth MB. Age, the Big Five, and time-of-day preference: a mediational model. *Person Ind Differ.* (2014) 56:170–4. doi: 10.1016/j.paid.2013.09.003
10. Robillard R, Naismith SL, Smith KL, Rogers NL, White D, Terpening Z, et al. Sleep-wake cycle in young and older persons with a lifetime history of mood disorders. *PLoS ONE*. (2014) 9:e87763. doi: 10.1371/journal.pone.0087763
11. Roenneberg T, Keller LK, Fischer D, Matera JL, Vetter C, Winnebeck EC. Human activity and rest in situ. *Methods Enzymol*. (2015) 552:257–83. doi: 10.1016/bs.mie.2014.11.028
12. Fischer D, Vetter C, Roenneberg T. A novel method to visualise and quantify circadian misalignment. *Sci Rep*. (2016) 6:38601. doi: 10.1038/srep38601
13. Cole RJ, Kripke DF, Gruen W, Mullaney DJ, Gillin JC. Automatic sleep/wake identification from wrist activity. *Sleep*. (1992) 15:461–9. doi: 10.1093/sleep/15.5.461
14. Matsuo M, Masuda F, Sumi Y, Takahashi M, Yamada N, Ohira MH, et al. Comparisons of portable sleep monitors of different modalities: potential as naturalistic sleep recorders. *Front Neurol*. (2016) 7:110. doi: 10.3389/fneur.2016.00110
15. Phillips AJK, Clerx WM, O'Brien CS, Sano A, Barger LK, Picard RW, et al. Irregular sleep/wake patterns are associated with poorer academic performance and delayed circadian and sleep/wake timing. *Sci Rep*. (2017) 7:3216. doi: 10.1038/s41598-017-03171-4
16. Lunsford-Avery JR, Engelhard MM, Navar AM, Kollins SH. Validation of the sleep regularity index in older adults and associations with cardiometabolic risk. *Sci Rep*. (2018) 8:14158. doi: 10.1038/s41598-018-32402-5
17. Witting W, Kwa IH, Eikelenboom P, Mirmiran M, Swaab DF. Alterations in the circadian rest-activity rhythm in aging and Alzheimer's disease. *Biol Psychiatry*. (1990) 27:563–72. doi: 10.1016/0006-3223(90)90523-5
18. Gonçalves BS, Cavalcanti PR, Tavares GR, Campos TF, Araujo JF. Nonparametric methods in actigraphy: an update. *Sleep Sci*. (2014) 7:158–64. doi: 10.1016/j.slsci.2014.09.013
19. Urbán R, Magyaródi T, Rigó A. Morningness-eveningness, chronotypes and health-impairing behaviors in adolescents. *Chronobiol Int.* (2011) 28:238–47. doi: 10.3109/07420528.2010.549599
20. Horne JA, Ostberg O. A self-assessment questionnaire to determine morningness-eveningness in human circadian rhythms. *Int J Chronobiol.* (1976) 4:97–110.
21. Roenneberg T, Wirz-Justice A, Merrow M. Life between clocks: daily temporal patterns of human chronotypes. *J Biol Rhythms*. (2003) 18:80–90. doi: 10.1177/0748730402239679
22. Zung WW. A self-rating depression scale. *Arch Gen Psychiatry*. (1965) 12:63–70. doi: 10.1001/archpsyc.1965.01720310065008
23. Fukuda K, Kobayashi S. *Japanese Version SDS (Self-Rating Depression Scale): Manual.* Kyoto: Sankyobo (1983).
24. Johns MW. A new method for measuring daytime sleepiness: the Epworth sleepiness scale. *Sleep*. (1991) 14:540–5. doi: 10.1093/sleep/14.6.540
25. Takegami M, Suzukamo Y, Wakita T, Noguchi H, Chin K, Kadotani H, et al. Development of a Japanese version of the Epworth Sleepiness Scale (JESS) based on item response theory. *Sleep Med*. (2009) 10:556–65. doi: 10.1016/j.sleep.2008.04.015
26. Buysse DJ, Reynolds CF, Monk TH, Berman SR, Kupfer DJ. The Pittsburgh Sleep Quality Index: a new instrument for psychiatric practice and research. *Psychiatry Res*. (1989) 28:193–213. doi: 10.1016/0165-1781(89)90047-4
27. Doi Y, Minowa M, Uchiyama M, Okawa M, Kim K, Shibui K, et al. Psychometric assessment of subjective sleep quality using the Japanese version of the Pittsburgh Sleep Quality Index (PSQI-J) in psychiatric disordered and control subjects. *Psychiatry Res*. (2000) 97:165–72. doi: 10.1016/S0165-1781(00)00232-8
28. Monk TH, Kupfer DJ, Frank E, Ritenour AM. The Social Rhythm Metric (SRM): measuring daily social rhythms over 12 weeks. *Psychiatry Res*. (1991) 36:195–207. doi: 10.1016/0165-1781(91)90131-8
29. Fukuhara S, Bito S, Green J, Hsiao A, Kurokawa K. Translation, adaptation, and validation of the SF-36 Health Survey for use in Japan. *J Clin Epidemiol*. (1998) 51:1037–44. doi: 10.1016/S0895-4356(98)00095-X
30. Kang SG, Lee YJ, Kim SJ, Lim W, Lee HJ, Park YM, et al. Weekend catch-up sleep is independently associated with suicide attempts and self-injury in Korean adolescents. *Compr Psychiatry*. (2014) 55:319–25. doi: 10.1016/j.comppsych.2013.08.023
31. Morita Y, Sasai-Sakuma T, Asaoka S, Inoue Y. Prevalence and correlates of insufficient sleep syndrome in Japanese young adults: a web-based cross-sectional study. *J Clin Sleep Med*. (2015) 11:1163–9. doi: 10.5664/jcsm.5092
32. Wei P, Lu X, Song J. Variable importance analysis: a comprehensive review. *Reliab Eng Syst Safety.* (2015) 142:399–432. doi: 10.1016/j.ress.2015.05.018
33. Cortes C, Vapnik V. Support-vector networks. *Mach Learn Vol*. (1995) 20:273–97. doi: 10.1007/BF00994018
34. Pontil M, Verri A. Support vector machines for 3D object recognition. *IEEE Transact Pattern Anal Mach Intell*. (1998) 20:637–46. doi: 10.1109/34.683777
35. Bock JR, Gough DA. Predicting protein–protein interactions from primary structure. *Bioinformatics*. (2001) 17:455–60. doi: 10.1093/bioinformatics/17.5.455
36. Amayri O, Bouguila N. A study of spam filtering using support vector machines. *Artif Intell Rev*. (2010) 34:73–108. doi: 10.1007/s10462-010-9166-x
37. Breiman L. Random forests. *Mach Learn*. (2001) 45:5–32. doi: 10.1023/A:1010933404324
38. Strobl C, Boulesteix AL, Kneib T, Augustin T, Zeileis A. Conditional variable importance for random forests. *BMC Bioinform*. (2008) 9:307. doi: 10.1186/1471-2105-9-307
39. Gallicchio L, Hoffman SC, Helzlsouer KJ. The relationship between gender, social support, and health-related quality of life in a community-based study in Washington County, Maryland. *Qual Life Res*. (2007) 16:777–86. doi: 10.1007/s11136-006-9162-4
40. Varì R, Scazzocchio B, D'Amore A, Giovannini C, Gessani S, Masella R. Gender-related differences in lifestyle may affect health status. *Ann Ist Super Sanita*. (2016) 52:158–66. doi: 10.4415/ANN_16_02_06
41. Lieberman HR, Wurtman JJ, Teicher MH. Circadian rhythms of activity in healthy young and elderly humans. *Neurobiol Aging*. (1989) 10:259–65. doi: 10.1016/0197-4580(89)90060-2
42. Jean-Louis G, Mendlowicz MV, Von Gizycki H, Zizi F, Nunes J. Assessment of physical activity and sleep by actigraphy: examination of gender differences. *J Womens Health Gend Based Med*. (1999) 8:1113–7. doi: 10.1089/jwh.1.1999.8.1113
43. Nakayama-Ashida Y, Takegami M, Chin K, Sumi K, Nakamura T, Takahashi K, et al. Sleep-disordered breathing in the usual lifestyle setting as detected with home monitoring in a population of working men in Japan. *Sleep*. (2008) 31:419–25. doi: 10.1093/sleep/31.3.419
44. Gerstner JR, Perron IJ, Riedy SM, Yoshikawa T, Kadotani H, Owada Y, et al. Normal sleep requires the astrocyte brain-type fatty acid binding protein FABP7. *Sci Adv*. (2017) 3:e1602663. doi: 10.1126/sciadv.1602663
45. van Tienoven TP, Minnen J, Daniels S, Weenas D, Raaijmakers A, Glorieux I. Calculating the social rhythm metric (SRM) and examining its use in interpersonal social rhythm therapy (IPSRT) in a healthy population study. *Behav Sci.* (2014) 4:265–77. doi: 10.3390/bs4030265
46. da Fonsêca RD, Lopes RD, de Souza Morais SA, Ferreira PR, Fernandes AB, Campos TF. Short form of the Social Rhythm Metric: a tool to evaluate the social and functional impact on stroke patients. *Sleep Biol Rhythms.* (2019) 17:19–26. doi: 10.1007/s41105-018-0179-1

47. Seiffert C, Khoshgoftaar TM, Van Hulse J, Napolitano A. RUSBoost: a hybrid approach to alleviating class imbalance. *IEEE Transact Syst Man Cybern Part A Syst Humans*. (2010) 40:185–97. doi: 10.1109/TSMCA.2009.2029559
48. Thurman SM, Wasylyshyn N, Roy H, Lieberman G, Garcia JO, Asturias A, et al. Individual differences in compliance and agreement for sleep logs and wrist actigraphy: a longitudinal study of naturalistic sleep in healthy adults. *PLoS ONE*. (2018) 13:e0191883. doi: 10.1371/journal.pone.0191883
49. Short MA, Gradisar M, Lack LC, Wright H, Carskadon MA. The discrepancy between actigraphic and sleep diary measures of sleep in adolescents. *Sleep Med*. (2012) 13:378–84. doi: 10.1016/j.sleep.2011.11.005
50. Turco M, Corrias M, Chiaromanni F, Bano M, Salamanca M, Caccin L, et al. The self-morningness/eveningness (Self-ME): an extremely concise and totally subjective assessment of diurnal preference. *Chronobiol Int*. (2015) 32:1192–200. doi: 10.3109/07420528.2015.1078807
51. Arrona-Palacios A, Díaz-Morales JF. Morningness–eveningness and sleep habits at school: a comparative study between Mexico and Spain. *Biol Rhythm Res*. (2016) 48:175. doi: 10.1080/09291016.2016.1245459
52. The Japan Medical Association (2009). *Survey on Private clinics in Japan (Japanese document)*. Tokyo: The Japan Medical Association.

Conflict of Interest: KF is with Quadlytics Inc. as well as Nagoya University. MK is with Quadlytics Inc. as well as Kyoto University. HK's laboratory, the Department of Sleep and Behavioral Sciences is an endowment department, supported with unrestricted grants from Fukuda Lifetech Co., Ltd., Fukuda Life Tech Keiji Co., Ltd., Tanaka Sleep Clinic, Akita Sleep Clinic, and Ai Ai Care Co., Ltd.

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

frontiers
in Public Health

SYSTEMATIC REVIEW
published: 11 January 2022
doi: 10.3389/fpubh.2021.715403

Systematic Review on Information Technology Approaches to Evaluate the Impact of Public Health Campaigns: Real Cases and Possible Directions

Rafael Pinto[1,2,3]*, *Lyrene Silva*[1], *Ricardo Valentim*[2,4], *Vivekanandan Kumar*[5], *Cristine Gusmão*[2,6], *Carlos Alberto Oliveira*[2,7] *and Juciano Lacerda*[2,8]

[1] Department of Informatics and Applied Mathematics, Federal University of Rio Grande do Norte, Natal, Brazil, [2] Laboratory of Technological Innovation in Health (LAIS), Federal University of Rio Grande do Norte, Natal, Brazil, [3] Information Systems Coordination, Federal Institute of Rio Grande do Norte, Natal, Brazil, [4] Department of Biomedical Engineering, Federal University of Rio Grande do Norte, Natal, Brazil, [5] School of Computing and Information Systems, Athabasca University, Athabasca, Canada, [6] Department of Biomedical Engineering, Federal University of Pernambuco, Recife, Brazil, [7] Multidisciplinary Department of Human Development with Technologies, State University of Rio de Janeiro, Rio de Janeiro, Brazil, [8] Department of Social Communication, Federal University of Rio Grande do Norte, Natal, Brazil

OPEN ACCESS

Edited by:
Priit Kruus,
Tallinn University of Technology, Estonia

Reviewed by:
Aliete Cunha-Oliveira,
Coimbra Nursing School, Portugal
Ana Suzina,
Loughborough University London, United Kingdom
Almudena Muñoz Gallego,
Complutense University of Madrid, Spain

***Correspondence:**
Rafael Pinto
rafael.pinto@lais.huol.ufrn.br

Specialty section:
This article was submitted to Digital Public Health, a section of the journal Frontiers in Public Health

Received: *26 May 2021*
Accepted: *16 December 2021*
Published: *11 January 2022*

Citation:
Pinto R, Silva L, Valentim R, Kumar V, Gusmão C, Oliveira CA and Lacerda J (2022) Systematic Review on Information Technology Approaches to Evaluate the Impact of Public Health Campaigns: Real Cases and Possible Directions. Front. Public Health 9:715403. doi: 10.3389/fpubh.2021.715403

Evaluating the success of a public health campaign is critical. It helps policy makers to improve prevention strategies and close existing gaps. For instance, Brazil's "Syphilis No!" campaign reached many people, but how do we analyze its real impact on population awareness? Are epidemiologic variables sufficient? This study examined literature on using of information technology approaches to analyze the impact of public health campaigns. We began the systematic review with 276 papers and narrowed it down to 17, which analyzed campaigns. In addition to epidemiological variables, other types of variables of interest included: level of (i) access to the campaign website, (ii) subject knowledge and awareness, based on questionnaires, (iii) target population's interest, measured from both online search engine and engagement with Social Network Service, and (iv) campaign exposure through advertising, using data from television commercials. Furthermore, we evaluated the impact by considering several dimensions such as: communication, epidemiology, and policy enforcement. Our findings provide researchers with an overview of various dimensions, and variables-of-interest, for measuring public campaign impact, and examples of how and which campaigns have used them.

Keywords: public health, campaign, evaluation, systematic review, communicable disease

INTRODUCTION

Public communications campaigns (PCC) are strategic to communicating important information to the public (1). In the context of this work, PCCs for Health Care will be called public health campaigns (PHC). PHC aims to influence behavior either on an individual level, through direct messages, or collectively with policies that inspire change (1).

PHC aims to promote awareness, increase knowledge, encourage the adoption of desirable attitudes and behaviors, and contribute to individual and collective health decision-making (2). PHCs are often sponsored by policy makers, and offer preventive recommendations addressing serious health problems, such as sexually transmitted diseases, alcohol, tobacco, and obesity, as well as the dangers of automobiles, guns and pharmaceuticals (3).

For example, the "Sífilis Não!" (Syphilis No!) campaign, communicated the importance of syphilis prevention, and recommended rapid syphilis testing or VDRL test, conducted at Basic Health Units (UBS) of the Unified Health System (SUS) in Brazil. The Ministry of Health sponsored the campaign, to fight the syphilis epidemic in the country. Between Nov 2018 and May 2019, a variety of content reached large audiences through various media platforms (e.g., television, radio, and online), as well as outdoor media, such as billboards and posters, and print media, such as magazines and newspapers.

Campaign evaluations serve several useful purposes. They help policy makers improve prevention strategies and close existing gaps. Analysis can guide the development of better campaigns, to improve the impact or intensify the message. Understanding the effectiveness of a campaign can help to better direct the use of funds and effort. To assess levels of engagement with a given topic on a platforms such as the Internet, researchers are increasingly using computational approaches. A variety of such techniques and technologies have been used to assess the impact of public health campaigns. Subsequently, a secondary study is beneficial both to understand the current state-of-the-art and to guide future research in this domain.

Accordingly, we performed a systematic literature review (SLR) according to Kitchenham and Charters (4) and Petersen et al. (5) to: (1) characterize scientific literature related to using information technology approaches to evaluate the impact of public health campaigns, and (2) summarize the knowledge gained from understanding the techniques employed, variables of interest analyzed, and the metrics used to validate these results. Results of this study help identify gaps and provide a direction to appropriately position new research activities in this domain.

Mass media campaigns have long been a tool for promoting public health (6) and there are many systematic literature reviews in this domain. Some of them focus on specific public campaigns: Te et al. (7) identifies social media health campaigns against the consumption of sugary drinks; Jones and Salazar (8) describes the use of Social Network Services (SNS) in the context of primary HIV prevention; Yadav and Kobayashi (9) assesses newer evidence from quantitative studies on the effectiveness of mass media campaigns for reducing alcohol-impaired driving (AID) and alcohol-related crashes; Vega and Roland (10) describes the social marketing approaches used to increase syphilis awareness in eight US cities.

Conversely, some studies do not target a specific public, such as: Shi et al. (11) aims to differentiate SNS from more traditional health communication approaches; Jacob et al. (12) determines the costs, benefits, and overall economic value of communication campaigns; Robinson et al. (13) evaluates the effectiveness of health communication campaigns that use multiple channels, including mass media, and distribute health-related products; Randolph et al. (14) discusses the effects of three particular campaign strategies, entertainment education, law enforcement, and mass media.

Dorfman et al. (1) proposed a taxonomy of communication campaigns, which includes three axes: purpose, scope, and maturity. By purpose, the authors understand that the goal of a campaign should directly affect the individuals, or the collective policies, that shape social behavior. Scope refers to the most visible part of a communication campaign: its size and extent (In what region? In what period? Is it local, state or national?). Maturity can come over time (campaigns can be unique events or could last for years). It may become more formal, with clearer goals, well-developed materials, and deeper integration into an organization's overall activities.

TABLE 1 | Properties added to the taxonomy proposed by Dorfman et al (1).

Property	Reasoning
Campaign name	The campaign's name can be able to highlight the action novelty's level. New campaigns bring similar slogans compared to previous campaigns, losing their impact.
Ads type	Detailing the advertisement type allows greater accuracy when making comparisons about the effects produced by them.
Topic area	Topic area can demonstrate that there are subjects that can have repercussions on audiences, depending on the values and social representations that they carry out in different societies, as well as knowing if a subject remains in evidence over time.
Launched/sponsor by	Understand whether a public campaign is launched by a government entity, an NGO or by the social action of a private company can produce differences in strategic and impact terms.
Amount spent	The amount spent compared to the identified impacts can be a strategic way of measuring the quality of investment in health communication actions.

This taxonomy was important to characterize a public campaign, but we missed other properties that we believe to be relevant. Such properties, presented in **Table 1**, in addition to Dorfman et al.'s taxonomy, helped us to identify public campaigns within the scope of our work. A catalog with the identification of the primary studies of our research is provided in **Appendix 1**, based on all identified properties.

METHODS

The research methodology for conducting this systematic literature review (SLR) was performed from the guidelines proposed by Kitchenham and Charters (4) and Petersen et al. (5). Although the PRISMA (Preferred Reporting Items for Systematic Reviews and Meta-Analyses) guidelines (15, 16) were not used in this study, the guidelines used follow similar stages, such as: (i) articles were identified through a database search; (ii) articles were screened according to the selection criteria concerning title and abstract; (iii) the eligibility of each full-text article was rated according to predetermined criteria (quality assessment); (iv) approved articles were included in the systematic review.

Research Questions

The goal of this research is to analyze information technology approaches that assess the impact of public health communications campaigns as a health care promotion strategic method. Thus, we framed the research questions using the PICOC (Population, Intervention, Comparison,

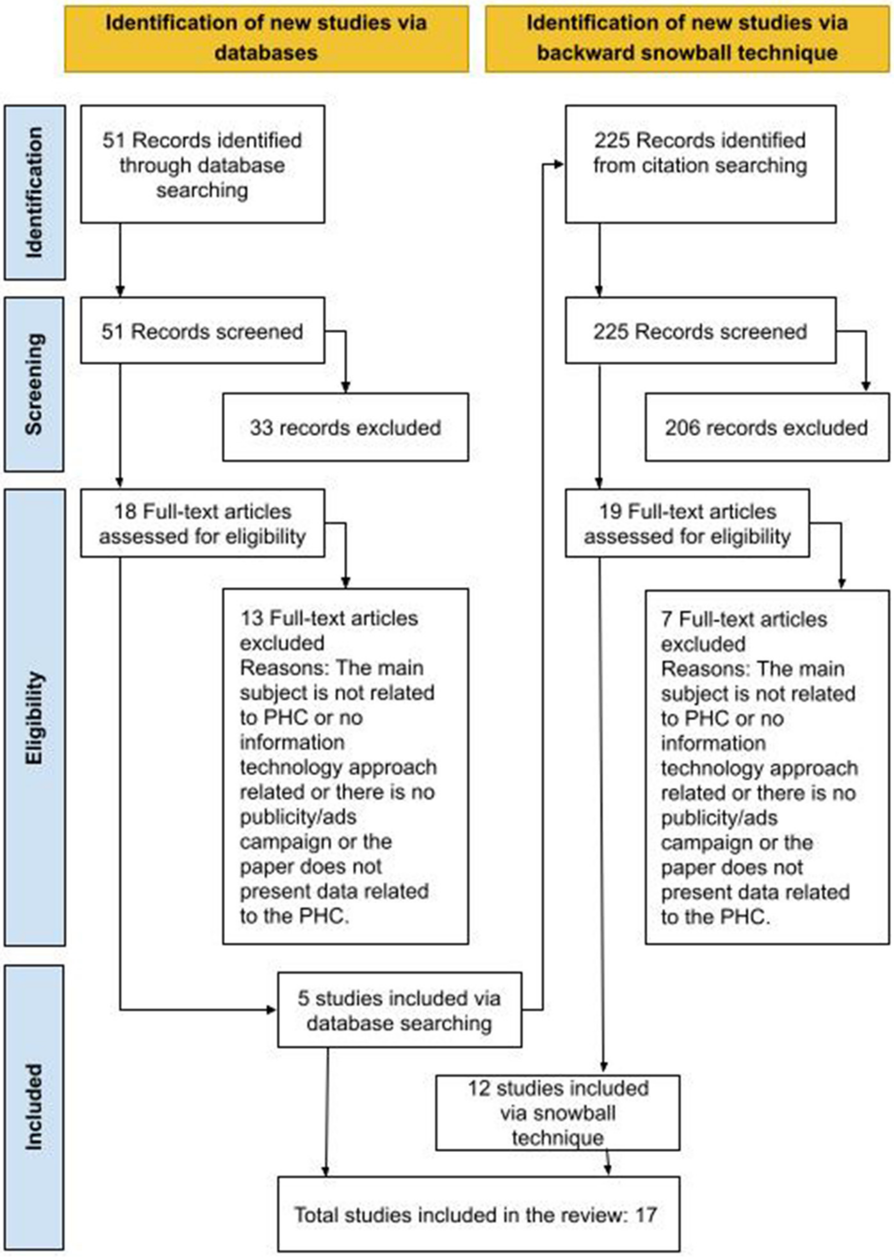

Identification of new studies via databases
Identification of new studies via backward snowball technique
Identification
Screening
Eligibility
Included
51 Records identified through database searching
225 Records identified from citation searching
51 Records screened
225 Records screened
33 records excluded
206 records excluded
18 Full-text articles assessed for eligibility
19 Full-text articles assessed for eligibility
13 Full-text articles excluded Reasons: The main subject is not related to PHC or no information technology approach related or there is no publicity/ads campaign or the paper does not present data related to the PHC.
7 Full-text articles excluded Reasons: The main subject is not related to PHC or no information technology approach related or there is no publicity/ads campaign or the paper does not present data related to the PHC.
5 studies included via database searching
12 studies included via snowball technique
Total studies included in the review: 17

FIGURE 1 | Flow diagram of the study selection process and results.

Outcome, and Context) criteria as suggested by Kitchenham and Charters (4).

- Population: public health campaigns.
- Intervention: the use of technology information to evaluate health campaigns.
- Comparison: characterization of the measures used to support the evaluation of health campaign.
- Outcomes: characterization of the produced output, emergent techniques, overlooked areas and open issues.
- Context: any case where technology information approaches were used to evaluate health campaign.

This leads to the following research questions (RQs):

- RQ1: When and where have the studies been published?
- RQ2: What are the characteristics (topic area, location and year) of the campaigns?
- RQ3: What variables of interest were analyzed to assess the impact of public health campaigns?

TABLE 2 | Quality scores.

Questions	S01	S02	S03	S04	S05	S06	S07	S08	S09	S10	S11	S12	S13	S14	S15	S16	S17
1. Is the campaign clearly characterized?																	
Does the campaign have a name?	1	0	1	1	1	1	0	0.5	1	1	1	1	0.5	0.5	0.5	1	1
Does the campaign have ads?	0	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1
Does the campaign have a target audience?	1	1	1	1	1	1	1	1	1	0	1	1	1	1	1	1	1
Does the campaign have a period of time?	1	1	1	1	1	1	0	1	1	1	1	1	1	1	1	1	1
Does the campaign have a level of organization?	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1
Does the campaign have a representative/sponsor?	0	1	1	1	1	1	0	0	1	1	1	1	1	1	1	1	1
Total	0.67	0.83	1.00	1.00	1.00	1.00	0.50	0.80	1.00	0.83	1.00	1.00	1.00	1.00	1.00	1.00	1.00
2. Is the scheme classification clearly defined as evaluation research?																	
Are techniques, methods, tools or other solutions evaluated in practice?	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1
Are the outcomes investigated?	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1
Total	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1
Overall quality	0.83	0.92	1.00	1.00	1.00	1.00	0.75	0.90	1.00	0.92	1.00	1.00	1.00	1.00	1.00	1.00	1.00

- RQ4: What techniques or tools have been used to support campaign impact analysis?

Search Process

Scopus database was the main source of scientific papers for the current study. We selected this digital library because it contains publications from major journals and conference proceedings, and it is one of the largest curated bibliographic abstract and citation databases of research literature (17). Moreover, recent bibliographic research indicated it as the most comprehensive and user-friendly database (18, 19).

In May 2020, we conducted a preliminary search with a set of terms related to the research topics based on Population and the Intervention previous defined. The search generated many results. However, few were relevant. We repeated this process until we found a search string with a set of relevant results. To maximize recall, we did not include the comparison, outcomes, and context criteria in the search term. We considered the context criterion in Inclusion Criteria (IC), while the outcomes and the comparison criteria defined our data extraction strategy.

The set of search terms included three aspects: (i) "campaign*" AND (ii) ("public health" OR "communicable disease" OR "disease transmission" OR "transmitted infection" OR "disease outbreak*" OR "illness outbreak*" OR "infectious disease*" OR "disease surveillance" OR "disease epidemiology") AND (iii) ("impact" OR "correlation" OR "assess" OR "effectiveness" OR "efficacy" OR "evaluation" OR "analysis").

The automated search looked for these terms in titles, abstracts, and keywords, in the Computer Science area. To increase the quality of the studies found, they had to be published in a peer- reviewed journal. In addition, a backward snowball search (20) was performed by scanning the bibliographies of all selected papers.

Inclusion and Exclusion Criteria

The systematic review included studies that met the following two inclusion criteria: (IC1) The study discusses the effectiveness of using information technology approaches to measure public health campaigns; (IC2) The paper evaluates a health care campaign.

Regarding the exclusion criteria (EC), studies were excluded if they presented at least one of the following criteria: (EC1) The study is not written in English; (EC2) The study is an older version of other study already considered; (EC3) The study is a short publication (<5 pages) or poster; (EC4) The study is a gray literature (For example: technical reports, white papers, and work in progress); (EC5) Paper full text is not available for download.

Paper Selection

Figure 1 shows the steps and quantities of papers returned in the entire search process. The process of selection and classification of the studies was performed by the first author of this article and the second author reviewed the process to avoid research bias.

Titles and abstracts of papers resulting from the automated search were evaluated, rejecting papers that were obviously not relevant (Stage 1-identification). If eligibility was unclear from the abstract, the study was retained for further evaluation (Stage 2–screening). Subsequently, in Stage 3-eligibility, the full texts were read to confirm the criteria and submitted to Stage 4–included for data extraction.

Fifty-one Studies Were Analyzed in Stage 2 and 18 in Stage 3. However, Only five Documents Were Included for Data Extraction. A Complete Analysis of 18 Documents Was Necessary to Understand the Research Context and to Apply the Inclusion and Exclusion Criteria More accurately.

Those five documents served as basis for a subsequent analysis, through their own references, using the backward snowball technique. Thus, 225 articles were identified from citation searching (Stage 1), and screened (Stage 2) in order to verify the inclusion and exclusion criteria. In Stage 3, 19 articles were selected for complete reading and, finally, 12 articles met

TABLE 3 | Data extraction form.

	Data item	Description	RQ
Paper	Paper ID	Integer	Overview
	Title	Name of the paper	Overview
	Author	Set of names of the authors	Overview
	Country	First author country	RQ1
	Venue	Name of publication venue	RQ1
	Year	Year of publication	RQ1
Campaign	Campaign name	Name/Title of the campaign	Overview/QA
	Ads type	Ads type used in the campaign	Overview/QA
	Topic area	Topic area addressed in the campaign	RQ2
	Target audience	General Public; Target Public; Policy-makers	Overview/QA
	Country campaign	Country where the campaign was carried out	RQ2
	Period of time	Campaign duration period	RQ2
	Level of organization	Local, state or national	Overview/QA
	Launched/sponsor by	Representative who launched or sponsored the campaign	Overview/QA
	Amount spent	Total amount spent in Campaign	Overview
Evaluation	Data sources	Data sources explored by the study	RQ3/QA
	Data source category	Category that best characterizes the data source	RQ3/QA
	Variables of interest	Variables of interest explored by the study	RQ3/QA
	Dimension	Area related to the variable of interest identified	RQ3
	Tools or technologies	Techniques, methods, tools or other solutions implemented	RQ4/QA
	Results	Main results reported	Overview

the determined criteria. Accordingly, 17 studies (five identified via database plus 12 identified from snowball technique) were selected for data extraction (listed in **Appendix A**).

Study Quality Assessment

Quality assessment is essential in systematic reviews to determine the rigor and relevance of the primary studies and should be applied in a similar way across the different types of studies identified (21).

To assess the quality of the studies, we developed a survey on six characteristics of communication campaigns, **Table 1** (3rd to 8th rows). In addition, we only consider studies classified as Evaluation Research, as defined by Wieringa et al. (22) that is: techniques, methods, tools or other solutions are implemented and evaluated in practice.

For each question, the study's quality was evaluated as "Yes," "Partially," or "No," and scored with the values 1, 0.5, and 0, respectively. We noted that the overall quality of the studies was good, ranging from 0.75 to 1. **Table 2** shows the quality assessment criteria found for the 17 selected studies. In **Appendix A**, these data are listed in detail.

Data Extraction

To extract data from the identified primary studies, we developed a template (**Table 3**) to register the main information in a spreadsheet. Each data item addresses a Research Question (RQ), Quality Assessment (QA) or an overview of the study.

Threats to Validity

The ability to choose digital libraries that will accurately represent a study can influence the validity of an SLR. Scopus has one of the largest abstracts and citation databases for research literature. Researchers consider it a high-quality, versatile, and respected research database (19).

This review created more restrictive filters for the area (Science Computer) and source type (Journals), to avoid an exhaustive review. Accordingly, some studies–although relevant–may not be included. We mitigated this limitation by scanning the reference list through a backward snowballing technique (20) of identified papers. This technique added 225 papers for the analysis of titles and abstracts, of which 19 were included in the final selection.

Regarding data extraction, the process of selection and classification of the studies was done by one researcher and all data extracted was checked by the other, and subsequently rechecked by the first researcher. At the end, we discussed the results and when doubts remained, a third researcher was consulted, to make the final decision.

RESULTS

In total, 276 papers were found by the automatic and snowball searches. However, after applying the inclusion/exclusion criteria, 17 papers remained in the set of relevant papers. We believe that these 17 papers present the necessary criteria to answer the research questions.

RQ1: When and where have the studies been published? Historically, the first publication was in 2003. Then, in 2008. After 3 years with no results found, publications returned in 2011, where there were annual publications until 2019. The countries carrying out this research were: Australia ($n = 6$), United States ($n = 6$), United Kingdom ($n = 3$), China ($n = 1$), and Norway ($n = 1$). Most studies were published in health and communication journals as shown in **Table 4**.

RQ2: Which topic areas were addressed to the campaign, location, and year? In the 17 selected studies, the topic areas addressed were Anti-Smoking [$n = 5$, (S03, S06, S07, S09, S10)], Children Vaccination [$n = 1$, (S01)], Newborn screening and biobanking programs [$n = 1$, (S14)], Overweight and Obesity [$n = 3$, (S05, S16, S17)], and Sexual Health Care [$n = 7$, (S02, S04, S08, S11, S12, S13, S15)].

Figure 2 shows the distribution of thematic areas targeted to the campaign in the period from 1995 to 2018. Each rectangle covers a period on a campaign topic by country. Each

TABLE 4 | Sources identified in primary studies, ordered by year.

Year	Author country	Venue name	Cite	Study
2003	United States	Health communication	(23)	S07
2008	Australia	American journal of public health	(24)	S06
2011	Australia	Sexual health	(25)	S15
2012	United States	American journal of preventive medicine	(26)	S08
2013	Australia	Sexual health	(27)	S13
2014	United Kingdom	Journal of medical internet research	(28)	S11
2014	United States	Journal of communication	(29)	S09
2015	United Kingdom	Data mining and knowledge discovery	(30)	S01
2015	United States	AIDS and Behavior	(31)	S12
2016	United States	JMIR public health surveill	(32)	S14
2016	Australia	Journal of health communication	(33)	S16
2017	United States	Tobacco control	(34)	S10
2017	Norway	International journal of e-health and medical communications	(35)	S02
2017	China	IEEE transactions on nanobioscience	(36)	S03
2018	Australia	Australian and New Zealand journal of public health	(37)	S17
2018	Australia	Social media and society	(38)	S05
2019	United Kingdom	Digital health	(39)	S04

country is designated by a color. Australia (green) covered long-term campaigns in the Anti-Smoking, Overweight and Obesity, and Sexual Health Care areas. The United States (blue) had campaigns in the areas of Anti-smoking, Newborn screening and biobanking programs, and Sexual Health Care. England (orange) in Sexual Health Care (Chlamydia) and Children Vaccination. Norway (yellow) had a short-term campaign in the Sexual Health Care area.

When the dates were not clearly specified in the article, an internet search was carried out to find approximate values. For example: In S01, influenza season (2013/2014) meant the period from 02/09/2013 to 13/04/2014 (It was dominated by the circulation of influenza A(H1N1) pdm09 virus, in Europe). When the articles specified only the month and year of the campaign, we considered the first (starting) and last (end) day of the month. One study (S07) did not report a period and due to this it was not shown in the **Figure 2**.

RQ3: What variables of interest were analyzed to assess the impact of public health campaigns? The 17 studies evaluated several types of variables, such as: amount of access to the campaign website, awareness and knowledge of the subject through the questionnaires, interest of the population through the online search engine, engagement through the SNS, and advertising exposure data of television commercials.

Most studies sought to cross-check the data obtained with socioeconomic and census data. The data source used to obtain the variables can be categorized as follows: questionnaire ($n = 10$), social network service ($n = 10$), Sexually Transmitted Infection (STI) testing ($n = 2$), television commercials ($n = 2$), campaign website ($n = 1$), pharmaceutical products ($n = 1$), smoke-free restaurant laws ($n = 1$), sociodemographic data ($n = 1$), tobacco prices ($n = 1$). **Table 5** shows data sources, variables of interest, dimensions, and methods used by each study.

Despite the variety of variables analyzed, we can observe a predominance related to i) assessing user engagement with the campaign subject through SNS, such as Facebook and Twitter (S01, S02, S03, S04, S08, S09, S10, S11, S12, S14); and ii) assessing public knowledge, attitudes and behavior through questionnaires (S04, S05, S06, S07, S08, S12, S13, S15, S16, S17), before and/or after campaign dissemination.

SNS were used to recruit participants to answer the questionnaires (S05, S08, S13), assessing the reach and knowledge of the impacted people. The authors argued that this approach produced good results due to accurate user geolocation and demographic data (e.g., gender, age, education, employment status, language spoken).

In addition, some studies (S02, S04, S05, S08, S12, S14) highlighted that Facebook advertisements significantly helped to increase both the number of people visiting a campaign webpage, and the number of users who could potentially improve their knowledge and health care behavior, because of the campaign.

Televised ads were assessed in two anti-tobacco campaigns using broadcasting time and its Nielsen rating (https://www.nielsen.com/ca/en/about-us/). S03 used TV ratings correlated with social media (Twitter) while S06 correlated TV ratings with questionnaire, tobacco prices, population use of pharmaceutical smoking cessation products, and smoke-free restaurant laws. In both cases, the results indicated success in the approaches used to assess the campaign impact.

Regarding STI Testing, S04 and S12 indicate that their campaign achieved its goal, i.e., to change sex care behavior and reach a large number of people. In fact, in the post-campaign period, the demand for test orders increased.

Grouping the variables of interest in dimensions (areas of context) provided better understanding about which dimension was most explored. Our findings show three dimensions:

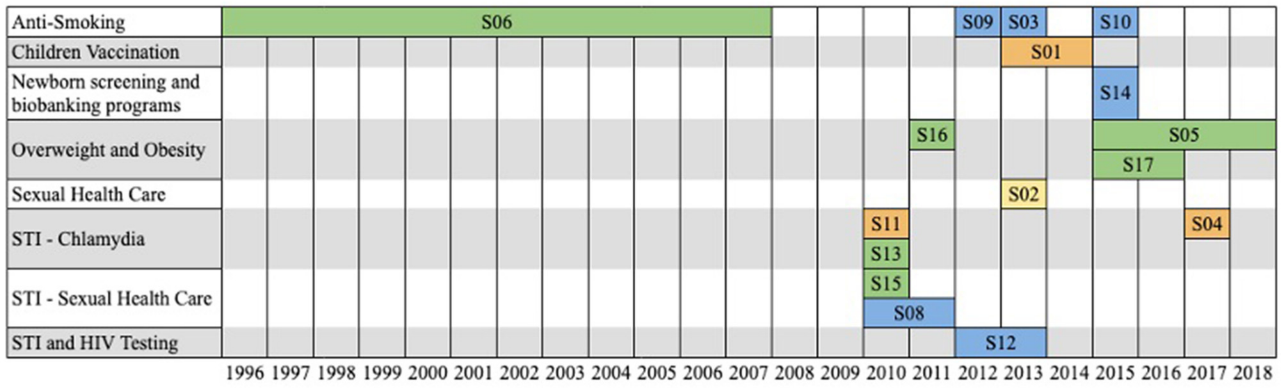


FIGURE 2 | Distribution of topic areas addressed to campaign by range of period.

communication ($n = 20$), epidemiology ($n = 10$), and public policy ($n = 2$).

Communication dimension was assigned when the variable of interest sought to assess engagement on the Internet/social media or surveys asked the user knowledge about the health campaign. Epidemiology dimension was assigned when the variable of interest sought to assess patterns of health and disease conditions in a defined population. Public policy dimension was detected when the variable of interest sought to assess control policies, such as changes in taxes on products, and new laws.

RQ4: What techniques or tools have been used to support campaign impact analysis? Considering that social networking services (Facebook, Twitter, and Instagram) were widely explored in the studies, the main tools were the dashboards of the respective sites (explored by S02, S04, S08, S11, S12, S14). SNS dashboards often provide enough data to verify the engagement of privately hosted websites.

Nevertheless, analyzing user-generated content on a social network, such as Twitter, and inferring correlations with campaign data requires a detailed evaluation of the content. Accordingly, the studies S03, S09, and S10 obtained data from Twitter through a licensed data provider (http://www.gnip.com). This vendor provides real-time access to 100% of all tweets and meta-data. These analyses and correlations were supported by general linear models used to assess the relationship between demographics data and campaign awareness. Furthermore, Google analytics was used to identify the number of unique users, total visits, page views, unique views, and average visit duration for each page of the campaign website.

Regarding software for statistical analysis and using data science to explore, visualize, model and make inferences in the data, studies S12, S13 and S16 used Stata (https://www.stata.com/) while studies S06, S08, and S17 used Statistical Analysis System–SAS (https://www.sas.com/enca/software/stat.html). NVivo (https://www.qsrinternational.com/nvivo/home), a qualitative data analysis software, was used by S05 for textual analysis of the questionnaires and by S11 to capture the campaign page content, using a simple counting method to describe user and moderator content, discussion thread patterns, and moderator intervention. S12 used a web-based survey administration and data management application called REDCap (https://www.project-redcap.org/).

In S05, all statistics analyzes were conducted with SPSS Statistics 22 (https://www.ibm.com/support/pages/spss-statistics-220-available-download), a software platform that offers advanced statistical analysis, a vast library of machine learning algorithms, text analysis, open-source extensibility, integration with big data and seamless deployment into applications.

S01 proposed a statistical framework for transforming user-generated content published on web platforms to an assessment of the impact of a health-oriented intervention and S07 used a statistical technique named ANOVA (analysis of variance) to test all variables defined in the study.

DISCUSSION

This paper assessed how researchers evaluate public health campaigns. Based on the studies found and the answers to the research questions, we observed that the studies sought to assess campaign impact through quantitative and qualitative analysis, using mainly SNS, especially Facebook. The popularity of social platforms such as Facebook–in the United States and worldwide–serves to justify this approach (40, 41). Facebook ads can reach and engage large, well-specified populations at relatively low cost (42, 43).

In addition, SNS are a rich source of data that can provide responses to PHC impact assessments. This may be because data is available to users through the dashboards, where it is possible to observe quantitative values. Detailed data can be easily obtained from some vendors, by paying fees, which facilitates data collection, allowing such methods and tools to do the "heavy lifting" (sometimes paid, such as Stata, SAS, NVivo or SPSS). A potential threat related to this approach is to identify real engagement of users. Some studies reported this issue (S4, S5, S8, S11, S14), which raised the following question: What is the real motivation to engage on campaign interventions pages on social networks? What about the "spiral of the silence" (44)

TABLE 5 | Data sources and variables of interest analyzed per study.

Paper	Data source	Variables of interest	Dimension	Method used
S01	Search engine	Bing search queries geo-located in the target vaccinated locations.	Communication	The authors presented a statistical framework for estimating the prevalence of an intervention campaign in the population from Internet data.
	Social network service	Number of Twitter postings.	Communication	
S02	Social network service	Number of visitors/visits, time spent.	Communication	The authors analyzed the impact of a Facebook fan page, a Facebook advertisement campaign, and posters through Facebook statistics dashboard and Google Analytics.
S03	Television commercials	Tweets related to the televised ads.	Communication	The authors presented a statistical framework (Advertising Social Influence Estimation-ASIE) which predict the probability of users posting tweets influenced by both TV broadcasting and friends in the online social network.
	Social network service		Communication	
S04	Questionnaire	Questions about attitudes toward sexual health promotion on social media, preferences for the content of promotional campaigns, and potential barriers to engagement with social media health promotion.	Epidemiology	The authors analyzed qualitative face-to-face interviews, and the engagement with chlamydia testing page through Facebook statistics dashboard and Google Analytics, as well as descriptive statistics to assess amount of ? chlamydia tests requested during the intervention period.
	Social network service	Number of visitors/visits and online actions.	Communication	
	STI testing	Number of chlamydia tests requested during the campaign period.	Epidemiology	
S05	Questionnaire	Questions about the experience and satisfaction with a health campaign in Facebook. Participants' self-reported postcode, and height and weight.	Communication	The authors conducted statistical tests to compare Facebook users' groups and explored its profile to examine the characteristics of fans of the page, as well as analyzed an online survey to investigate how users were interacting with the campaign page and others health pages on Facebook.
S06	Questionnaire	Smoking prevalence was estimated from survey.	Epidemiology	The group used time-series autoregressive integrated moving average analysis in a statistical software (SAS) to estimate the effect of antitobacco advertising and tobacco policies on monthly smoking prevalence.
	Television commercials	Occurrences of all tobacco-related advertisements appearing on television.	Communication	
	Tobacco prices	Cigarette costliness was measured with the ratio of the average recommended retail price per cigarette pack to the average weekly earnings.	Public policy	
	Pharmaceutical products	Population use of pharmaceutical smoking cessation products.	Epidemiology	
	Smoke-free restaurant laws	Population exposure to smoke-free laws was expressed as the percentage of the total sample that was subject to such laws.	Public policy	
S07	Questionnaire	Questions about the participants perception related to campaign messages and their behavioral intent post- campaign.	Communication	The authors conducted statistical tests to analyze three primary variables of interest (source evaluation, message evaluation, and behavioral intention), based on the theory of psychological reactance.
S08	Questionnaire	Questions about sexual health care behavior.	Epidemiology	The authors analyzed data related to the engagement with the Facebook intervention page through Google Analytics and an online survey, using a statistical software (SAS).
	Social network service	Engagement with the Facebook page.	Communication	
S09	Social network service	Message acceptance, rejection, and disregard from each tweet identified as Tips-relevant.	Communication	The authors present an analysis of Twitter messages about a health campaign using an analytic framework.

(Continued)

TABLE 5 | Continued

Paper	Data source	Variables of interest	Dimension	Method used
S10	Social network service	Messages were labeled as anti, pro, or neutral campaign.	Communication	The authors analyzed the content of Twitter messages about a health campaign through statistical analysis.
S11	Social network service	Volume of interaction. The total number of fans, wall posts, and comments over time, fan demographics. Website access numbers and viewing patterns.	Communication	The authors examined quantitative and qualitative data on a Facebook page about a health campaign. Google analytics was used to describe the number of people using the page and viewing patterns.
S12	Social network service	Data related to numbers of "likes," visits, and number of "followers."	Communication	The group used standard descriptive statistics to assess a health campaign by: tracking website/social media use, online survey, and comparing rates of STI testing.
	Campaign website	Data related to the engagement on campaign website.	Communication	
	Questionnaire	Information on age, how they heard about the campaign, assessed knowledge of STIs, and if the campaign influenced intention to get tested.	Communication and Epidemiology	
	STI testing	Checking STI Testing Pre- and Post-Campaign.	Epidemiology	
S13	Questionnaire	Questions about demographic variables, height and weight, how they found out about the study, sexual history, experience and knowledge of STIs.	Epidemiology	The group assessed the feasibility of using SNSs to recruit young women to complete a health-related survey.? Data were analyzed using a statistical software (STATA).
S14	Social network service	Engagement with the Facebook: Page likes, views, posts and photo album engagement (likes, connections, shares, conversation, and comments).	Communication	The authors presented a framework for examining a spectrum of Facebook engagement outcomes from observation to conversation. Data were provided by Facebook dashboard.
S15	Questionnaire	Number of clients that contacted the service after ads exposure and ads type used.	Communication	The authors presented a simple descriptive study evaluating the effectiveness of different advertising methods.
S16	Questionnaire	Questions about demographic variables, risk factors, if the respondent was aware of the "Swap It" campaign, attitudes and behaviors regarding diet and exercise, and behavioral intentions and actions.	Communication and epidemiology	The authors evaluated a health campaign via cross-sectional serial telephone surveys. Data were analyzed using a statistical software (STATA).
S17	Questionnaire	Questions about campaign awareness, knowledge, attitudes, and intentions. Current behavior and recent behavior change. Demographic variables, body mass index category and risk index score.	Communication and Epidemiology	The authors presented a cohort design study and sed generalized linear mixed models in a statistical software (SAS) to examine campaign awareness, knowledge, attitudes, intentions, and behaviors over time

which suggests that interventions can change the attitudes and behaviors of real interested users behind the screen, even without their interaction?

It is intriguing to consider the paradox of trying to reach a target audience through the SNS to raise awareness about an infectious disease, since some parts of the population cannot be reached through technologies and the Internet. In Brazil, for example, internet use reached 152 million people, representing 81% of the population. This percentage drops to 67% when we looked at the lower social classes (45). However, at the same time, it is likely that no communication channel will be able to reach 100% of the target audience by itself. The technological bias present in many campaigns is clear when only digital channels are activated, mainly excluding lower social classes, homeless people and people deprived of liberty, who in turn are part of the target audience of this type of campaign.

In addition, questionnaires targeting impacted users was another effective form of evaluating campaign impact. This approach is useful for measuring target audience awareness, attitudes and behavior, and has been used in eight out of 17 primary studies.

In contrast to most of the studies analyzed, article S15 reported that advertisements through paper-based methods, text messaging and social networking sites were not effective. In this study, only 28 rural youth contacted the health care service (campaign's goal) over the 11-month period. Twenty young people were reached by nurses (15 from school nurses, five from community health nurses), six through the campaign webpage, one through Facebook and one through the student diary. No clients were recruited through other advertising methods. Arguably, the limited publicity could be because the target audience were young people living in rural areas in Victoria, Australia. Therefore, although SNS have a great potential to provide data to analyze the impact of campaigns, attention must be paid to the public and the conditions of access and web-behavior of that public.

Despite the widespread use of social networking sites, only one study (S01) sought a second source of data based on user-generated Internet content (Microsoft's Bing search engine). The study aimed to introduce a complementary framework for evaluating the impact of a targeted intervention, such as a vaccination campaign against an infectious disease, through statistical analysis of user-generated content submitted on web platforms (Twitter and Bing search engine). The results from Twitter data demonstrated less sensitivity across similar controls relative to Bing data, suggesting a greater reliability.

In general, we observed a gap in assessing the impact of public health campaigns, regarding the use of online data (i.e., online news) and others user-generated Internet content. Google Search and Yahoo should be used in addition to Bing search engine to check the increase of news related to the campaign. In addition, a variety of questions should be asked, such as what can the analysis of this content bring? Are campaigns helping to grow spontaneous news on the related topic? When evaluating a population over time, Google Trends is arguably an excellent tool to demonstrate how the population has sought to know about a certain topic (46–48). However, it was not explored in the context of these works.

Moreover, databases of scientific papers such as Springer Nature, Wiley Blackwell, Taylor and Francis, IEEE, American Physical Science and Elsevier and its indexers such as Scopus, Web of Science (WoS) and Google Scholar could be used to demonstrate the interest of the academy in developing research on the campaign theme in a comparison of time (before, during and after the campaign). These variables of interest can show a new dimension, Education.

Health campaigns are common worldwide. Nevertheless, few studies have reported concrete campaign data. We would like to encourage future research to better specify analyzed data in their studies showing properties raised by Dorfman et al. (1) and complemented for us, to enable a comparison on health campaigns analysis, as present in **Appendix A**, stimulating the Campaign dimension.

Notably, by using variables of interest that were grouped into three dimensions (as discussed in this review), we analyzed the reach of the "Syphilis No!" campaign (49), assessing data related to the campaign, online news, search engine activity, online courses, serological tests, medication distribution and case notification rates. Results of this analysis show positive changes over time in communication, education, and epidemiological surveillance dimensions, especially after the campaign propagation.

CONCLUSIONS

The paper provided an overview of studies that evaluate the impact of public health campaigns. The analysis focused on identifying variables of interest, techniques and tools used. We discussed the results, presented new questions, and discovered unexplored variables of interest.

Public health campaigns play a strategic role promoting awareness, increasing knowledge, and encouraging the target population to adopt desirable attitudes and behaviors. As observed, its impact must be measured in several dimensions such as: i) communication (engagement in social networks, questionnaires assessing the user's knowledge about the campaign), ii) epidemiology (disease screening test, cases' notification, questionnaires assessing the user's knowledge about the disease/health issues), and iii) policy enforcement (law strategies on health promotion).

This multidimensional analysis provided a complete evaluation of public health campaigns, to understand its scope, find correlations between different variables of interest and expand possibilities for analysis.

Accordingly, we hope this work will motivate future researchers to explore other variables of interest and dimensions that are possibly overlooked in assessing the impact of public health campaigns based on multidimensional aspects.

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

RP, LS, and VK: conceptualization. RP, LS, RV, and VK: methodology. RP, LS, RV, VK, CG, CO, and JL: validation and writing–review and editing. RP and LS: formal analysis. RP: investigation, data curation, and writing–original draft preparation. LS, RV, and VK: supervision. RV: project administration and funding acquisition. All authors contributed to the article and approved the submitted version.

FUNDING

Brazilian Ministry of Health–Number's project: 54/2017.

ACKNOWLEDGMENTS

This research was carried out in cooperation agreement between Federal University of Rio Grande do Norte and Athabasca University. We express our sincere gratitude to all leaders that was involved with this agreement.

SUPPLEMENTARY MATERIAL

The Supplementary Material for this article can be found online at: https://www.frontiersin.org/articles/10.3389/fpubh.2021.715403/full#supplementary-material

REFERENCES

1. Dorfman L, Ervice J, Woodruff K. *Voices For Change: A Taxonomy of Public Communications Campaigns and Their Evaluation Challenges*. Washington, DC: Communications Consortium Media Center, Media Evaluation Project (2002).
2. Bucchi M, Trench B. Health campaign research: Enduring challenges and new developments. In: *Routledge handbook of public communication of science and technology*. Routledge (2014). p. 214–29.
3. Freudenberg N. Public health advocacy to change corporate practices: implications for health education practice and research. *Health Educ Behav.* (2005) 32:298–319. doi: 10.1177/1090198105275044
4. Kitchenham B, Charters S. *Guidelines For Performing Systematic Literature Reviews in Software Engineering*. Keele; Keele University (2007).
5. Petersen K, Feldt R, Mujtaba S, Mattsson M. Systematic mapping studies in software engineering. *EASE*. (2008) 8:68–77. doi: 10.14236/ewic/EASE2008.8
6. Noar SM. A 10-year retrospective of research in health mass media campaigns: where do we go from here? *J Health Commun.* (2006) 11:21–42. doi: 10.1080/10810730500461059
7. Te V, Ford P, Schubert L. Exploring social media campaigns against sugar-sweetened beverage consumption: a systematic search. *Cogent Medicine.* (2019) 6:1607432. doi: 10.1080/2331205X.2019.1607432
8. Jones J, Salazar LF. A review of hiv prevention studies that use social networking sites: implications for recruitment, health promotion campaigns, and efficacy trials. *AIDS Behav.* (2016) 20:2772–81. doi: 10.1007/s10461-016-1342-9
9. Yadav RP, Kobayashi M. A systematic review: effectiveness of mass media campaigns for reducing alcohol-impaired driving and alcohol-related crashes. *BMC Public Health.* (2015) 15:857. doi: 10.1186/s12889-015-2088-4
10. Vega MY, Roland EL. Social marketing techniques for public health communication: a review of syphilis awareness campaigns in 8 US cities. *Sex Transm Dis.* (2005) 32:S30–6. doi: 10.1097/01.olq.0000180461.30725.f4
11. Shi J, Poorisat T, Salmon CT. The use of social networking sites (SNSS) in health communication campaigns: review and recommendations. *Health Commun.* (2018) 33:49–56. doi: 10.1080/10410236.2016.1242035
12. Jacob V, Chattopadhyay SK, Elder RW, Robinson MN, Tansil KA, Soler RE, et al. Economics of mass media health campaigns with health-related product distribution: a community guide systematic review. *Am J Prev Med.* (2014) 47:348–59. doi: 10.1016/j.amepre.2014.05.031
13. Robinson MN, Tansil KA, Elder RW, Soler RE, Labre MP, Mercer SL, et al. Mass media health communication campaigns combined with health-related product distribution: a community guide systematic review. *Am J Prev Med.* (2014) 47:360–71. doi: 10.1016/j.amepre.2014.05.034
14. Randolph KA, Whitaker P, Arellano A. The unique effects of environmental strategies in health promotion campaigns: a review. *Eval Program Plann.* (2012) 35:344–53. doi: 10.1016/j.evalprogplan.2011.12.004
15. Moher D, Liberati A, Tetzlaff J, Altman DG, Group P. Preferred reporting items for systematic reviews and meta-analyses: the prisma statement. *PLoS Med.* (2009) 6:e1000097. doi: 10.1371/journal.pmed.1000097
16. Moher D, Shamseer L, Clarke M, Ghersi D, Liberati A, Petticrew M, et al. Preferred reporting items for systematic review and meta-analysis protocols (prisma-p) 2015 statement. *Syst Rev.* (2015) 4:1–9. doi: 10.1186/2046-4053-4-1
17. Baas J, Schotten M, Plume A, Côté G, Karimi R. Scopus as a curated, high-quality bibliometric data source for academic research in quantitative science studies. *Quantitative Science Studies.* (2020) 1:377–86. doi: 10.1162/qss_a_00019
18. Harzing AW, Alakangas S. Google scholar, scopus and the web of science: a longitudinal and cross-disciplinary comparison. *Scientometrics.* (2016) 106:787–804. doi: 10.1007/s11192-015-1798-9
19. Mongeon P, Paul-Hus A. The journal coverage of web of science and scopus: a comparative analysis. *Scientometrics.* (2016) 106:213–28. doi: 10.1007/s11192-015-1765-5
20. Greenhalgh T, Peacock R. Effectiveness and efficiency of search methods in systematic reviews of complex evidence: audit of primary sources. *BMJ.* (2005) 331:1064–5. doi: 10.1136/bmj.38636.593461.68
21. Wohlin C, Runeson P, Neto PAdMS, Engström E, do Carmo Machado I, De Almeida ES. On the reliability of mapping studies in software engineering. *J Syst Softw*. (2013) 86:2594–610. doi: 10.1016/j.jss.2013.04.076
22. Wieringa R, Maiden N, Mead N, Rolland C. Requirements engineering paper classification and evaluation criteria: a proposal and a discussion. *Requir Eng.* (2006) 11:102–7. doi: 10.1007/s00766-005-0021-6
23. Grandpre J, Alvaro EM, Burgoon M, Miller CH, Hall JR. Adolescent reactance and anti-smoking campaigns: a theoretical approach. *Health Commun.* (2003) 15:349–66. doi: 10.1207/S15327027HC1503_6
24. Wakefield MA, Durkin S, Spittal MJ, Siahpush M, Scollo M, Simpson JA, et al. Impact of tobacco control policies and mass media campaigns on monthly adult smoking prevalence. *Am J Public Health.* (2008) 98:1443–50. doi: 10.2105/AJPH.2007.128991
25. Gamage DG, Fuller CA, Cummings R, Tomnay JE, Chung M, Chen M, et al. Advertising sexual health services that provide sexually transmissible infection screening for rural young people–what works and what doesn't. *Sex Health.* (2011) 8:407–11. doi: 10.1071/SH10144
26. Bull SS, Levine DK, Black SR, Schmiege SJ, Santelli J. Social media–delivered sexual health intervention: a cluster randomized controlled trial. *Am J Prev Med.* (2012) 43:467–74. doi: 10.1016/j.amepre.2012.07.022
27. Ahmed N, Jayasinghe Y, Wark JD, Fenner Y, Moore EE, Tabrizi SN, et al. Attitudes to chlamydia screening elicited using the social networking site facebook for subject recruitment. *Sex Health.* (2013) 10:224–8. doi: 10.1071/SH12198
28. Syred J, Naidoo C, Woodhall SC, Baraitser P. Would you tell everyone this? Facebook conversations as health promotion interventions. *J Med Internet Res.* (2014) 16:e108. doi: 10.2196/jmir.3231
29. Emery SL, Szczypka G, Abril EP, Kim Y, Vera L. Are you scared yet? Evaluating fear appeal messages in tweets about the tips campaign. *J Commun.* (2014) 64:278–95. doi: 10.1111/jcom.12083
30. Lampos V, Yom-Tov E, Pebody R, Cox IJ. Assessing the impact of a health intervention via user- generated internet content. *Data Min Knowl Discov.* (2015) 29:1434–57. doi: 10.1007/s10618-015-0427-9
31. Dowshen N, Lee S, Lehman BM, Castillo M, Mollen C. Iknowushould2: Feasibility of a youth-driven social media campaign to promote STI and HIV testing among adolescents in Philadelphia. *AIDS Behav.* (2015) 19:106–11. doi: 10.1007/s10461-014-0991-9
32. Platt T, Platt J, Thiel DB, Kardia SL. Facebook advertising across an engagement spectrum: a case example for public health communication. *JMIR Public Health Surveill.* (2016) 2:e27. doi: 10.2196/publichealth.5623
33. O'Hara BJ, Grunseit A, Phongsavan P, Bellew W, Briggs M, Bauman AE. Impact of the swap it, don't stop it Australian national mass media campaign on promoting small changes to lifestyle behaviors. *J Health Commun.* (2016) 21:1276–85. doi: 10.1080/10810730.2016.1245803
34. Allem JP, Escobedo P, Chu KH, Soto DW, Cruz TB, Unger JB. Campaigns and counter campaigns: reactions on twitter to e-cigarette education. *Tob Control.* (2017) 26:226–9. doi: 10.1136/tobaccocontrol-2015-052757
35. Gabarron E, Luque LF, Schopf TR, Lau AY, Armayones M, Wynn R, et al. Impact of facebook ads for sexual health promotion via an educational web app: a case study. *Healthcare Policy and Reform.* (2019) 990–1003. doi: 10.4018/978-1-5225-6915-2.ch045
36. Zhan Q, Zhang J, Philip SY, Emery S, Xie J. Inferring social influence of anti-tobacco mass media campaign. *IEEE Trans Nanobioscience.* (2017) 16:356–66. doi: 10.1109/TNB.2017.2707075
37. Kite J, Gale J, Grunseit A, Bellew W, Li V, Lloyd B, et al. Impact of the make healthy normal mass media campaign (phase 1) on knowledge, attitudes and behaviours: a cohort study. *Aust N Z J Public Health.* (2018) 42:269–76. doi: 10.1111/1753-6405.12779
38. Kite J, McGill B, Freeman B, Vineburg J, Li V, Berton N, et al. User perceptions of the make healthy normal campaign facebook page: a mixed methods study. *Soc Media Soc.* (2018) 4:1–10. doi: 10.1177/20563051187 94639
39. Nadarzynski T, Burton J, Henderson K, Zimmerman D, Hill O, Graham C. Targeted advertisement of chlamydia screening on social media: a mixed-methods analysis. *Digital Health.* (2019) 5:1–10. doi: 10.1177/2055207619827193
40. Capurro D, Cole K, Echavarr ia MI, Joe J, Neogi T, Turner AM. The use of social networking sites for public health practice and research: a systematic review. *J Med Internet Res.* (2014) 16:e79. doi: 10.2196/jmir.2679
41. Madden M, Lenhart A, Cortesi S, Gasser U. *Pew Internet and American Life Project*. Washington, DC: Pew Research Center (2010).

42. Platt J, Platt T, Thiel D, Kardia S. 'Born in michigan? you're in the biobank': Engaging population biobank participants through facebook advertisements. *Public Health Genomics.* (2013) 16:145–58. doi: 10.1159/000351451

43. Ramo DE, Prochaska JJ. Broad reach and targeted recruitment using facebook for an online survey of young adult substance use. *J Med Internet Res.* (2012) 14:e28. doi: 10.2196/jmir.1878

44. Hampton KN, Rainie H, Lu W, Dwyer M, Shin I, Purcell K. *Social Media and the "Spiral of Silence" (PewResearchCenter).* Washington, DC: Pew Research Center (2014). Available online at: https://www.pewresearch.org/internet/2014/08/26/social-media-and-the-spiral-of-silence/

45. UOL T. *Brasil chega a 152 mi de usuários de internet; idosos estão mais conectados.* (2021). Available online at: https://www.uol.com.br/tilt/noticias/redacao/2021/08/18/tic-domicilios-2020-idosos-usaram-mais-internet-uso-de-smart-tv-cresceu.htm (accessed November 20, 2021).

46. Dreher PC, Tong C, Ghiraldi E, Friedlander JI. Use of google trends to track online behavior and interest in kidney stone surgery. *Urology.* (2018) 121:74–8. doi: 10.1016/j.urology.2018.05.040

47. Glynn RW, Kelly JC, Coffey N, Sweeney KJ, Kerin MJ. The effect of breast cancer awareness month on internet search activity-a comparison with awareness campaigns for lung and prostate cancer. *BMC Cancer.* (2011) 11:442. doi: 10.1186/1471-2407-11-442

48. Ling R, Lee J. Disease monitoring and health campaign evaluation using google search activities for HIV and aids, stroke, colorectal cancer, and marijuana use in Canada: a retrospective observational study. *JMIR Public Health Surveill.* (2016) 2:e156. doi: 10.2196/publichealth.6504

49. de Morais Pinto R, de Medeiros Valentim RA, Fernandes da Silva L, Lima GFdMS, Kumar V, Pereira de Oliveira CA, et al. Analyzing the reach of public health campaigns based on multidimensional aspects: the case of the syphilis epidemic in brazil. *BMC Public Health.* (2021) 21:1–13. doi: 10.1186/s12889-021-11588-w

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.

TYPE Original Research
PUBLISHED 09 September 2022
DOI 10.3389/fninf.2022.771965

OPEN ACCESS

EDITED BY
Giovanni Pellegrino,
McGill University Health Centre,
Canada

REVIEWED BY
Ailiang Miao,
Nanjing Brain Hospital Affiliated to
Nanjing Medical University, China
Stavros I. Dimitriadis,
Integrative Neuroimaging Lab, Greece
Pradeep Modur,
University of Texas at Austin,
United States

*CORRESPONDENCE
Xueping Peng
xueping.peng@uts.edu.au
Ting Wu
lvwu123@163.com
Jiayang Guo
guojy@xmu.edu.cn
Yijie Pan
pyj20@mails.tsinghua.edu.cn

RECEIVED 07 September 2021
ACCEPTED 03 August 2022
PUBLISHED 09 September 2022

CITATION
Zhao X, Peng X, Niu K, Li H, He L, Yang F, Wu T, Chen D, Zhang Q, Ouyang M, Guo J and Pan Y (2022) A multi-head self-attention deep learning approach for detection and recommendation of neuromagnetic high frequency oscillations in epilepsy. *Front. Neuroinform.* 16:771965. doi: 10.3389/fninf.2022.771965

A multi-head self-attention deep learning approach for detection and recommendation of neuromagnetic high frequency oscillations in epilepsy

Xiangyu Zhao[1,2], Xueping Peng[3]*, Ke Niu[4], Hailong Li[5], Lili He[5], Feng Yang[1], Ting Wu[6,7]*, Duo Chen[8], Qiusi Zhang[1], Menglin Ouyang[9], Jiayang Guo[10,11]* and Yijie Pan[12,13]*

[1]Information Technology Research Center, Beijing Academy of Agriculture and Forestry Sciences, Beijing, China, [2]National Engineering Research Center for Information Technology in Agriculture, Beijing, China, [3]Australian Artificial Intelligence Institute, Faculty of Engineering and Information Technology, University of Technology Sydney, Ultimo, NSW, Australia, [4]Computer School, Beijing Information Science and Technology University, Beijing, China, [5]Department of Radiology, Imaging Research Center, Cincinnati Children's Hospital Medical Center, Cincinnati, OH, United States, [6]Department of Radiology, Jiangsu Province Hospital of Chinese Medicine, Affiliated Hospital of Nanjing University of Chinese Medicine, Nanjing, China, [7]Department of Magnetoencephalography, Nanjing Brain Hospital, Affiliated to Nanjing Medical University, Nanjing, China, [8]School of Artificial Intelligence and Information Technology, Nanjing University of Chinese Medicine, Nanjing, China, [9]The Affiliated Hospital of Medical School, Ningbo University, Ningbo, China, [10]National Institute for Data Science in Health and Medicine, Xiamen University, Xiamen, China, [11]Department of Hematology, School of Medicine, Xiamen University, Xiamen, China, [12]Department of Computer Science and Technology, Tsinghua University, Beijing, China, [13]Ningbo Institute of Information Technology Application, Chinese Academy of Sciences, Ningbo, China

Magnetoencephalography is a noninvasive neuromagnetic technology to record epileptic activities for the pre-operative localization of epileptogenic zones, which has received increasing attention in the diagnosis and surgery of epilepsy. As reported by recent studies, pathological high frequency oscillations (HFOs), when utilized as a biomarker to localize the epileptogenic zones, result in a significant reduction in seizure frequency, even seizure elimination in around 80% of cases. Thus, objective, rapid, and automatic detection and recommendation of HFOs are highly desirable for clinicians to alleviate the burden of reviewing a large amount of MEG data from a given patient. Despite the advantage, the performance of existing HFOs rarely satisfies the clinical requirement. Consequently, no HFOs have been successfully applied to real clinical applications so far. In this work, we propose a multi-head self-attention-based detector for recommendation, termed MSADR, to detect and recommend HFO signals. Taking advantage of the state-of-the-art multi-head self-attention mechanism in deep learning, the proposed MSADR achieves a more superior accuracy of 88.6% than peer machine learning models in both detection and recommendation tasks. In addition, the robustness of MSADR is also extensively assessed with various ablation tests, results of which further demonstrate the effectiveness and generalizability of the proposed approach.

KEYWORDS

high frequency oscillations (HFOs), magnetoencephalography, MEG, deep learning, multi-head self-attention, HFOs detection, HFOs recommendation

1. Introduction

About 30% of pediatric patients with epilepsy are medically intractable and require respective neurosurgery to gain seizure freedom (Durnford et al., 2011; Yamakawa et al., 2020). Recording epileptic activities are crucial to the pre-operative localization of epileptogenic zones and the optimization of the diagnosis of epilepsy. The success of epilepsy surgery depends on the pre-operative localization of epileptogenic zones (Guo et al., 2018). Although intracranial electroencephalography (iEEG) is commonly treated as the gold standard for the localization of epileptogenic zones, it may bring a risk of infection and bleeding during implantation (Hu et al., 2015). Thus, a noninvasive detection method for epileptogenic zones is preferred to epilepsy surgery. Magnetoencephalography (MEG) is a noninvasive technology for the detection of epileptic activities. MEG has a higher spatial resolution to localize epileptic activities for epilepsy surgery than other noninvasive approaches, such as electroencephalography (EEG) (Nakasato et al., 1994).

Localizing epileptogenic zones play a central role in epilepsy surgery. However, to date, there are no robust biomarkers that are able to accurately capture the location of epileptogenic zones (Tamilia et al., 2017). A variety of diagnostic indicators are introduced in the current clinical practice to estimate the epileptogenic zones. Despite the progress, existing methods, which heavily rely on epileptic spikes (typically $\leq$70 Hz), fail to reduce seizure frequency in approximately 50% of the cases, which greatly limits their applications in epilepsy surgery (Stigsdotter-Broman et al., 2014; Olan Çocuklarda and Öncesi, 2015; Oldham et al., 2015; Reinholdson et al., 2015; Verdinelli et al., 2015). High frequency oscillations (HFOs) (typically 80-500 Hz) can be used to localize the epileptogenic zones as biomarkers. Recent studies (Xiang et al., 2010; Ontario, 2012; Modur, 2014; Van Klink et al., 2014; Van't Klooster et al., 2015; Leung et al., 2018; Nevalainen et al., 2020) show that applying pathological HFOs to localize the epileptogenic zones leads to a significant reduction in seizure frequency, even seizure elimination in about 80% of cases. Thus, pathological HFOs have been associated with epileptogenic zones (Xiang et al., 2009; Miao et al., 2014). There is increasing evidence to show that HFOs are putative biomarkers to identify epileptic regions of the brain, which may improve the surgical diagnosis and surgical outcomes of patients with epilepsy.

Recent reports (Papadelis et al., 2009, 2016; Van Klink et al., 2016; Von Ellenrieder et al., 2016; Hedrich et al., 2017; Fan et al., 2021; Guo et al., 2022) have shown that MEG can detect epileptic spikes and HFOs effectively. In the presurgical diagnosis process, it is critical to accurately detect the HFOs in MEG signals for improving the post-surgical outcomes of patients with epilepsy. Visual reviews of HFOs in MEG signals by human experts play an important role in current clinical practices. However, visual identification of HFOs is usually subjective, time-consuming, and error prone due to the large volume of MEG signal data (Zelmann et al., 2012; Roehri et al., 2017; Fujiwara et al., 2020). Consequently, a number of automatic approaches (Gardner et al., 2007; Zelmann et al., 2010; Jacobs et al., 2012; Burnos et al., 2014) has been proposed to enable HFO detection so as to assist human experts for the visual review of iEEG and MEG signals. During the detection tasks, a universally two-step framework is applied by most of these methods: (1) The whole recording data is divided into a large number of signal segments. (2) The HFO detectors extract certain signal features for decision making. The handcrafted features that are manually designed based on observation or statistical analysis play as the solution for the feature of HFO signals. For example, Van Klink et al. (2017) proposed an automatic HFO detection and visualization approach in MEG. Similarly, in another work (Burnos et al., 2014), handcrafted features (e.g., high frequency peak and low frequency peak) were proposed to automatically distinguish HFOs in EEG signals. In these works, a cutoff for handcrafted features is often required to recognize an HFO signal segment. It is clear that these approaches based on handcrafted features require to be adjusted or re-optimized when the detectors are applied to similar neuroimaging data from different populations. This circumstance hinders the generalizability of HFOs in unseen conditions. Recently, machine learning provides a possible opportunity for improving the performance of HFO detections and reducing human interference. Traditional machine learning algorithm (Elahian et al., 2017), such as logistic regression, has been used for the identification of the epileptogenic zones. More recently, a deep learning approach SMO detector (Guo et al., 2018) was proposed. Such deep learning based HFO detector requires minimal human interference by using a golden standard dataset to train the detector.

Objective and automatic detection of HFOs in MEG signals with advanced deep learning algorithms may serve as a promising clinical decision support system to assist human experts for the visual review of MEG signals (Guo et al., 2018; Kong et al., 2019). In addition to the correct detection of HFOs in the MEG signals, recommending the possible results to clinicians is also crucial for an accurate and timely clinical evaluation. The recommended HFO list may not only serve as evidence for the particular patients but also serve as clinical diagnosis cases for future data retrieval purpose (e.g., teaching and research). In this study, our overall goal is set to develop a deep learning model to detect HFO signals with high confidence among a large amount of MEG data and recommend these findings to clinicians as a clinical decision support system. To that effect, we propose a multi-head self-attention deep neural network as the HFO detector. Compared to the existing algorithms (e.g., SMO detector Guo et al., 2018), we introduce the popular multi-head self-attention mechanism in this paper to enable an HFO detector to jointly pay attention to important information from various representation subspaces at multiple positions. Instead of computing the attention once, this multi-head self-attention strategy is able to compute the importance of each feature multiple times in parallel. Our hypothesis is that an HFO detector with multi-head self-attention mechanism is able to outperform the existing detector based on deep neural networks. Our newly developed HFO detector enables clinicians to objectively and automatically observe and localize HFOs for the preparation of epilepsy surgery without human designed signal features. According to the output probability values of our proposed detector, the MEG signals can be sorted in descending order. In order to accurately and timely understand the patient's condition, we can recommend *N* signals with the highest HFO signal probability value (e.g., top-10) to the clinician and assist in developing a treatment plan. This process is also known as a top-*N* recommendation task (Zhao et al., 2013, 2014; Wang et al., 2017; Zhang et al., 2017).

The following sections of this paper are organized as follows: First, we describe the patients and their associated MEG data in this work and the detailed MSADR framework in the Section 2. Second, experiment setups, such as model evaluation, peer machine learning models, and developmental environment are described. Third, we present the performance of detection and recommendation of MSADR for HFO signals. Then, ablation studies are also conducted to test our MSADR approach. Fourth, the discoveries and limitation of this work are discussed. Finally, we conclude the paper by summarizing the contributions and future directions.

2. Materials and methods

2.1. MEG data

2.1.1. Data acquisition

In this retrospective study, we obtained interictal MEG data from 20 clinical patients with epilepsy consisting of 10 females and 10 males (age: 6–60 years, mean age: 32 years), who were affected by focal seizures arising from one part of the brain. The Institutional Review Board was approved, and written informed consents were obtained from all subjects.

Full details of MEG data acquisition can be found in our prior study (Guo et al., 2018). Briefly, MEG recordings were performed using a 306-channel, whole-head MEG system (VectorView, Elekta Neuromag, Helsinki, Finland) in a magnetically shielded room. Sleep deprivation and reduction of anti-epileptic drugs were used to increase the chance for capture HFOs during MEG recordings, as one part of the pre-surgical evaluation. An approximate 1 h of MEG data was recorded for all patients. The sampling rate of MEG data was 2,400 Hz. The noise floor in our MEG systems was calculated with MEG data acquired without a subject (empty room). The noise floor was used to identify MEG system noise. The noise level was about 3–5 fT/Hz. The empty room measurements were also used to compute the noise covariance matrix for localizing epileptic activities (i.e., HFOs). A three-dimensional coordinate frame relative to the subject's head was derived from these positions. The system allowed head localization to an accuracy of 1 mm. The changes in head location before and after acquisition were required to be less than 5 mm for the study to be accepted. To identify the system and environmental noise, we routinely recorded one background MEG dataset without patients just before the experiment.

2.1.2. Segment

A public available software MEG Processor (Xiang et al., 2009) was used to correct and label the MEG data. In the current work, the MEG data were segmented into about 11,016,000 signal segments[1] with 2 s window size without overlap. Each signal segment was a MEG signal intensity vector with 4,800 data points in the time domain. These segments were first filtered automatically, which segments with a goodness-of-fit value of $< 85\%$ or confidence volume of $>3\text{mm}^3$ were dropped. Second, two band-pass filter, an 80–250 Hz one for ripples and a 250–500 Hz one for fast ripples, were introduced to filter high frequency MEG data into candidate segment set while the low frequency ones were automatically dropped. Both physiologic and pathologic high frequency neuromagnetic signals were included in the candidate segment set. The physiologic HFOs were manually rejected, and then the pathologic ones were selected by comparing MEG ripples and iEEG recordings at source levels (Wu et al., 2014) by two human experts. A total of 660 HFO signal segments selected by human experts, together with 660 normal control (NC) signal segments randomly selected from the rest segments, were compiled into our gold standard dataset. Figure 1 shows examples of MEG data, HFO, and NC segments.

2.2. Method

2.2.1. Overview framework for HFO detection and recommendation

In Figure 2, we display the overview of our multi-head self-attention based detector for recommendation (MSADR), which consists of MEG data acquisition, signal segmentation (purple box), multi-head self-attention based detector (orange box), HFO signal probability (green box), and detection and recommendation list.

The structure of multi-head self-attention based detector (MSAD) is given in Figure 3. It consists of layers of dense,

[1] The number of segments is calculated as $20(patients) \times 306(channels) \times 60(min) \times 60(s) \div 2(s/window)$.

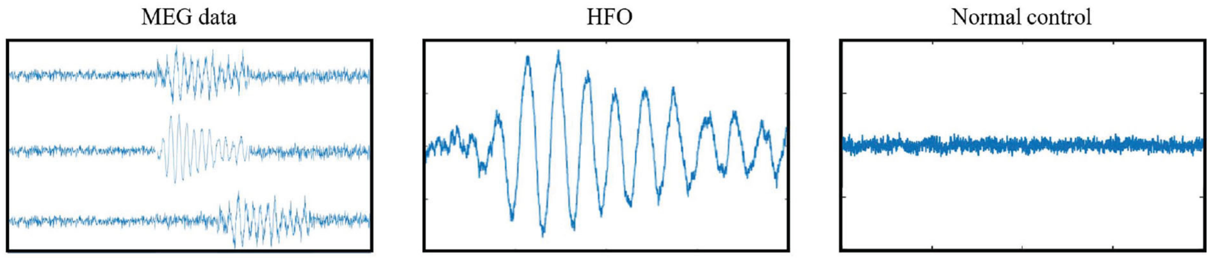


FIGURE 1
Examples of gold standard signals.

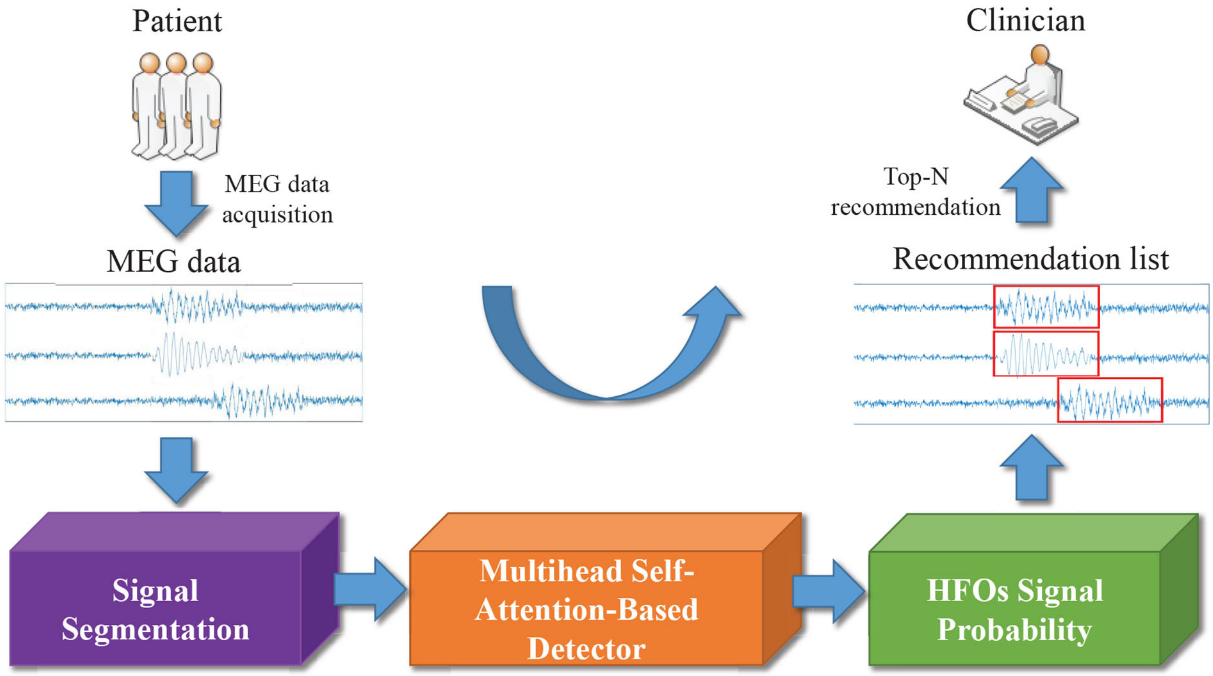


FIGURE 2
Overview of multi-head self-attention based detector for recommendation of neuromagnetic high frequency oscillations in epilepsy.

normalization, multi-head self-attention, self-attention. The various components are described in the following sections.

2.2.2. Dropout, dense, and normalization

There is one dropout layer, three dense layers, and two normalization layers in our proposed model. Figure 4 shows the computation details of these layers.

A dropout layer, which prevents over-fitting during model training, is applied to input data, i.e., HFO sequence. The white circle in Figure 4 indicates dropped units according to dropout probability. The dropout layer is followed by a dense layer, whose hidden units are 512, and the activation function is "relu", to reduce the dimension of the previous layer. The normalization layer (Ioffe and Szegedy, 2015) is used to accelerate deep network training by reducing internal covariate shift.

2.2.3. Self-attention

The attention is proposed to compute an alignment score between elements from two sources (Shen et al., 2018). In particular, given a sequence of HFOs, $\boldsymbol{x} = [x_1, x_2, ..., x_n]$ and a representation of a query $q \in \mathbb{R}^d$, the attention computes the alignment score between q and each element x_i using a compatibility function $f(x_i, q)$. A softmax function then transforms the alignment scores $[f(x_i, q)]_{i=1}^{n}$ to a probability distribution $p(z|\boldsymbol{x}, q)$, where z represents the importance degree to q. That is, a large $p(z = i|\boldsymbol{x}, q)$ means that x_i contributes

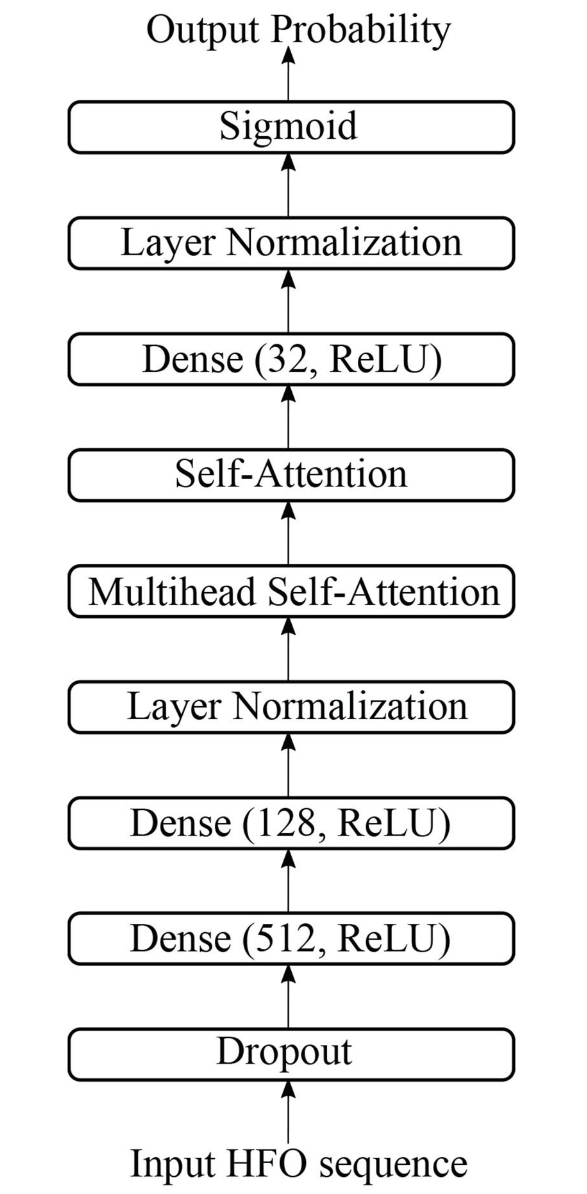


FIGURE 3
The structure of multi-head self-attention-based detector in this study.

important information to q. This attention process can be formalized as follows:

$$\alpha = [f(x_i, q)]_{i=1}^{n}, \tag{1}$$

$$p(z = i|\boldsymbol{x}, q) = softmax(\alpha). \tag{2}$$

The output *Attention* is the weighted element according to its importance, i.e.,

$$Attention(q, \boldsymbol{x}) = p(z = i|\boldsymbol{x}, q)\boldsymbol{x}. \tag{3}$$

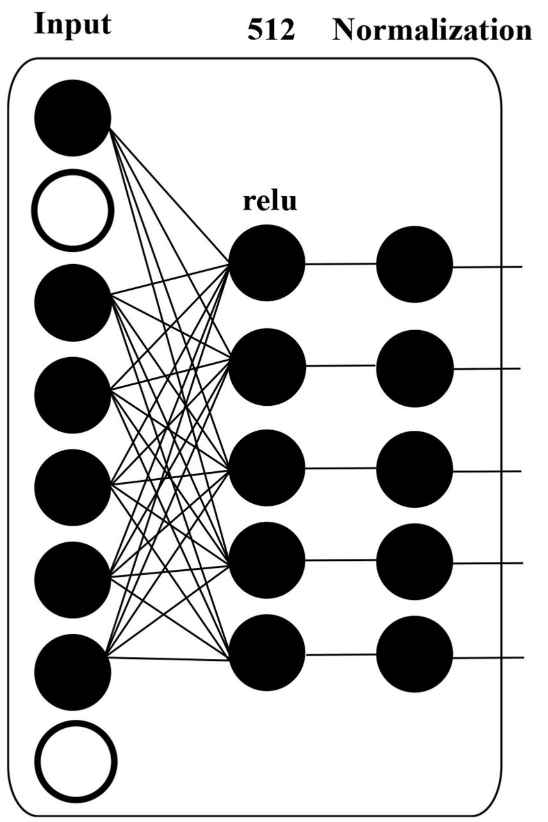


FIGURE 4
Details of MLP block in DANN structure.

Additive attention (Bahdanau et al., 2015; Shang et al., 2015) is a commonly-used attention mechanism where the compatibility function $f(\cdot)$ is parameterized by a MLP, *i.e.*:

$$f(x_i, q) = w^T\sigma(W^{(1)}x_i + W^{(2)}q), \tag{4}$$

where $W^{(1)} \in \mathbb{R}^{d\times d}$, $W^{(2)} \in \mathbb{R}^{d\times d}, w \in \mathbb{R}^d$ are learnable parameters, d is the number of columns of x_i, and $\sigma(\cdot)$ is an activation function. Compared with multiplicative attention (Rush et al., 2015; Sukhbaatar et al., 2015) using cosine similarity or inner product as the compatibility function for $f(x_i, q)$, *i.e.*:

$$f(x_i, q) = \langle\, W^{(1)}x_i,\ W^{(2)}q\,\rangle, \tag{5}$$

Though additive attention is expensive in time cost and memory consumption, it achieves better empirical performance for downstream tasks.

Self-attention (Liu et al., 2016; Lin et al., 2017; Peng et al., 2021) explores the importance of each feature to the entire HFOs given a specific task. In particular, q is removed from the common compatibility function which is formally written as the following equation:

$$f(x_i) = w^T\sigma(W^{(1)}x_i), \tag{6}$$

$$\alpha = [f(x_i)]_{i=1}^{n}, \tag{7}$$

$$p(z = i|\boldsymbol{x}) = softmax(\alpha). \tag{8}$$

The output *Attention* is the weighted element according to its importance, i.e.,

$$Attention(\boldsymbol{x}) = p(z = i|\boldsymbol{x})\boldsymbol{x}. \tag{9}$$

2.2.4. Multi-head self-attention

Multi-head self-attention allows the model to jointly attend to information from different representation subspaces at different positions. We use the multi-head version with k heads, as introduced in Vaswani et al. (2017),

$$MultiHead(\boldsymbol{x}) = Concat(head_1, \ldots, head_k)W^{(O)}, \tag{10}$$

$$\text{where } head_i = Attention(\boldsymbol{x}W^{(x)}), \tag{11}$$

where projections using learned parameter matrices $W^{(x)} \in \mathbb{R}^{d\times d/k}$, and $W^{(O)} \in \mathbb{R}^{d\times d}$.

2.2.5. Loss function

A standard cross-entropy loss is used as the training objective of MSADR, defined as

$$\mathcal{L} = -y\log(p) - (1-y)\log(1-p), \tag{12}$$

where y is the target label (0 or 1) and p is the predicted probability between 0 and 1 given an HFO sequence.

3. Experiment setup

We evaluate the proposed model on two tasks including the classification of whole patients and recommendation for each individual patient.

3.1. Model evaluation

We conduct a comprehensive evaluation in this study by employing the proposed MSADR on the HFO dataset to classify HFO data and recommend detected HFO sequences to medical experts. We employ the evaluation strategy of leave-one-out cross-validation in our experiments. In the 10-fold leave-one-out cross-validation, the HFO dataset is separated to two parts. One consists of 90% of the whole as training data while the rest part is regarded as test data.

For the detection task, we put all segments from 20 patients together to separate the training data and test data. We first calculate true positive (TP), false positive (FP), true negative (TN), and false negative (FN) by comparing the predicted labels and gold-standard labels. Then, we calculate accuracy, recall, precision, and F-score by

$$\begin{aligned}
Accuracy &= \frac{TP+TN}{TP+TN+FP+FN},\\
Recall &= \frac{TP}{TP+FN},\\
Precision &= \frac{TP}{TP+FP},\\
F-score &= 2\times\frac{Precision\times Recall}{Precision+Recall}.
\end{aligned}$$

For the recommendation task, the dataset is separated according to patients, in each split, segments from 18 patients are selected into the training set while segments from the rest 2 patients as test data. We use top-N precision (P@N) (Choi et al., 2016) to evaluate the ability of the algorithm to recommend detected HFOs for individual patients in the test set, defined as follows:

$$\text{P@}N = \frac{\text{TP@}N}{N},$$

where "TP@N" in the formula stands for TP HFOs in top-N recommendation task. Top-N precision mimics the behavior of doctors conducting differential diagnoses, where doctors list most probable diagnoses and treat patients accordingly to identify the patient status. Therefore, a machine with a high top-N precision translates to a doctor with an effective diagnostic skill. This makes top-N precision an attractive performance metric for our problem (Choi et al., 2016).

3.2. Peer machine learning models

To compare our proposed model MSADR with existing machine learning models, we also implemented random forest (RF) (Breiman, 2001; Nissen et al., 2018), support vector machine (SVM) models (Ak et al., 1999; Zhang et al., 2020), and SMO detector (Guo et al., 2018).

- **Random Forest (RF)**: RF is a classic ensemble learning methods by learning multiple decision trees and employing averaging to improve classification performance and control over-fitting. The number of trees in the forest was optimized from empirical values [20, 40, 60, 80, and 100]. We set maximal depth of the tree as 10.
- **Support Vector Machine (SVM)**: A SVM model is developed to perform classification by using vectorized FC features. We apply a linear kernel and search for the margin penalty with empirical values [0.2, 0.4, 0.6, 0.8, and 1.0].
- **SMO detector (Guo et al., 2018)**: In terms of the existing deep learning model, we compared our model with SMO

detector, a DNN model developed previously for HFO detection. Briefly, we implemented the SMO detector model as a 7-layer DNN, with input number of HFO sequence in the input layer, followed by dropout, dense (512, ReLU), dense (128, ReLU), normalization, dense (32, ReLU), normalization, and the sigmoid layer to generate one output unit. A cross entropy loss function is applied to supervise the network learning adopted. The learning rate is set as 0.0001. A total of 10 epochs are applied to ensure the convergence of the model.

3.3. Developmental environment

The proposed DANN and peer machine learning models are all implemented in Python 3.7 environment. To build the deep learning related models, we apply TensorFlow (2.0.0-rc1) backend. For the traditional models, we adopt the models from Sklearn 0.20.2.

All the experiments are conducted on a workstation with 10 cores of Intel Core i9 CPU and 64GB RAM. Due to the high computation cost of deep learning algorithm, we use one GPU (Nvidia TITAN Xp, 12GB RAM) to accelerate the training speed of the models.

4. Results

We evaluate the performance of detection and recommendation for each set of experiments. There are two sets of experiments to be conducted, which consist of overall performance compared with baseline models and effectiveness of varying head number of multi-head self-attention.

4.1. Overall performance comparison

4.1.1. Detection

We first compare the HFO detection performance of the proposed MSADR model and multiple peer machine learning models, including RF, SVM, and SMO. The results are derived on a leave-one-out cross-validation experiment by using the entire dataset. As shown in Table 1, our proposed MSADR takes the lead place (the bold value) in all metrics of HFO detection accuracy (0.886), recall (0.840), and F-score (0.859) among compared machine learning models, while the RF model returns the lowest detection performance on recall and F-score, and SVM on accuracy. Our model outperforms the SVM model by 0.126 on accuracy, the RF model by 0.263 on recall, and 0.142 on F-score.

4.1.2. Recommendation

We then compare the HFO recommendation performance of the proposed MSADR model and baseline models including RF, SVM, and SMO. The experiment setting is almost the same as the detection task, except for an additional recommendation module to generate a ranking list. As shown in Table 2, our proposed MSADR obtains the best HFO recommendation performance on P@1 (0.967), P@3 (0.858), and P@5 (0.879) among the compared machine learning models, whereas the RF model returns the lowest recommendation performance on P@3, P@5, and SVM on P@1. Our model increases the performance of SVM by 0.1 on P@1, the RF model by 0.162 and 0.215 on P@3 and P@5, respectively.

4.1.3. Computational costs

The computational costs of the proposed MSADR model and baseline models are provided in Table 3. It is noticed that our proposed MSADR takes a longer time for model training and inference than baseline models. The main reason is the complexity of the deep learning models (MSADR and SMO) and the huge number of the parameters. The MSADR has 411,325 parameters to be trained, which obtains a stronger learning capacity to get the best model performance. MSADR uses a learned model to conduct the detection and recommendation task, which theoretically take more time.

The results in Tables 1, 2 also show a trend that deep learning models (MSADR and SMO) achieve improved performance compared to the traditional model, such as SVM and RF, demonstrating the superior capability of the deep learning

TABLE 1 Detection comparison of random forest (RF), support vector machine (SVM), SMO, and multi-head self-attention-based detector for recommendation (MSADR) trained using leave-one-out cross-validation on the entire dataset.

Method	Accuracy	Recall	Precision	F-Score
RF	0.779	0.577	**0.951**	0.717
SVM	0.760	0.743	0.764	0.753
SMO detector (Guo et al., 2018)	0.845	0.732	**0.951**	0.826
MSADR	**0.886**	**0.840**	0.881	**0.859**

The bold values represent the lead place of the corresponding metric.

TABLE 2 Recommendation comparison of RF, SVM, SMO, and MSADR trained using leave-one-out cross-validation on the entire dataset.

Method	P@1	P@3	P@5
RF	0.893	0.696	0.664
SVM	0.867	0.800	0.760
SMO detector (Guo et al., 2018)	0.893	0.811	0.793
MSADR	**0.967**	**0.858**	**0.879**

The bold values represent the lead place of the corresponding metric.

TABLE 3 Computational cost comparison of RF, SVM, SMO, and MSADR.

Method	Training time (ms)	Inference time (ms)	No. of parameters
RF	284	6	-
SVM	83	5	2,400
SMO detector (Guo et al., 2018)	2,506	140	318,449
MSADR	12,317	525	411,325

model on complex data patterns, such as HFO. In addition, the inference time of MSADR is about 0.5 s (525 ms). This is an acceptable time cost while it can bring about 16.6% accuracy improvement on detection and 11.5% P@1 improvement on recommendation toward the fastest model (SVM). Another trend can be observed that the P@1 returns the best recommendation score a cross all models, which demonstrates that the machine learning model is a promising alternative approach to assist clinicians to make decisions.

4.2. Effectiveness of varying head number of multi-head self-attention

The effectiveness of our MSADR is further tested by varying head number ($k = [2, 4, 8, 16]$) of multi-head self-attention on two tasks of detection and recommendation. The results in this set of experiments are calculated based on leave-one-out cross-validation by using the entire dataset.

4.2.1. Detection

The HFO detection performance of the proposed MSADR model has been evaluated by varying head number of multi-head self-attention. Figure 5A displays plots of the accuracy, recall, precision, and F-score of the proposed MSADR over different strategies of varying head number.

It is apparent that the proposed MSADR achieves the state-of-the-art in terms of accuracy, recall, and F-score when the head number is set as 8, while the performance on precision is the worst. Performance accuracy, recall, and F-score decrease as the head number increase due to the over-fitting of self-attention, whereas performance on precision slightly increases. Overall, the proposed MSADR has the best three indicators out of four when the head number of multi-head self-attention is set to 8.

4.2.2. Recommendation

The effectiveness of the proposed MSADR model over HFO recommendation task has been compared in terms of P@*N* by varying head number of multihead self-attention. Figure 5B displays plots of the P@1, P@3, and P@5 of MSADR over different strategies of varying head number. As shown in the figure, our proposed MSADR achieves the highest HFO recommendation performance on P@1 and P@5, while the performance of the two indicators decreases as the head number increases due to the over-fitting of multi-head self-attention. Overall, the proposed MSADR has the best two indicators out of three when the head number of multi-head self-attention is set to 8.

4.3. Ablation study

A detailed ablation study is performed to examine the contributions of the model's components to the tasks of detection and recommendation. There are four configurations of replaceable components in this model. The two components are (1) multi-head self-attention layer and (2) the self-attention layer. The four configurations based on MSADR are

- **raw (DNN):** (1) and (2) are removed from MSADR, which becomes a pure DNN, one of our peer baseline models (SMO detector);
- **Attn_1:** (2) is removed and (1) is remained in MSADR;
- **Attn_2:** (1) is removed and (2) is remained in MSADR;
- **MSADR:** our proposed model.

All models are trained with 10 epochs and a batch size of 32. The head number of multi-head self-attention is empirically set to 8.

4.3.1. Detection

From Table 4, we find that the MSADR model obtains the best performance on detection task compared to the ablated models, except for the performance of Attn_1 on precision. Moreover, we note that Attn_1 and Attn_2 outperform raw, which gives us the confidence to apply self-attention to learn the relationship between HFO signals. It is clear that the single self-attention model provides comparable information to the performance of the Attn_1 and Attn_2 model. In particular, MSADR outperforms the best ablated model for Accuracy by 1.1%, for Recall by 2.1%, and for F-score by 1.0%.

4.3.2. Recommendation

Table 5 shows the recommendation performance for the ablated models and our proposed model. As can be seen from the table, the proposed model achieves the best performance compared to the ablated models on the recommendation task. We observe that Attn_1 and Attn_2 outperform raw on P@1 and P@3, which again demonstrates that self-attention is a vital component to learn the relationship

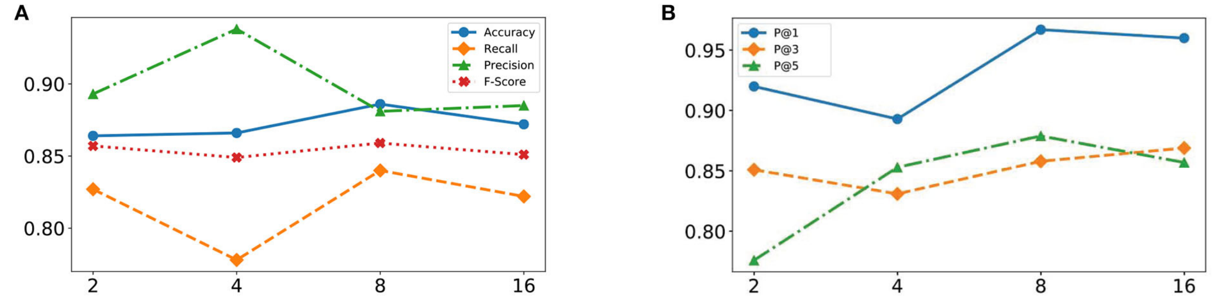


FIGURE 5
Effectiveness of varying head number of multi-head self-attention from 2 to 16. **(A)** Detection task. **(B)** Recommendation task.

TABLE 4 Detection comparison of ablated models trained using leave-one-out cross-validation on the entire dataset.

Method	Accuracy	Recall	Precision	F-Score
raw	0.845	0.732	0.951	0.826
Attn_1	0.845	0.755	**0.975**	0.849
Attn_2	0.875	0.819	0.928	0.847
MSADR	**0.886**	**0.840**	0.881	**0.859**

The bold values represent the lead place of the corresponding metric.

TABLE 5 Recommendation comparison of ablated models trained using leave-one-out cross-validation on the entire dataset.

Method	P@1	P@3	P@5
raw	0.893	0.811	0.793
Attn_1	0.913	0.844	0.759
Attn_2	0.920	0.847	0.873
MSADR	**0.967**	**0.858**	**0.879**

The bold values represent the lead place of the corresponding metric.

between HFO signals. MSADR outperforms the best ablated model by 4.7%, 1.1%, and 0.6% on P@1, P@3, and P@5, respectively.

5. Discussion

Since first discovered in the 1990s, HFOs have been considered a promising biomarker to locating the seizure onset zone and improving postsurgical outcomes in patients with epilepsy (Huang and White, 1989; Fan et al., 2021). Noninvasive brain recording technologies (i.e., scalp EEG and MEG) were a milestone in human HFO research and have provided the possibility to investigate this brain activity in a wider range (Papadelis et al., 2009, 2016; Van Klink et al., 2016; Von Ellenrieder et al., 2016; Hedrich et al., 2017). Due to excellent temporal resolution and acceptable spatial resolution, MEG is able to effectively record HFOs and localize epileptic activities for epilepsy surgery (Fan et al., 2021). After noninvasive recording, the detection of HFOs is the next crucial task for onset zone detection. Although visual identification is still considered to be the gold standard for HFO detection, it still faces the problem of highly time-consuming and subjective (Frauscher et al., 2017).

This study mainly focuses on the automatic detection and recommendation of HFOs from interictal MEG data. The MEG data of clinical epileptic patients were recorded with a multi-channel whole-head MEG system (Xiang et al., 2010; Guo et al., 2018), and then segmented into signal segments with 2 s window size without overlap. The labeled HFO segments by human experts and randomly selected NC segments from the complementary set of labeled HFO set was compiled into our gold standard dataset. With the gold standard data, we trained the proposed MSADR algorithm for the detection and recommendation model of HFOs. For a new patient, the trained model can detect HFOs from the segmented MEG data and recommend HFO signals to clinicians, alleviating the burden on reviewing the large amount of MEG data. The effectiveness of our proposed detection and recommendation approaches were demonstrated by the cross-validation experimental results. The proposed MSADR can improve the detection accuracy by at least 13.7% and the top-1 recommendation precision by 8.2% compared with the traditional machine learning methods (RF and SVM) while improving the detection accuracy by 4.8% and the top-1 recommendation precision by 8.2% compared with another deep learning method (SMO detector). The computational costs are important, especially in real world applications. Though MSADR is a time-consuming method, the acceptable inference time (0.5 s) can guarantee the user experience. In addition, sliding window with overlap can be used for segmenting, so as to improve the possibility of HFO locating in the center of segments in real world applications.

There are some limitations to this study. First, the experiment was built on a small dataset with 20 patients. A larger data set is required to further validate the effectiveness and efficiency of MSADR. Second, this work only focused on the detection and recommendation of HFOs from interictal MEG data. Performance of the MSADR approach on other neuromagnetic data (i.e., ictal MEG, iEEG, and EEG) remains unclear. We will test our method in future work. Third, the MEG segments from different patients or channels are treated equally and independently in this paper. However, there are complex timing and co-occurrence relationships among segments. Mining and utilizing these relationships may improve the effectiveness of HFO detection and recommendation. Finally, since the current approach requires signal segmentation of MEG data, it is only able to differentiate HFOs and NC segments with a pre-defined fixed signal length. It cannot directly detect HFOs in an automatic way on the raw MEG data (e.g., start and end positions).

Due to the high cost and lack of automatic detection technology with broad applicability, traditional MEG has limited availability (Guo et al., 2018; Kong et al., 2019). However, the technological innovations in MEG have been progressing. New MEG systems with optically pumped magnetometers do not require cooling with liquid helium and can be worn more conveniently (Boto et al., 2019). This may reduce the cost of MEG data recording and expand the scope of application. The research and application of automatic or semi-automatic HFO detection methods with broad applicability will make more efficient use of MEG data (Guo et al., 2018), and the integration into clinical review software can effectively enhance clinical value, including preoperative localization of epileptogenic regions, the assessment of disease severity, predicting seizures, monitoring treatment, evaluating treatment effects, and assessing epileptic susceptibility after brain injury (Fan et al., 2021).

6. Conclusion

In this study, we develop an MSADR detector for the detection and recommendation of HFO signals by using the multi-head self-attention mechanism. By comparing our model with traditional machine learning models (RF and SVM) and deep learning model (SMO detector), the proposed MSADR detector is proved to reach state-of-the-art performance in both detection and recommendation tasks. The robustness of our detector is also extensively assessed with multiple ablation tests. The MSADR is supposed to detect HFOs from a large amount of segmented MEG data and recommend HFO signals to help clinicians locate epileptogenic regions and assist in treatment. Our future directions may focus on extending our model to capture the timing and co-occurrence relationships among segments to further improve the effectiveness of HFO detection and recommendation.

Data availability statement

The datasets presented in this article are not readily available because of a confidentiality agreement that prevents them from being disclosed to the public. Requests to access the datasets should be directed to JG, guojy@xmu.edu.cn.

Ethics statement

The studies involving human participants were reviewed and approved by Nanjing Brain Hospital. The patients/participants provided their written informed consent to participate in this study.

Author contributions

XZ, XP, KN, JG, and YP contributed to the conception and design of the study. XP and FY developed the DL algorithms. TW made a contribution to data collection. XZ, DC, QZ, MO, and JG performed the statistical analysis. XZ and KN wrote the initial draft of the manuscript. All authors contributed to manuscript revision, read, and approved the submitted version.

Funding

This work was supported by the National Key Research and Development Program of China (Nos. 2020YFD1100605 and 2019YFD1101100), Beijing Information Science and Technology University Qin Xin Talents Cultivation Program (No. QXTCP C202112), Research Level Improvement Project (No. 2020KYNH214), Beijing Educational Science Planning Project of China (No. CHCA2020102), National Institutes of Health (Nos. R01-EB029944 and R01-EB030582), National Natural Science Foundation of China (Nos. 82172022 and 62006100), 2025 Key Technology Innovation Program of Ningbo City (No. 2019B10082) Fundamental Research Funds for the Chinese Central Universities (0070ZK1096 to JG), and Ningbo Public Welfare Technology Plan Project (No. 2019C50081).

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.

The reviewer AM declared a shared affiliation with the author TW to the handling editor at the time of review.

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.

References

Ak, J., Suykens, J., and Vandewalle (1999). Least squares support vector machine classifiers. *Neural Process. Lett.* 9, 293–300.

Bahdanau, D., Cho, K. H., and Bengio, Y. (2015). "Neural machine translation by jointly learning to align and translate," in *Proceedings of the 3rd International Conference on Learning Representations, ICLR 2015* (San Diego, CA).

Boto, E., Seedat, Z. A., Holmes, N., Leggett, J., Hill, R. M., Roberts, G., et al. (2019). Wearable neuroimaging: combining and contrasting magnetoencephalography and electroencephalography. *Neuroimage* 201, 116099–116099. doi: 10.1016/j.neuroimage.2019.116099

Breiman, L. (2001). Random forests. *Mach. Learn.* 45, 5–32. doi: 10.1023/A:1010933404324

Burnos, S., Hilfiker, P., Sürücü,, O., Scholkmann, F., Krayenbühl, N., Grunwald, T., et al. (2014). Human intracranial high frequency oscillations (hfos) detected by automatic time-frequency analysis. *PLoS ONE* 9, e94381. doi: 10.1371/journal.pone.0094381

Choi, E., Bahadori, M. T., Schuetz, A., Stewart, W. F., and Sun, J. (2016). "Doctor ai: predicting clinical events via recurrent neural networks," in *Machine Learning for Healthcare Conference* (Los Angeles, CA), 301–318.

Durnford, A. J., Rodgers, W., Kirkham, F. J., Mullee, M. A., Whitney, A., Prevett, M., et al. (2011). Very good inter-rater reliability of engel and ilae epilepsy surgery outcome classifications in a series of 76 patients. *Seizure* 20, 809–812. doi: 10.1016/j.seizure.2011.08.004

Elahian, B., Yeasin, M., Mudigoudar, B., Wheless, J. W., and Babajani-Feremi, A. (2017). Identifying seizure onset zone from electrocorticographic recordings: a machine learning approach based on phase locking value. *Seizure* 51, 35–42. doi: 10.1016/j.seizure.2017.07.010

Fan, Y., Dong, L., Liu, X., Wang, H., and Liu, Y. (2021). Recent advances in the noninvasive detection of high-frequency oscillations in the human brain. *Rev. Neurosci.* 32, 305–321. doi: 10.1515/revneuro-2020-0073

Frauscher, B., Bartolomei, F., Kobayashi, K., Cimbalnik, J., van 't Klooster, M. A., Rampp, S., et al. (2017). High-frequency oscillations: the state of clinical research. *Epilepsia* 58, 1316–1329. doi: 10.1111/epi.13829

Fujiwara, K., Huang, Y., Hori, K., Nishioji, K., Kobayashi, M., Kamaguchi, M., et al. (2020). Over- and under-sampling approach for extremely imbalanced and small minority data problem in health record analysis. *Front. Public Health* 8, 178. doi: 10.3389/fpubh.2020.00178

Gardner, A. B., Worrell, G. A., Marsh, E., Dlugos, D., and Litt, B. (2007). Human and automated detection of high-frequency oscillations in clinical intracranial eeg recordings. *Clin. Neurophysiol.* 118, 1134–1143. doi: 10.1016/j.clinph.2006.12.019

Guo, J., Xiao, N., Li, H., He, L., Li, Q., Wu, T., et al. (2022). Transformer-based high-frequency oscillation signal detection on magnetoencephalography from epileptic patients. *Front. Mol. Biosci.* 9, 822810–822810. doi: 10.3389/fmolb.2022.822810

Guo, J., Yang, K., Liu, H., Yin, C., Xiang, J., Li, H., et al. (2018). A stacked sparse autoencoder-based detector for automatic identification of neuromagnetic high frequency oscillations in epilepsy. *IEEE Trans. Med. Imaging* 37, 2474–2482. doi: 10.1109/TMI.2018.2836965

Hedrich, T., Pellegrino, G., Kobayashi, E., Lina, J.-M., and Grova, C. (2017). Comparison of the spatial resolution of source imaging techniques in high-density eeg and meg. *Neuroimage* 157, 531–544. doi: 10.1016/j.neuroimage.2017.06.022

Hu, J., Wang, C. S., Wu, M., Du, Y. X., He, Y., and She, J. (2015). Removal of eog and emg artifacts from eeg using combination of functional link neural network and adaptive neural fuzzy inference system. *Neurocomputing* 151, 278–287. doi: 10.1016/j.neucom.2014.09.040

Huang, C., and White, L. (1989). High-frequency components in epileptiform EEG. *J. Neurosci. Methods* 30, 197–201. doi: 10.1016/0165-0270(89)90130-1

Ioffe, S., and Szegedy, C. (2015). "Batch normalization: accelerating deep network training by reducing internal covariate shift," in *Proceedings of the 32nd International Conference on Machine Learning* (Lille), 448–456.

Jacobs, J., Staba, R., Asano, E., Otsubo, H., Wu, J., Zijlmans, M., et al. (2012). High-frequency oscillations (hfos) in clinical epilepsy. *Progr. Neurobiol.* 98, 302–315. doi: 10.1016/j.pneurobio.2012.03.001

Kong, Y., Gao, J., Xu, Y., Pan, Y., Wang, J., and Liu, J. (2019). Classification of autism spectrum disorder by combining brain connectivity and deep neural network classifier. *Neurocomputing* 324, 63–68. doi: 10.1016/j.neucom.2018.04.080

Leung, H., Poon, W. S., Kwan, P. K., Lui, C. H., Poon, T. L., Chan, E. L., et al. (2018). Ictal intracranial electroencephalography using wavelet analysis of high-frequency oscillations in chinese patients with refractory epilepsy. *Hong Kong Med. J.* 24, 21–23.

Lin, Z., Feng, M., dos Santos, C. N., Yu, M., Xiang, B., Zhou, B., et al. (2017). A structured self-attentive sentence embedding. *arXiv:1703.03130.* doi: 10.48550/arXiv.1703.03130

Liu, Y., Sun, C., Lin, L., and Wang, X. (2016). Learning natural language inference using bidirectional lstm model and inner-attention. *arXiv:1605.09090.* doi: 10.48550/arXiv.1605.09090

Miao, A., Xiang, J., Tang, L., Ge, H., Liu, H., Wu, T., et al. (2014). Using ictal high-frequency oscillations (80-500 hz) to localize seizure onset zones in childhood absence epilepsy: a meg study. *Neurosci. Lett.* 566, 21–26. doi: 10.1016/j.neulet.2014.02.038

Modur, P. N. (2014). High frequency oscillations and infraslow activity in epilepsy. *Ann. Indian Acad. Neurol.* 17(Suppl. 1), S99. doi: 10.4103/0972-2327.128674

Nakasato, N., Levesque, M. F., Barth, D. S., Baumgartner, C., Rogers, R. L., and Sutherling, W. W. (1994). Comparisons of meg, eeg, and ecog source localization in neocortical partial epilepsy in humans. *Electroencephalogr, Clin, Neurophysiol.* 91, 171–178. doi: 10.1016/0013-4694(94)90067-1

Nevalainen, P., von Ellenrieder, N., Klimes, P., Dubeau, F., Frauscher, B., and Gotman, J. (2020). Association of fast ripples on intracranial eeg and outcomes after epilepsy surgery. *Neurology* 95, 10468. doi: 10.1212/WNL.0000000000010468

Nissen, I. A., Stam, C. J., Van, S., Viktor, W., Reijneveld, J. C., Baayen, J. C., et al. (2018). Localization of the epileptogenic zone using interictal meg and machine learning in a large cohort of drug-resistant epilepsy patients. *Front. Neurol.* 9, 647. doi: 10.3389/fneur.2018.00647

Olan Çocuklarda, M. N. D. E., and Öncesi, C. (2015). Presurgical evaluation and epilepsy surgery in mri negative resistant epilepsy of childhood with good outcome. *Turk Neurosurg* 25, 905–913. doi: 10.5137/1019-5149.JTN.12093-14.0

Oldham, M. S., Horn, P. S., Tsevat, J., and Standridge, S. (2015). Costs and clinical outcomes of epilepsy surgery in children with drug-resistant epilepsy. *Pediatr. Neurol.* 53, 216–220. doi: 10.1016/j.pediatrneurol.2015.05.009

Ontario, H. Q. (2012). Epilepsy surgery: an evidence summary. *Ont. Health Technol. Assess. Ser.* 12, 1.

Papadelis, C., Poghosyan, V., Fenwick, P. B., and Ioannides, A. A. (2009). Meg's ability to localise accurately weak transient neural sources. *Clin. Neurophysiol.* 120, 1958–1970. doi: 10.1016/j.clinph.2009.08.018

Papadelis, C., Tamilia, E., Stufflebeam, S., Grant, P. E., Madsen, J. R., Pearl, P. L., et al. (2016). Interictal high frequency oscillations detected with simultaneous magnetoencephalography and electroencephalography as biomarker of pediatric epilepsy. *J. Vis. Exp.* 6, e54883. doi: 10.3791/54883

Peng, X., Long, G., Shen, T., Wang, S., and Jiang, J. (2021). "Sequential diagnosis prediction with transformer and ontological representation," in *2021 IEEE International Conference on Data Mining (ICDM)* (Auckland: IEEE), 489–498.

Reinholdson, J., Olsson, I., Edelvik, A., Hallböök, T., Lundgren, J., Rydenhag, B., et al. (2015). Long-term follow-up after epilepsy surgery in infancy and early childhood-a prospective population based observational study. *Seizure* 30, 83–89. doi: 10.1016/j.seizure.2015.05.019

Roehri, N., Pizzo, F., Bartolomei, F., Wendling, F., and Bénar, C.-G. (2017). What are the assets and weaknesses of hfo detectors? a benchmark framework based on realistic simulations. *PLoS ONE* 12, e0174702. doi: 10.1371/journal.pone.0174702

Rush, A. M., Chopra, S., and Weston, J. (2015). A neural attention model for abstractive sentence summarization. *arXiv:1509.00685*. doi: 10.18653/v1/D15-1044

Shang, L., Lu, Z., and Li, H. (2015). "Neural responding machine for short-text conversation," in *Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)* (Beijing: Association for Computational Linguistics), 1577–1586.

Shen, T., Zhou, T., Long, G., Jiang, J., Pan, S., and Zhang, C. (2018). "Disan: directional self-attention network for rnn/cnn-free language understanding," in *AAAI'18/IAAI'18/EAAI'18: Proceedings of the Thirty-Second AAAI Conference on Artificial Intelligence and Thirtieth Innovative Applications of Artificial Intelligence Conference and Eighth AAAI Symposium on Educational Advances in Artificial Intelligence* (New Orleans, LA).

Stigsdotter-Broman, L., Olsson, I., Flink, R., Rydenhag, B., and Malmgren, K. (2014). Long-term follow-up after callosotomy–a prospective, population based, observational study. *Epilepsia* 55, 316–321. doi: 10.1111/epi.12488

Sukhbaatar, S., Weston, J., and Fergus, R. (2015). "End-to-end memory networks," in *NIPS* (Montreal, QC), 2440–2448.

Tamilia, E., Madsen, J. R., Grant, P. E., Pearl, P. L., and Papadelis, C. (2017). Current and emerging potential of magnetoencephalography in the detection and localization of high-frequency oscillations in epilepsy. *Front. Neurol.* 8, 14. doi: 10.3389/fneur.2017.00014

Van Klink, N., Hillebrand, A., and Zijlmans, M. (2016). Identification of epileptic high frequency oscillations in the time domain by using meg beamformer-based virtual sensors. *Clin. Neurophysiol.* 127, 197–208. doi: 10.1016/j.clinph.2015.06.008

Van Klink, N., Van Rosmalen, F., Nenonen, J., Burnos, S., Helle, L., Taulu, S., et al. (2017). Automatic detection and visualisation of meg ripple oscillations in epilepsy. *Neuroimage Clin.* 15, 689–701. doi: 10.1016/j.nicl.2017.06.024

Van Klink, N., Van't Klooster, M., Zelmann, R., Leijten, F., Ferrier, C., Braun, K., et al. (2014). High frequency oscillations in intra-operative electrocorticography before and after epilepsy surgery. *Clin. Neurophysiol.* 125, 2212–2219. doi: 10.1016/j.clinph.2014.03.004

Van't Klooster, M. A., Leijten, F. S., Huiskamp, G., Ronner, H. E., Baayen, J. C., Van Rijen, P. C., et al. (2015). High frequency oscillations in the intra-operative ecog to guide epilepsy surgery ("the hfo trial"): study protocol for a randomized controlled trial. *Trials* 16, 422. doi: 10.1186/s13063-015-0932-6

Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., et al. (2017). "Attention is all you need," in *NeurIPS* (Long Beach, CA), 5998–6008.

Verdinelli, C., Olsson, I., Edelvik, A., Hallböök, T., Rydenhag, B., and Malmgren, K. (2015). A long-term patient perspective after hemispherotomy-a population based study. *Seizure* 30, 76–82. doi: 10.1016/j.seizure.2015.05.016

Von Ellenrieder, N., Pellegrino, G., Hedrich, T., Gotman, J., Lina, J.-M., Grova, C., et al. (2016). Detection and magnetic source imaging of fast oscillations (40-160 hz) recorded with magnetoencephalography in focal epilepsy patients. *Brain Topogr.* 29, 218–231. doi: 10.1007/s10548-016-0471-9

Wang, S., Hu, L., and Cao, L. (2017). "Perceiving the next choice with comprehensive transaction embeddings for online recommendation," in *ECML_PKDD* (Springer), 285–302.

Wu, T., Ge, S., Zhang, R., Liu, H., Chen, Q., Zhao, R., et al. (2014). Neuromagnetic coherence of epileptic activity: an meg study. *Seizure* 23, 417–423. doi: 10.1016/j.seizure.2014.01.022

Xiang, J., Liu, Y., Wang, Y., Kirtman, E. G., Kotecha, R., Chen, Y., et al. (2009). Frequency and spatial characteristics of high-frequency neuromagnetic signals in childhood epilepsy. *Epileptic Disord.* 11, 113–125. doi: 10.1684/epd.2009.0253

Xiang, J., Wang, Y., Chen, Y., Liu, Y., Kotecha, R., Huo, X., et al. (2010). Noninvasive localization of epileptogenic zones with ictal high-frequency neuromagnetic signals: case report. *J. Neurosurg. Pediatr.* 5, 113–122. doi: 10.3171/2009.8.PEDS09345

Yamakawa, T., Miyajima, M., Fujiwara, K., Kano, M., Suzuki, Y., Watanabe, Y., et al. (2020). Wearable epileptic seizure prediction system with machine-learning-based anomaly detection of heart rate variability. *Sensors* 20, 3987. doi: 10.3390/s20143987

Zelmann, R., Mari, F., Jacobs, J., Zijlmans, M., Chander, R., and Gotman, J. (2010). "Automatic detector of high frequency oscillations for human recordings with macroelectrodes," In *2010 Annual International Conference of the IEEE Engineering in Medicine and Biology* (Buenos Aires: IEEE), 2329–2333.

Zelmann, R., Mari, F., Jacobs, J., Zijlmans, M., Dubeau, F., and Gotman, J. (2012). A comparison between detectors of high frequency oscillations. *Clin. Neurophysiol.* 123, 106–116. doi: 10.1016/j.clinph.2011.06.006

Zhang, J., Richardson, J. D., and Dunkley, B. T. (2020). Classifying post-traumatic stress disorder using the magnetoencephalographic connectome and machine learning. *Sci. Rep.* 10:5937. doi: 10.1038/s41598-020-62713-5

Zhang, Y., Ai, Q., Chen, X., and Croft, W. B. (2017). "Joint representation learning for top-n recommendation with heterogeneous information sources," in *CIKM '17: Proceedings of the 2017 ACM on Conference on Information and Knowledge Management* (Singapore: ACM), 1449–1458.

Zhao, X., Niu, Z., and Chen, W. (2013). Interest before liking: two-step recommendation approaches. *Knowledge Based Syst.* 48, 46–56. doi: 10.1016/j.knosys.2013.04.009

Zhao, X., Niu, Z., Chen, W., Shi, C., Niu, K., and Liu, D. (2014). A hybrid approach of topic model and matrix factorization based on two-step recommendation framework. *J. Intell. Inf. Syst.* 44, 335–353. doi: 10.1007/s10844-014-0334-3

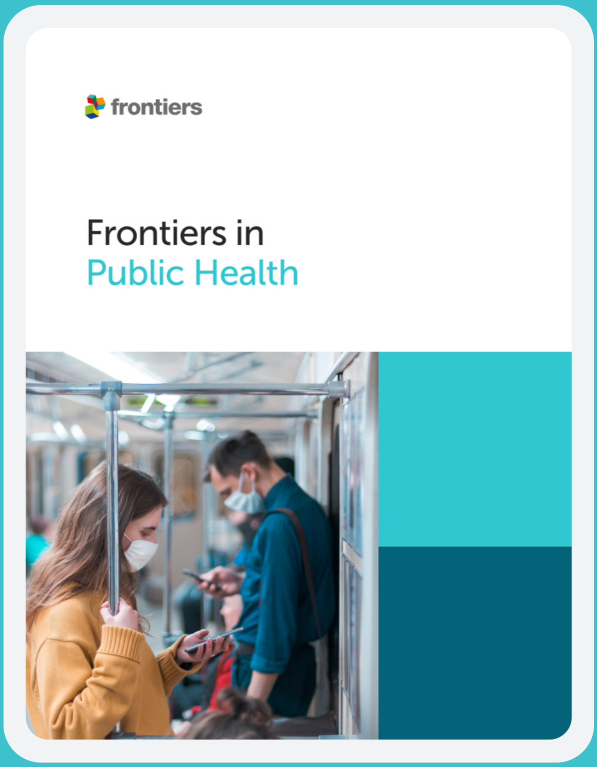
Frontiers in
Public Health

Explores and addresses today's fast-moving healthcare challenges

One of the most cited journals in its field, which promotes discussion around inter-sectoral public health challenges spanning health promotion to climate change, transportation, environmental change and even species diversity.

Discover the latest Research Topics

See more →



frontiers

Frontiers in
Public Health

Frontiers

Avenue du Tribunal-Fédéral 34
1005 Lausanne, Switzerland
frontiersin.org

Contact us

+41 (0)21 510 17 00
frontiersin.org/about/contact