frontiers | Research Topics

Novel methods and technologies for the evaluation of drug outcomes and policies

Edited by
Dalia M. Dawoud, Blythe Adamson, Grammati Sarri, Amr Makady and Zaheer-Ud-Din Babar

Coordinated by
Omneya Mohamed

Published in
Frontiers in Pharmacology

FRONTIERS EBOOK COPYRIGHT STATEMENT

The copyright in the text of individual articles in this ebook is the property of their respective authors or their respective institutions or funders. The copyright in graphics and images within each article may be subject to copyright of other parties. In both cases this is subject to a license granted to Frontiers.

The compilation of articles constituting this ebook is the property of Frontiers.

Each article within this ebook, and the ebook itself, are published under the most recent version of the Creative Commons CC-BY licence. The version current at the date of publication of this ebook is CC-BY 4.0. If the CC-BY licence is updated, the licence granted by Frontiers is automatically updated to the new version.

When exercising any right under the CC-BY licence, Frontiers must be attributed as the original publisher of the article or ebook, as applicable.

Authors have the responsibility of ensuring that any graphics or other materials which are the property of others may be included in the CC-BY licence, but this should be checked before relying on the CC-BY licence to reproduce those materials. Any copyright notices relating to those materials must be complied with.

Copyright and source acknowledgement notices may not be removed and must be displayed in any copy, derivative work or partial copy which includes the elements in question.

All copyright, and all rights therein, are protected by national and international copyright laws. The above represents a summary only. For further information please read Frontiers' Conditions for Website Use and Copyright Statement, and the applicable CC-BY licence.

ISSN 1664-8714
ISBN 978-2-8325-4764-9
DOI 10.3389/978-2-8325-4764-9

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

Novel methods and technologies for the evaluation of drug outcomes and policies

Topic editors

Dalia M. Dawoud — National Institute for Health and Care Excellence, United Kingdom

Blythe Adamson — Flatiron Health, United States

Grammati Sarri — Cytel, United Sates

Amr Makady — Janssen Pharmaceutica NV, Belgium

Zaheer-Ud-Din Babar — University of Huddersfield, United Kingdom

Topic Coordinator

Omneya Mohamed — Baxter Internation Inc, United Arab Emirates

Citation

Dawoud, D. M., Adamson, B., Sarri, G., Makady, A., Babar, Z.-U.-D., Mohamed, O., eds. (2024). *Novel methods and technologies for the evaluation of drug outcomes and policies*. Lausanne: Frontiers Media SA. doi: 10.3389/978-2-8325-4764-9

Table of contents

frontiers | Frontiers in Pharmacology

TYPE Editorial
PUBLISHED 18 April 2024
DOI 10.3389/fphar.2024.1396034

Check for updates

OPEN ACCESS

EDITED AND REVIEWED BY
Marcus Tolentino Silva,
University of Brasilia, Brazil

*CORRESPONDENCE
Blythe Adamson,
badamson@flatiron.com

RECEIVED 05 March 2024
ACCEPTED 28 March 2024
PUBLISHED 18 April 2024

CITATION
Adamson B, Makady A, Sarri G, Mohamed O, Babar Z and Dawoud DM (2024), Editorial: Novel methods and technologies for the evaluation of drug outcomes and policies. *Front. Pharmacol.* 15:1396034. doi: 10.3389/fphar.2024.1396034

COPYRIGHT
© 2024 Adamson, Makady, Sarri, Mohamed, Babar and Dawoud. This is an open-access article distributed under the terms of the Creative Commons Attribution License (CC BY). The use, distribution or reproduction in other forums is permitted, provided the original author(s) and the copyright owner(s) are credited and that the original publication in this journal is cited, in accordance with accepted academic practice. No use, distribution or reproduction is permitted which does not comply with these terms.

Editorial: Novel methods and technologies for the evaluation of drug outcomes and policies

Blythe Adamson[1,2]*, Amr Makady[3], Grammati Sarri[4], Omneya Mohamed[5], Zaheer Babar[6] and Dalia M. Dawoud[7,8]

[1]Flatiron Health, New York, NY, United States, [2]The Comparative Health Outcomes, Policy and Economics (CHOICE) Institute, University of Washington, Seattle, WA, United States, [3]Janssen-Cilag B.V., Breda, Netherlands, [4]Cytel, London, United Kingdom, [5]Baxter AG Scientific Office, Dubai, United Arab Emirates, [6]Department of Pharmacy, University of Huddersfield, Huddersfield, United Kingdom, [7]Faculty of Pharmacy, Cairo University, Cairo, Egypt, [8]National Institute for Health and Care Excellence (NICE), London, United Kingdom

KEYWORDS

methods, analytic, artificial intelligence, causal inference, health technology assessment (HTA), HEOR

Editorial on the Research Topic
Novel methods and technologies for the evaluation of drug outcomes and policies

Globally, providing quality, equitable healthcare by accelerating patient access to new, promising health technologies while balancing the impact of their increased expenditures remains a global challenge. In parallel, the landscape of techniques and tools available to evaluate the safety and effectiveness of drugs is rapidly evolving with the advent of novel technologies and methodologies, thereby re-inventing the way we evaluate health outcomes and policies. This Research Topic of Frontiers in Pharmacology presents a compelling collection of scientific papers that delve into these advancements, offering insights into the latest developments in this dynamic field.

Artificial intelligence (AI) and machine learning (ML) methodologies are key themes explored in three papers within this Research Topic. The study by Zitu et al. on the generalizability of ML methods in detecting adverse drug events from clinical narratives in electronic medical records is a testament to the potential of AI in enhancing drug safety monitoring. Adamson et al. application of AI and ML in extracting real-world data from electronic health records (EHRs) is a stride forward in oncology research. This approach exemplifies how technology can enhance the curation of health records into valuable data sources. Vithlani et al. systematically review the conduct and reporting of health economic evaluations for AI-based healthcare interventions. Their work reveals the rapid growth in this area and the necessity for specific reporting standards to enhance transparency and decision-making in AI intervention evaluations. Some believe the increasing use of AI and ML raises ethical concerns regarding data privacy, bias, and transparency.

There was vital discussion around dominant Research Topic in Health Technology Assessment (HTA); integrating real-world evidence (RWE) in HTA, expanding analytical approaches (cost-effectiveness, equity-informed analyses) to include considerations beyond clinical and economic value drivers, and exploring new, reactive HTA approaches. Claire et al. bring forward insights, drawn from a multi-stakeholder workshop, to address the slow adoption of RWE in HTA compared to regulatory processes and the underlying reasons for

staying behind. They emphasize the need for developing resources to promote best practices for conducting RWE studies, comprehensive training, stakeholder collaboration, and impactful research projects to bridge this gap, thereby enhancing HTA's evidence base for informed healthcare decisions. Muir et al. review on integrating additional value elements in HTA modeling methods is a call to broaden the scope of health technology assessments. By incorporating societal values and health equity, their work advocates for a more holistic approach to evaluating new therapies. In the same direction, Zebrowska et al. and team's groundbreaking work on quantifying the impact of novel metastatic cancer therapies on health inequalities is a sobering reminder of the disparities in healthcare by offering an example of equity-informed analysis. Their study highlights how advancements in treatments may inadvertently widen the survival gap among different patient groups, emphasizing the need for more equitable healthcare solutions.

Cheyne et al. draw parallels between "living" clinical practice guidelines and HTA. Their reflections on incorporating continuous evidence synthesis in HTA processes offer a new paradigm in healthcare evaluation, ensuring that HTA remains responsive and current in a rapidly evolving evidence landscape.

Moving to Research Topic on advanced RWD analysis techniques in health economics and outcomes research (HEOR), the selected articles presented solutions for evaluating effectiveness and safety for new drugs in rare and very rare diseases and presented case study applications in causal inference and pharmacoepidemiology. Mackay and Springford advocacy for Bayesian methods in evaluating treatments for rare indications addresses a critical gap in HEOR. They argue for the use of Bayesian approaches to overcome challenges in small sample sizes and disconnected evidence networks, paving the way for more nuanced and robust analysis in rare disease settings. Franchini et al. Introduce an innovative approach in discrete event simulation, focusing on event-specific probabilities and distributions, especially in the context of censored data. Their methodological advancements in modeling competing events hold significant promise for more accurate and nuanced analysis in pharmacoeconomic studies.

Causal inference principles applied by Polito et al. and team to external control analysis in observational data is a noteworthy contribution. By defining the estimand attributes and selecting appropriate estimators, their study offers a refined approach to evaluating long-term survival outcomes in metastatic non-small cell lung cancer. Jeong et al. and team's use of network analysis to elucidate the dynamic landscape of drug-drug interactions offers a novel perspective. Their work underscores the potential of computational methods in identifying key research areas and informing clinical practice. The analysis of the FDA Adverse Event Reporting System by Su et al. offers a deep dive into the adverse event profiles of Denosumab and Zoledronic acid. Their findings provide invaluable insights for clinicians and policymakers, highlighting the importance of ongoing safety monitoring in pharmacovigilance.

Finally, the living systematic review by Elvidge et al. provides a crucial update on the economic evaluations of COVID-19 diagnostics and treatments emphasizing the need for a real time, regularly updated decision-making. Two years into the pandemic, their work synthesizes cost-effectiveness evidence for various interventions, highlighting the importance of making informed healthcare decisions in the rapidly changing landscape of COVID-19 management as new data emerges. This study underscores the ongoing need for living robust economic evaluations in guiding healthcare strategies, especially in a rapidly-changing pandemic setting.

This Research Topic not only reflects the recent trends in HEOR and rapid advancements in drug evaluation and policy research but also underscores the need for continuous adaptation and integration of novel methods in healthcare decision-making. It remains a potential challenge in accessing and implementing novel technologies, particularly in resource-limited settings. From the economic evaluations of emerging therapies to the cutting-edge use of AI and ML in data analysis, these studies collectively push the boundaries of current knowledge, paving the way for more informed, efficient, and equitable healthcare systems.

Author contributions

BA: Conceptualization, Investigation, Writing–original draft, Writing–review and editing. AM: Writing–review and editing. GS: Writing–review and editing. OM: Writing–review and editing. ZB: Writing–review and editing. DD: Supervision, Writing–review and editing.

Conflict of interest

Author BA was employed by Flatiron Health. Author AM was employed by Janssen-Cilag B.V. Author GS was employed by the Cytel. OM was employed by Baxter AG.

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.

The author(s) declared that they were an editorial board member of Frontiers, at the time of submission. This had no impact on the peer review process and the final decision.

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.

Supplementary material

The Supplementary Material for this article can be found online at: https://www.frontiersin.org/articles/10.3389/fphar.2024.1396034/full#supplementary-material

SUPPLEMENTARY FIGURE S1
Art by Catherine Au Yeung.

frontiers | Frontiers in Pharmacology

TYPE Original Research
PUBLISHED 12 July 2023
DOI 10.3389/fphar.2023.1218679

Check for updates

OPEN ACCESS

EDITED BY
Blythe Adamson,
Flatiron Health, United States

REVIEWED BY
Wen Zhang,
Huazhong Agricultural University, China
Robin Linzmayer,
Flatiron Health, United States

*CORRESPONDENCE
Lang Li,
✉ Lang.Li@osumc.edu

RECEIVED 07 May 2023
ACCEPTED 26 June 2023
PUBLISHED 12 July 2023

CITATION
Zitu MM, Zhang S, Owen DH, Chiang C and Li L (2023), Generalizability of machine learning methods in detecting adverse drug events from clinical narratives in electronic medical records. *Front. Pharmacol.* 14:1218679. doi: 10.3389/fphar.2023.1218679

COPYRIGHT
© 2023 Zitu, Zhang, Owen, Chiang and Li. This is an open-access article distributed under the terms of the Creative Commons Attribution License (CC BY). The use, distribution or reproduction in other forums is permitted, provided the original author(s) and the copyright owner(s) are credited and that the original publication in this journal is cited, in accordance with accepted academic practice. No use, distribution or reproduction is permitted which does not comply with these terms.

Generalizability of machine learning methods in detecting adverse drug events from clinical narratives in electronic medical records

Md Muntasir Zitu[1], Shijun Zhang[1], Dwight H. Owen[2], Chienwei Chiang[1] and Lang Li[1]*

[1]Department of Biomedical Informatics, College of Medicine, The Ohio State University, Columbus, OH, United States, [2]Department of Internal Medicine, College of Medicine, The Ohio State University, Columbus, OH, United States

We assessed the generalizability of machine learning methods using natural language processing (NLP) techniques to detect adverse drug events (ADEs) from clinical narratives in electronic medical records (EMRs). We constructed a new corpus correlating drugs with adverse drug events using 1,394 clinical notes of 47 randomly selected patients who received immune checkpoint inhibitors (ICIs) from 2011 to 2018 at The Ohio State University James Cancer Hospital, annotating 189 drug-ADE relations in single sentences within the medical records. We also used data from Harvard's publicly available 2018 National Clinical Challenge (n2c2), which includes 505 discharge summaries with annotations of 1,355 single-sentence drug-ADE relations. We applied classical machine learning (support vector machine (SVM)), deep learning (convolutional neural network (CNN) and bidirectional long short-term memory (BiLSTM)), and state-of-the-art transformer-based (bidirectional encoder representations from transformers (BERT) and ClinicalBERT) methods trained and tested in the two different corpora and compared performance among them to detect drug–ADE relationships. ClinicalBERT detected drug–ADE relationships better than the other methods when trained using our dataset and tested in n2c2 (ClinicalBERT F-score, 0.78; other methods, *F*-scores, 0.61–0.73) and when trained using the n2c2 dataset and tested in ours (ClinicalBERT F-score, 0.74; other methods, *F*-scores, 0.55–0.72). Comparison among several machine learning methods demonstrated the superior performance and, therefore, the greatest generalizability of findings of ClinicalBERT for the detection of drug–ADE relations from clinical narratives in electronic medical records.

KEYWORDS

adverse drug events, electronic health records, machine learning, natural language processing, relation extraction

1 Introduction

Adverse drug events (ADEs) are unintended harmful effects of taking medication (Hohl et al., 2018), which is a leading cause of death in the United States (Classen et al., 1997; Binkheder et al., 2022) and responsible for the hospitalization of 9,440,757 patients from 2008 to 2011, with an increasing trend over time (Poudel et al., 2017). The estimated annual

cost of drug-related morbidity and mortality resulting from non-optimized medication therapy was $528.4 billion, equivalent to 16% of total US healthcare expenditures, in 2016 (Watanabe et al., 2018). Patients with ADEs have demonstrated significantly longer hospital stays and an almost two-fold greater risk of death than patients without ADEs (Classen et al., 1997). Nevertheless, ADEs are mostly preventable (Rommers et al., 2007), and early detection can substantially reduce morbidity and, thereby, decrease associated healthcare costs (Classen et al., 1997; Kaushal et al., 2006; Handler et al., 2007).

ADEs are largely detected after marketing, so timely surveillance at this time is important for patient safety (Botsis et al., 2011; Polepalli Ramesh et al., 2014). Pharmacovigilance has traditionally employed spontaneous reporting systems (SRSs), but as many as 90% of ADEs may remain unreported in this voluntary scheme (Hazell and Shakir, 2006). In contrast, electronic health records (EHR) represent a potentially great source for post-marketing surveillance of drug safety, accommodating real-time clinical data gathered from routine clinical care (Coloma et al., 2013). One study revealed relevant ADE information, for example, in the structured data of 9,020 of 31,531 patients (28.6%) with side effects of statin documented in provider notes (Skentzos et al., 2011). Furthermore, clinical notes in EHRs provide longitudinal information related to drug-induced adverse events, but the manual review and extraction of ADEs from enormous clinical narratives is labor intensive, and clinical notes in EHRs vary from patient to patient, physician to physician, and hospital to hospital. Therefore, an automated system that utilizes artificial intelligence (AI) is needed to extract ADEs from clinical notes, and attempts have been made to build such a system.

The Medication and Adverse Drug Events Challenge (MADE1.0) (Jagannatha et al., 2019) aimed to automatically identify clinical concepts and relations from clinical narratives that included ADEs. The Challenge included three tasks: 1) naming the recognized entity (NER) and identifying the medication and its route, dosage, duration, frequency, and indication, as well as associated ADEs and their severity; 2) identifying relations (RI) of medications with ADEs, indications, and other entities; and 3) performing the NER and RI tasks jointly. The Challenge released 1,089 fully de-identified clinical notes from 21 randomly selected patients with cancer at the University of Massachusetts Memorial Hospital that included 2,612 drug–ADE relations. Methods used to classify relations ranged from statistical machine learning (ML)-based methods, such as support vector machine learning (SVM), random forest, and others, to neural-network-based bidirectional long short-term memory (BiLSTM). The best-performing model for the classification of ADE–drug name relations achieved an *F*-score of 0.72.

Another effort, the n2c2 Shared Task Challenge (Henry et al., 2020), mirrored MADE1.0 and included similar tasks. The n2c2 dataset comprised 505 discharge summaries taken from the Medical Information Mart for Intensive Care-III (MIMIC-III) clinical care database (Johnson et al., 2016). Records were selected by searching ADEs in the International Classification of Diseases (ICD) code descriptions of the records, which yielded a total of 1,840 ADE-drug relations. Methods used for the relation-classification task ranged from SVM to attention-based BiLSTM, with the best-performing model for ADE-drug name relation yielding an *F*-score of 0.85.

The use of data from a single EHR source in the two challenges allowed NLP approaches developed from these models to be data-specific. The model that performed the best (Wei et al., 2020) in relation extraction for the n2c2 Shared Task, for example, contained a post-processing technique that was dependent on the n2c2 annotation guideline to improve the performance of their BiLSTM conditional random field (CRF) algorithm. This n2c2 paper (Wei et al., 2020) also suggested that language models like BERT (Devlin et al., 2018) in the biomedical domain may further improve the NLP performance, and it remains to be an interesting topic for future research. In the best-performing model (Chapman et al., 2019) from the MADE 1.0 Challenge, the best model was random forest. The paper (Chapman et al., 2019) stated that generalizability of the best-performing model was unclear.

Though transformer-based methods like BERT (Devlin et al., 2018), BioBERT (Lee et al., 2020), and ClinicalBERT (Alsentzer et al., 2019) have become popular in recent years, they have not been applied to identify ADEs from the clinical notes of EHRs. Therefore, it is critical to implement the same model on data from different data sources to assess the generalizability of a model. No study has been conducted yet that used EHRs from different data sources to address the issue of generalizability. BERT-based methods have performed better than other machine and deep learning methods (Sun et al., 2019; González-Carvajal and Garrido-Merchán, 2020; Minaee et al., 2022) in mining biomedical texts, and methods that integrate biomedical corpora, such as ClinicalBERT, outperformed BERT models (Huang et al., 2019). In this article, we investigate the performance of deep learning (CNN, BiLSTM) and transformer-based (BERT and ClinicalBERT) methods, as well as classical SVM, in clinical notes of two different EHR systems.

The preprocessing of data is important in NLP in clinical text and challenged by the inherent variations in EHRs, and the preprocessing of clinical text in EHRs, such as tokenization, which breaks the text into meaningful elements and detects sentence boundaries, is different and more challenging than the processing of data in the literature (Griffis et al., 2016). Publications often inadequately present the end-to-end data preprocessing method, but in this paper, we sufficiently detail the end-to-end data preprocessing for all NLP methods used.

Our primary goal is to address whether and how ADE NLP algorithms developed from the n2c2 Shared Task can be generalized to a drug-specific ADE in a different EHR. There are two types of generalizability in the NLP algorithm development for detecting drug-induced ADEs. First, we want to know whether the NLP model developed in one dataset maintains a comparable performance in a different dataset. Second, if one NLP method has a better performance than the others, will it maintain its supreme performance in a different dataset? In this paper, we study both types of generalizability for NLP algorithms designed for drug–ADE relationship detection. We will use ADEs induced by immune checkpoint inhibitors (ICI) (Nashed et al., 2021) as recorded in clinical notes as examples for analysis. MADE1.0 (Jagannatha et al., 2019) data were not available to us when we conducted this study.

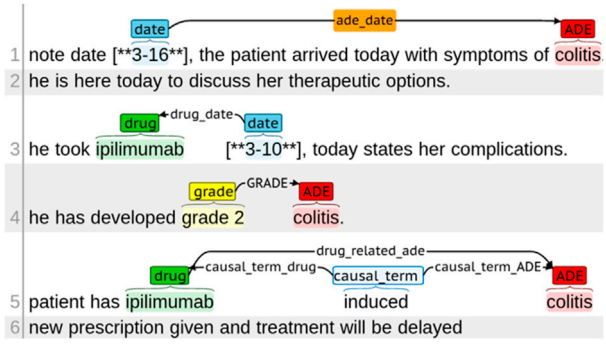

FIGURE 1
Annotation example for different entities in the ICI-OSU corpus.

2 Materials and methods

2.1 Datasets

We used two datasets to analyze the generalizability of ADE detection from EHRs; one was developed from the EHR system of The Ohio State University James Cancer Hospital, and the second was that of the n2c2 Shared Task (Henry et al., 2020).

2.2 ICI-OSU corpus

We built the ICI-OSU corpus by manually annotating 1,394 clinical notes of 47 randomly selected patients who received immune checkpoint inhibitors (ICIs) from 2011 to 2018 at The Ohio State University James Cancer Hospital. Supplementary materials contain a detailed annotation guideline that we developed to assist the annotators with manual annotation. Supplementary Table S1 contains the annotation guideline for entity annotation, whereas Supplementary Table S2 contains the guideline for relation annotation. Figure 1 shows different types of entity tags and relation annotations.

The entities included drug names, mentions of ADEs, dates drugs were taken, dates of ADEs, terms drawing causal relation between a drug and an ADE, and grades of ADE intensity. It is worth mentioning that *reasons for drug administration* were not considered as ADEs. Supplementary Figure S1 shows the difference between ADE and *reason for drug administration* with an example. Relations were annotated for entities located within a single sentence and across sentences. Two annotators with informatics skills and knowledge in cancer clinical trials independently annotated each note, and a third annotator performed the validation for inconsistent annotations between the two annotators. One annotator was a resident physician with hands-on experience with EHRs; the second had a master's degree in biology with 7 years of experience in corpus development and annotation; the third annotator, who performed the validation, was a graduate student in biomedical informatics. The institutional review board of The Ohio State University approved this study (#2020C0145).

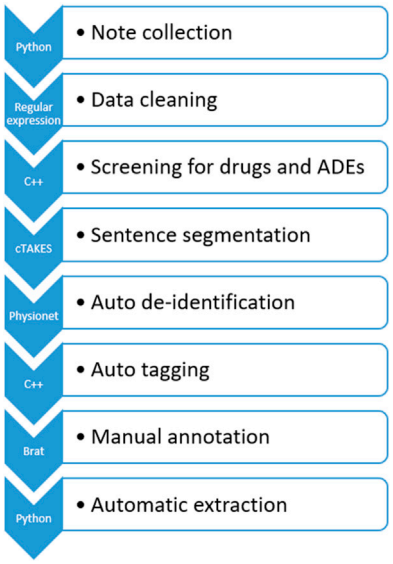

FIGURE 2
ICI-OSU data processing pipeline. Python and C++ programming were used at different steps to automate the process.

TABLE 1 List of drug names and adverse drug events. We used all possible mentions of these drugs and ADEs in our study.

Drug	ADE
Atezolizumab	Colitis
Ipilimumab	Pneumonitis
Nivolumab	Thyroid
Pembrolizumab	Abnormalities
Tremelimumab	Rash/dermatitis
Avelumab	Hepatitis
Durvalumab	Myalgia/arthralgia
Cemiplimab	Cardiotoxicity

2.3 ICI-OSU data processing pipeline

Figure 2 illustrates the data processing pipeline we built that feeds processed data to the ADE NLP models for generalizability analysis.

In the first step, we collected the clinical notes of the targeted patient. The ICI-OSU corpus contains the patient's notes for the first 12 months from the date of the first ICI dose in the structured data. The order of note dates is maintained to keep track of longitudinal data, such as the date of first drug use, date of first ADE, and date of drug discontinuation.

In the second step, we cleaned data using regular expression techniques (Wang et al., 2019), which included but were not limited to normalizing uneven spaces and drug names and manipulating abbreviations. For drug name normalization, we used the DrugBank

database (Wishart et al., 2008) as our primary source, and we built a lookup table that contained all possible mentions (such as the generic and brand name) of a drug. Supplementary Figure S2 shows a raw original clinical note, and Supplementary Figure S3 shows the cleaned data.

In the third step, we performed automatic screening for drugs and ADEs, tagging predefined drug and ADE terms for follow-up manual annotation. Table 1 delineates the drug names and the ADE list of our study. We used the Common Terminology Criteria for Adverse Events (CTCAE) (Freites-Martinez et al., 2021) and the DrugBank database as our primary guidelines for ADE and drug mentions, respectively. Then, we built a lookup table for screening that contained all possible mentions of these drugs and ADEs found in EHRs. We discussed with a physician and annotators who had hands-on experience working with EHRs, thus enriching and verifying our lookup table to ensure accuracy.

In the fourth step, we used the Apache clinical Text Analysis and Knowledge Extraction System (cTAKES)™ (Savova et al., 2010), an open-source NLP tool, to segment sentences in the EHRs, which is one of the most challenging tasks because of variations, such as in use of punctuation and abbreviations, that are unique to the recorder (Griffis et al., 2016). To optimize cTAKES™ performance on sentence segmentation, we also encrypted the unexpected line breakers to be consistent with the cTAKES™ segmentation rules. After segmentation, we decrypted the data to preserve their originality. Supplementary Figure S4 shows the data's appearance after sentence segmentation in cTAKES™.

In the fifth step, we performed automatic de-identification using de-identification software (Neamatullah et al., 2008) from PhysioNet (Goldberger et al., 2000) followed by manual de-identification by the annotators to ensure accuracy. The recall value of PhysioNet was over 98% in the de-identification task on our dataset. Supplementary Figure S5 shows the data after de-identification.

In the sixth step, we performed automatic tagging, annotating the drug name, mention of ADE, and date of ADE, to reduce the burden of subsequent manual annotation. For drug and ADE annotation, we performed exact matching of the data with our predefined drug and ADE lists. We used a rule-based method to annotate the dates in the clinical text automatically, and this annotation was consistent with that of the web-based brat rapid annotation tool (BRAT) (Stenetorp et al., 2012), which we used later for manual annotation. Though automatic annotation greatly reduced the burden of manual annotation, the annotators were free to annotate any new form of a drug name, mention of ADE, or date of ADE. The annotators could also correct anything incorrectly tagged in automatic tagging. Supplementary Figure S6 shows the automatically annotated notes.

In the seventh step, our annotators performed manual annotation in BRAT following the well-defined guideline mentioned previously and included in the supplementary materials. Supplementary Figure S7 shows the manually annotated notes.

In the eighth and final step, we extracted relevant annotation information automatically after manual annotation to build our corpus. Information extracted to prepare the data for machine learning models included drug–ADE relations, drug–ADE pairs, and neighboring words. We built an automatic system to extract that information from the annotated corpus based on the input format of the ML models. Due to the repetitive nature of EHRs, several drug–ADE relations were repeated when the text between a drug and an ADE, as well as the context, were exactly the same. We removed those duplicates in the automatic extraction.

2.4 n2c2 Shared Task corpus

The n2c2 data consisted of information from 505 discharge summaries taken originally from the MIMIC-III clinical care database (Johnson et al., 2016). The data provider described their process as first searching for ADEs in the ICD code description of each record and then manually screening the records with at least one ADE and dividing the data into a training set comprising 303 annotated files and a testing set that included 202 files (Henry et al., 2020). The n2c2 data contained several clinical concepts and relations as well as drug and ADE annotations, and we performed preprocessing as described previously to prepare the data for the ML models, first cleaning the n2c2 data, then segmenting sentences using cTAKES™, and finally using our automatic system to extract relevant annotations.

2.5 The definition of positive and negative drug–ADE relations in the n2c2 and ICI-OSU corpora

Like other researchers (Wei et al., 2020), we considered all possible combinations of drugs and ADEs to build positive and negative data for training and validating NLP models. Our generalizability analysis focused on drug–ADE relations within a single sentence; so, for example, for a sentence containing the drugs *d1* and *d2* and the ADEs *a1* and *a2,* the four possible drug–ADE combinations are (*d1*, *a1*), (*d1*, *a2*), (*d2*, *a1*), and (*d2*, *a2*). A drug–ADE relation was considered positive if the drug induced the ADE. We collected the positive samples directly from annotation to build the positive dataset. A relation was considered negative if the drug did not induce the ADE and was, therefore, not annotated in the corpus. We derived the negative dataset from all the drug–ADE combinations by subtracting the annotated positive set from the corpus. After removing duplicates, we obtained 189 positive samples and 698 negative samples from our annotated ICI-OSU data. The default n2c2 training and test data yielded 1,355 positive and 865 negative samples after duplicates were removed.

2.6 Machine learning deep learning models

We implemented several machine learning, deep learning, and transformer-based models, including SVM (Joachims, 1998), CNN (Kim, 2014), BiLSTM (Sherstinsky, 2018; Xu et al., 2019), BERT (Devlin et al., 2018), and ClinicalBERT (Alsentzer et al., 2019), to analyze the n2c2 and ICI-OSU datasets, and we trained these models on one dataset and validated them on the other to analyze the generalizability of their findings.

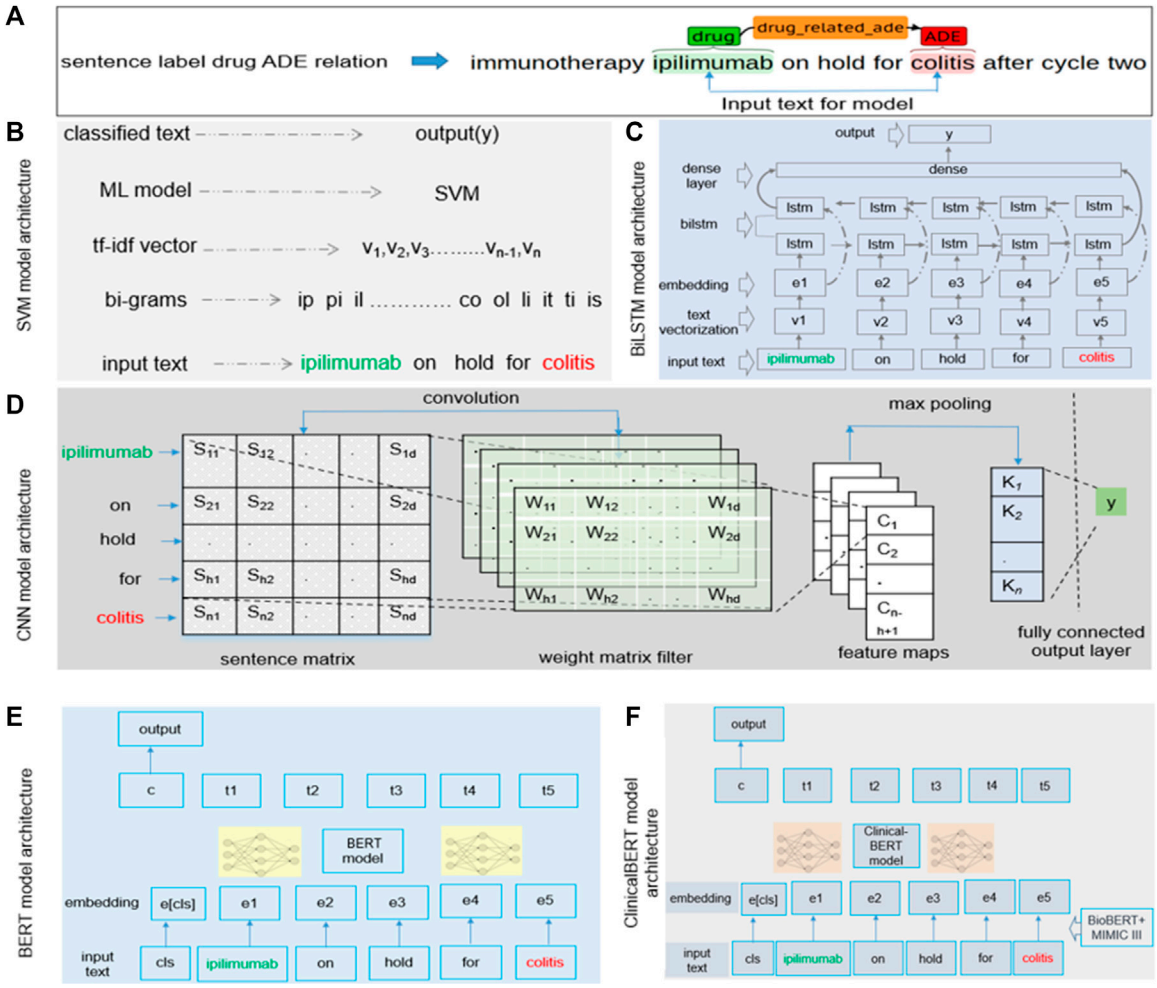

FIGURE 3
(A) Input text processing from single-sentence drug–ADE relation. **(B–F)** Implementation details of **(B)** SVM, **(C)** BiLSTM, **(D)** CNN, **(E)** BERT, and **(F)** clinicalBERT models.

2.6.1 Training and validation data

Training and validation data were speculated under the intra- and inter-dataset settings. The intra-dataset setting comprised training and validation data allocated from the same data source, either n2c2 or ICI-OSU. When training and validating the model on the same n2c2 dataset in the intra-set setting, we used the default training and validation data of the n2c2 dataset given by the data providers. Because the ICU-OSU dataset was relatively smaller, we applied five-fold cross-validation to avoid overfitting when we trained and validated our model on the same ICU-OSU dataset in the intra-set setting. In the inter-set setting, training was performed on one dataset, and its validation was performed in the other. We also withheld 30 percent of the data from the training set to serve as the internal validation set in the inter-dataset setting.

2.6.2 Hyperparameter selection and embedding

In the deep learning and transformer-based models, we studied different dropout rates ranging between 0.1 and 0.8 and learning rates (0.1, 0.01, 0.001, and 0.0001) with batch sizes of 16, 32, and 64. In those models, we used embedding (Mikolov et al., 2013) techniques and pre-trained word-embedding models, such as Stanford's global vectors for word representation (GloVe) (300-dimensional) (Pennington et al., 2014). In our transformer-based model, we also incorporated biomedical domain knowledge.

2.6.3 Preparation of input data, feature selection, and model implementation

We focused on the contextual features while building the models. Because the text between a drug and an ADE contains most of the contextual information regarding a drug–ADE relation, we considered the drug name, the ADE, and the text between them as the input for all models. However, feature selection varied from model to model. Our SVM model, for example, used character-level *n*-grams, whereas the CNN model considered *n*-grams at the token level. Figure 3A shows how we extracted the input text from a drug–ADE relation.

2.7 Implementation of the support vector machine model

We generated character *n*-grams from the first character of a sample to the last character of the input text using a range of values for *n*, converted those *n*-grams in term frequency (TF)-inverse document frequency (IDF) (Qaiser and Ali, 2018) vectorization, and then finally fed the feature vector into the model to predict the output *y* (0 or 1). We searched a range of values for *c* and gamma to obtain the best hyperparameter set and used the radial basis function (RBF) kernel in our SVM model. Figure 3B details the implementation of our SVM model with an example bigram.

2.8 Bidirectional long short-term memory (BiLSTM)

Understanding the context of a sentence is critical and requires that classification of the sentence includes information in both directions, from the beginning of the sentence to its end and from its end to its beginning. Previous studies showed promising results using BiLSTM to extract contextual information (Xu et al., 2019). BiLSTM algorithms can learn long-term dependencies and work in both directions of text and learn contextual features in a given time stamp. For our model, the input was a sequence from the start to the end of an entity of a sample. Figure 3C shows the architecture of the BiLSTM model. We performed text vectorization on the input and then used the pre-trained word embedding of GloVe (300d) (Pennington et al., 2014) in the embedding layer. The BiLSTM layer was used over the embedding layer, and finally, the dense layer was used, producing the output, *y*.

2.9 Convolutional neural network (CNN)

We implemented the standard CNN model (Kim, 2014), in which we focused on contextual information while extracting features. The CNN model applies a filter to extract features from text and uses those features to classify the text. As mentioned earlier, we used the text between the start of an entity and its end to analyze sequential words to learn features to extract. We used multiple filters of different sizes (2, 3, and 4) to examine different *n*-grams within the text. Figure 3D shows the architecture of the CNN model. We built a sentence matrix, with rows indicating the tokens of a sentence and columns indicating the features in which we implemented pre-trained word embedding. S_n is the number of tokens in a sentence; S_d is the feature dimension of a token; W_h is the kernel size. We then applied weight filters for the convolution operation and feature mapping and, finally, applied max pooling and fully connected the output layer to generate output. We used the rectified linear unit (ReLU) activation function and the pre-trained word embedding of GloVe (300d) in the embedding layer (Pennington et al., 2014).

TABLE 2 Agreement between annotators.

Type	*F*-score
Drug	99.00%
ADE	95.12%
Grade	70.66%
Causal term	73.58
Drug–ADE	70.94

2.10 Bidirectional encoder representations from transformers (BERT)

The BERT (Lee et al., 2020) model is based on a transformer encoder that uses a self-attention mechanism for sequence modeling (Vaswani et al., 2017). For our classification task, the sequential information of the text could be important. We selected a segment from the start to the end of an entity of a sample and then used the tokenized segment as input in the BERT model. The class token *cls* was added with the input. To obtain embeddings for our text, we used the pre-trained transformer model rather than an embedding layer. For our analysis, we used BERT-base, which consists of transformer blocks of 12 layers with 12 self-attention heads, a hidden size of 768, and 110 M trainable parameters (Lee et al., 2020). Figure 3E shows the architecture of the BERT model.

2.11 ClinicalBERT

The implementation of the ClinicalBERT (Alsentzer et al., 2019) model was similar to that of our BERT model, but ClinicalBERT incorporated pre-trained biomedical domain knowledge. We used the model that was initialized on BioBERT (Lee et al., 2020) and trained on all notes of the MIMIC-III dataset (Johnson et al., 2016). The model had a batch size of 32 and a maximum sequence length of 128 (Alsentzer et al., 2019). Figure 3F shows the architecture of the ClinicalBERT model.

3 Results

Annotation performance: we performed annotation in two rounds. In round one, we annotated 118 single-sentence positive drug–ADE relations and 24 cross-sentence positive drug–ADE relations. In round two, we labeled 163 single-sentence positive drug–ADE relations and 27 cross-sentence positive drug–ADE relations. Table 2 shows the average inter-annotator agreement (IAA) results of our two rounds of annotations. We calculated Cohen's kappa (McHugh, 2012) to measure IAA. The results indicate considerable disagreement between the two annotators regarding the identification of drug–ADE relations, which is probably attributable to the diverse nature of ADE mentions in clinical notes. Supplementary Table S3 shows more annotation results and findings details of our corpus. Our OSU-ICI corpus is the first drug class-specific drug–ADE corpus. By specifically targeting ICIs, it also becomes a golden standard for developing immunotherapy-induced adverse event phenotypes.

FIGURE 4
Comparison of precision, recall, and *F*-score among different models in intra- and inter-dataset settings. **(A, B)** show the performance of the different machine learning models when trained and tested on the same data source (intra dataset). **(C, D)** show the performance of the different machine learning models when trained and tested on different data source (inter dataset).

3.1 Performance evaluation and error analysis

To analyze the generalizability of findings among the models, we trained and tested each model on the dataset of the same data source (intra-data) and a different source (inter-data). Figures 4A, B show the performance of the different models in the intra-dataset setting. The CNN-based deep learning model performed the best, with an *F*-score of 80% for the ICI-OSU dataset, whereas the transformer-based ClinicalBERT model achieved the best *F*-score (87%) for the n2c2 data. The SVM model was also competitive and showed stable performance in the intra-dataset setting. Our results demonstrate that we did not get the best performance from either the BERT or ClinicalBERT model when training and testing on the ICI-OSU dataset. This is probably because the transformer models usually perform better on a large dataset, and the ICI-OSU dataset is small. Figures 4C, D show the performance of the different models in the inter-dataset setting, with the transformer-based models showing superior performance. The ClinicalBERT model achieved the highest F-score in both combinations of inter-data training and testing, probably because of the incorporation of domain knowledge as ClinicalBERT was trained on the MIMIC-III dataset. These performances indicate the importance of incorporating domain knowledge in models. Figure 5 shows the cluster map for the results of different models in the inter-dataset setting. In this NLP analysis of four combinations of training sets and test sets from n2c2 and ICI-OSU datasets, it is evident that ClinicalBERT has the best performance in three out of four combinations.

Figures 5A, B show the cluster map while training on the n2c2 dataset and testing on the ICI-OSU dataset. The BERT and ClinicalBERT models performed similarly for positive sample prediction, whereas the SVM and BiLSTM models clustered together with similar results. BiLSTM and ClinicalBERT models performed similarly for negative sample prediction. Figures 5C, D show the cluster map while training on the ICI-OSU dataset and testing on the n2c2 dataset. The BERT and ClinicalBERT models performed similarly for both positive and negative sample prediction, and the SVM and BiLSTM models clustered together with similar results. Comparison of the results of Figures 4 and 5 demonstrates the better performances of the machine learning and deep learning models in the intra-dataset setting and of the transformer-based models in the inter-dataset settings. Along with the internal capacity of the models to recognize patterns, the variable length of the datasets might contribute to such performance. Supplementary Tables S4 and S5 further detail the results.

3.2 Factors contributing to error 1 (differences in data sources)

The n2c2 dataset was the more general of the two sets and looked for all drug mentions in a clinical note, whereas the ICI-OSU dataset

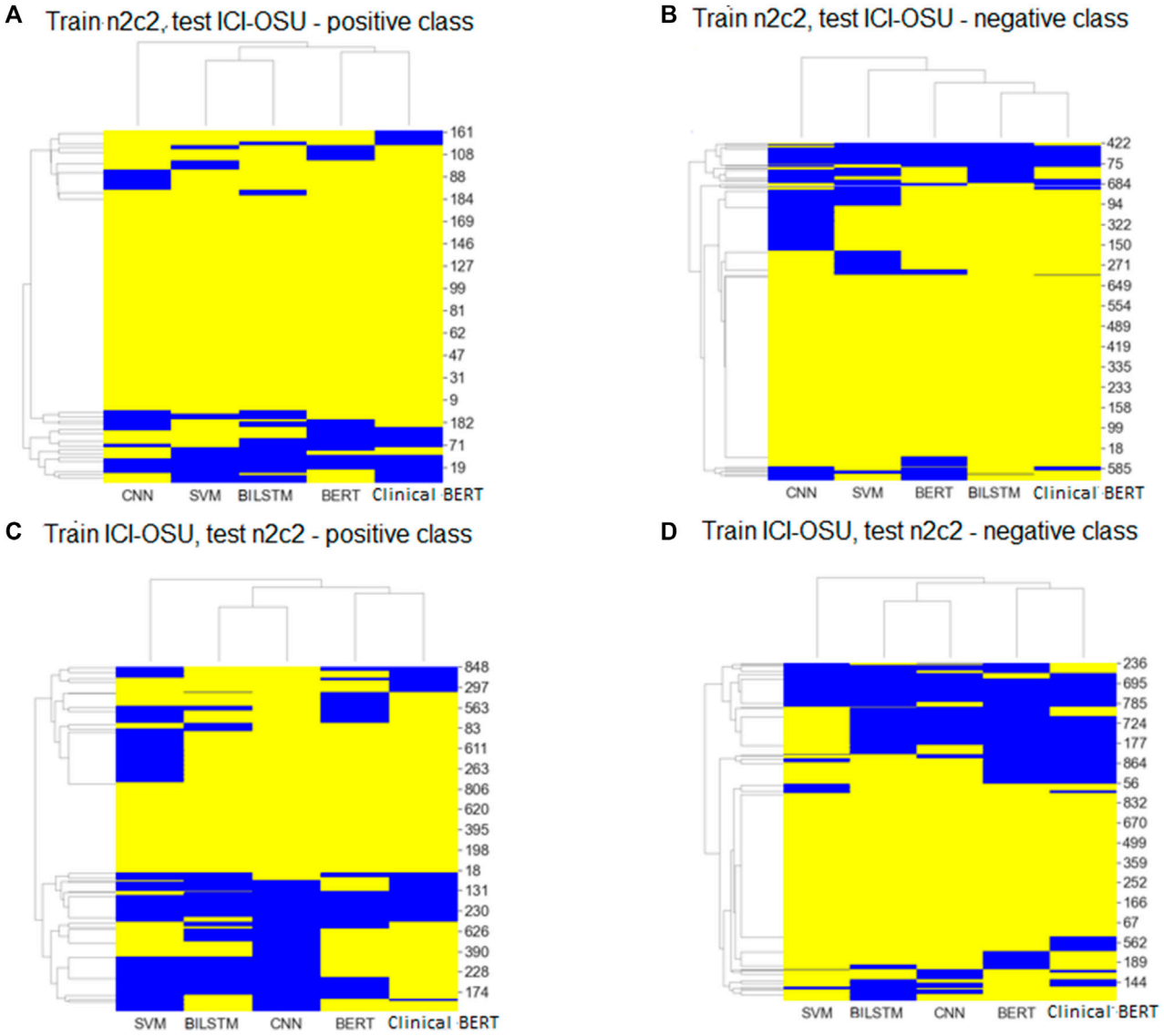

FIGURE 5
Cluster maps of results comparing predicted labels with true labels for every sample of positive and negative classes in inter dataset settings. Yellow: predicted and true label matched; blue: labels did not match. **(A, C)** shows the results for positive class. **(B, D)** shows the results for negative class.

focused on ICI drugs and annotated a specific set of ADEs. The use of these differing types of data challenged training on one source and testing on the other.

3.3 Factors contributing to error 2 (differences in annotation guidelines and causal terms)

The annotation guidelines for the two datasets differed. The guideline for the n2c2 dataset, for example, looked for the closest entity rather than causation to draw a relation, whereas the guideline for the ICI-OSU dataset looked for the presence of causal terms to identify a causal relation between entities.

4 Discussion

4.1 Incorporation of domain knowledge

We attempted to analyze the generalizability of ADE detection from clinical notes using several machine learning, deep learning, and transformer-based models and observed promising performance, particularly when we applied state-of-the-art transformer-based models. The superior performance of ClinicalBERT indicated the importance of incorporating domain knowledge when using pre-trained data. Thus, future studies should incorporate more domain knowledge to further enhance the performance of the models.

4.2 Cross-sentence relation

Our study focused on single-sentence drug–ADE relations, in which the drug and ADE occurred in the same sentence, and after sentence segmentation and annotation, we ended with only a few cross-sentence relations. However, it is also important to identify cross-sentence relations. Our primary challenge in identifying cross-sentence relations was the imbalance within a dataset that could pose a very large number of negative relations against a very small number of positive relations. In addition, our experience in manual annotation showed that a drug and ADE could be distantly related across sentences with numerous sentences in between. Nevertheless, limiting the number of sentences between two entities to draw a relation could help limit the search space to accommodate most of the positive relations across sentences. It would also keep the number of negative relations considerably low.

4.3 Variation in sentence length

Variations in sentence lengths, some only a few words and some extraordinarily long, made it difficult to train the model and contributed greatly to the error. The ICI-OSU dataset included two positive samples consisting of 30 tokens each, where every single model except BERT classified them incorrectly. BERT was able to predict the true label of one of those two samples correctly. A lack of similar training data probably contributed to the error. Having more training data of similar length or building a separate rule-based approach could facilitate the management of extraordinarily long sentences.

Data availability statement

The datasets presented in this article are not readily available because the clinical notes of EHRs were used in this study. Therefore, data could not be published at this point. Requests to access the datasets should be directed to LL, Lang.Li@osumc.edu.

Ethics statement

The studies involving human participants were reviewed and approved by the institutional review board of The Ohio State University (#2020C0145). Written informed consent for participation was not required for this study in accordance with the national legislation and the institutional requirements.

Author contributions

MZ and LL were responsible for the overall study design and writing the manuscript. MZ and SZ performed the end-to-end annotation task from the definition of annotation guidelines to the validation of manual annotation. MZ performed the implementation task, which included data processing and modeling. DO provided, enriched, and validated the drug and ADE lists. CC provided logistic support and made the original data available for study with the environmental setup. LL supervised the overall study. All authors contributed to the article and approved the submitted version.

Acknowledgments

We would like to acknowledge and thank Andrew Nashed for his support in this work.

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.

Supplementary material

The supplementary material for this article can be found online at: https://www.frontiersin.org/articles/10.3389/fphar.2023.1218679/full#supplementary-material

References

Alsentzer, E., Murphy, J. R., Boag, W., Weng, W., Jin, D., Naumann, T., et al. 2019. Publicly available ClinicalBERT embeddings

Binkheder, S., Wu, H-Y., Quinney, S. K., Li, L., Gao, Y., Skaar, T. C., et al. (2022). PhenoDEF: A corpus for annotating sentences with information of phenotype definitions in biomedical literature. *J. Biomed. Semant.* 13 (1), 17. doi:10.1186/s13326-022-00272-6

Botsis, T., Nguyen, M. D., Woo, E. J., Markatou, M., and Ball, R. (2011). Text mining for the vaccine adverse event reporting system: Medical text classification using informative feature selection. *J. Am. Med. Inf. Assoc.* 18 (5), 631–638. doi:10.1136/amiajnl-2010-000022

Chapman, A. B., Peterson, K. S., Alba, P. R., DuVall, S. L., and Patterson, O. V. (2019). Detecting adverse drug events with rapidly trained classification models. *Drug Saf.* 42 (1), 147–156. doi:10.1007/s40264-018-0763-y

Classen, D. C., Pestotnik, S. L., Evans, R. S., Lloyd, J. F., and Burke, J. P. (1997). Adverse drug events in hospitalized Patients<subtitle>Excess length of stay, extra costs, and attributable mortality</subtitle>. *JAMA* 277 (4), 301–306. doi:10.1001/jama.1997.03540280039031

Coloma, P. M., Trifirò, G., Patadia, V., and Sturkenboom, M. (2013). Postmarketing safety surveillance: Where does signal detection using electronic healthcare records fit into the big picture? *Drug Saf.* 36 (3), 183–197. doi:10.1007/s40264-013-0018-x

Devlin, J., Chang, M.-W., Lee, K., and Toutanova, K. (2018). Bert: Pre-training of deep bidirectional transformers for language understanding. https://arxiv.org/abs/1810.04805.

Freites-Martinez, A., Santana, N., Arias-Santiago, S., and Viera, A. (2021). Using the Common Terminology Criteria for adverse events (CTCAE - version 5.0) to evaluate the severity of adverse events of anticancer therapies. *Actas Dermosifiliogr. Engl. Ed.* 112 (1), 90–92. doi:10.1016/j.ad.2019.05.009

Goldberger, A. L., Amaral, L. A., Glass, L., Hausdorff, J. M., Ivanov, P. C., Mark, R. G., et al. (2000). PhysioBank, PhysioToolkit, and PhysioNet: Components of a new research resource for complex physiologic signals. *Circulation* 101 (23), E215–E220. doi:10.1161/01.cir.101.23.e215

González-Carvajal, S., and Garrido-Merchán, E. C., 2020. Comparing BERT against traditional machine learning text classification.

Griffis, D., Shivade, C., Fosler-Lussier, E., and Lai, A. M., 2016. *A quantitative and qualitative evaluation of sentence boundary detection for the clinical domain*. AMIA Jt Summits Transl Sci Proc, p.88–97.

Handler, S. M., Altman, R. L., Perera, S., Hanlon, J. T., Studenski, S. A., Bost, J. E., et al. (2007). A systematic review of the performance characteristics of clinical event monitor signals used to detect adverse drug events in the hospital setting. *J. Am. Med. Inf. Assoc.* 14 (4), 451–458. doi:10.1197/jamia.M2369

Hazell, L., and Shakir, S. A. W. (2006). Under-reporting of adverse drug reactions: A systematic review. *Drug Saf.* 29 (5), 385–396. doi:10.2165/00002018-200629050-00003

Henry, S., Buchan, K., Filannino, M., Stubbs, A., and Uzuner, O. n2c2 Shared Task Participants 2020 2018 n2c2 shared task on adverse drug events and medication extraction in electronic health records. *J. Am. Med. Inf. Assoc.* 27 (1), 3–12. doi:10.1093/jamia/ocz166

Hohl, C. M., Small, S. S., Peddie, D., Badke, K., Bailey, C., and Balka, E. (2018). Why clinicians don't report adverse drug events: Qualitative study. *JMIR Public Health Surveillance* 4 (1), e21. doi:10.2196/publichealth.9282

Huang, K., Altosaar, J., and Ranganath, R., 2019. ClinicalBERT: Modeling clinical notes and predicting hospital readmission.

Jagannatha, A., Liu, F., Liu, W., and Yu, H. (2019). Overview of the first natural language processing challenge for extracting medication, indication, and adverse drug events from electronic health record notes (MADE 1.0). *Drug Saf.* 42 (1), 99–111. doi:10.1007/s40264-018-0762-z

Joachims, T. (1998). "Text categorization with support vector machines: Learning with many relevant features," in *Machine learning: ECML-98. Ecml 1998*. Editors C. Nédellec, and C. Rouveirol (Berlin, Heidelberg: Springer), 1398. doi:10.1007/BFb0026683

Johnson, A. E. W., Pollard, T. J., Shen, L., Lehman, L. W. H., Feng, M., Ghassemi, M., et al. 2016. MIMIC-III, a freely accessible critical care database. *Sci. Data.* 3, 160035doi:10.1038/sdata.2016.35

Kaushal, R., Jha, A. K., Franz, C., Glaser, J., Shetty, K. D., Jaggi, T., et al. (2006). Return on investment for a computerized physician order entry system. *J. Am. Med. Inf. Assoc.* 13 (3), 261–266. doi:10.1197/jamia.M1984

Kim, Y. (2014). "Convolutional neural networks for sentence classification," in *Proceedings of the 2014 conference on empirical methods in natural language processing (EMNLP)* (Doha, Qatar: Association for Computational Linguistics), 1746–1751.

Lee, J., Yoon, W., Kim, S., Kim, D., Kim, S., So, C. H., et al. (2020). BioBERT: A pre-trained biomedical language representation model for biomedical text mining. *Bioinformatics* 36 (4), 1234–1240. doi:10.1093/bioinformatics/btz682

McHugh, M. L. (2012). Interrater reliability: The kappa statistic. *Biochem. Med. Zagreb.* 22 (3), 276–282. doi:10.11613/bm.2012.031

Mikolov, T., Sutskever, I., Chen, K., Corrado, G. S., and Dean, J. (2013). "Distributed representations of words and phrases and their compositionality," in *Advances in neural information processing systems*, 26.

Minaee, S., Kalchbrenner, N., Cambria, E., Nikzad, N., Chenaghlu, M., and Gao, J. (2022). Deep learning-based text classification: A comprehensive review. *ACM Comput. Surv. (CSUR)* 54 (3), 1–40. doi:10.1145/3439726

Nashed, A., Zhang, S., Chiang, C. W., Hwang, A., Riaz, N., Presley, C. J., et al. (2021). Comparative assessment of manual chart review and ICD claims data in evaluating immunotherapy-related adverse events. *Cancer Immunol. Immunother.* 70 (10), 2761–2769. doi:10.1007/s00262-021-02880-0

Neamatullah, I., Douglass, M. M., Lehman, L.-w. H., Reisner, A., Villarroel, M., Long, W. J., et al. (2008). Automated de-identification of free-text medical records. *BMC Med. Inf. Decis. Mak.* 8, 32. doi:10.1186/1472-6947-8-32

Pennington, J., Socher, R., and Manning, C. (2014). "GloVe: Global vectors for word representation," in *Proceedings of the 2014 conference on empirical methods in natural language processing (EMNLP)* (Doha, Qatar.

Polepalli Ramesh, B., Belknap, S. M., Li, Z., Frid, N., West, D. P., and Yu, H. (2014). Automatically recognizing medication and adverse event information from food and drug administration's adverse event reporting system narratives. *JMIR Med. Inf.* 2 (1), e10. doi:10.2196/medinform.3022

Poudel, D. R., Acharya, P., Ghimire, S., Dhital, R., and Bharati, R. (2017). Burden of hospitalizations related to adverse drug events in the USA: A retrospective analysis from large inpatient database. *Pharmacoepidemiol. Drug Saf.* 26 (6), 635–641. doi:10.1002/pds.4184

Qaiser, S., and Ali, R. (2018). Text mining: Use of TF-IDF to examine the relevance of words to documents. *Int. J. Comput. Appl.* 181 (1), 25–29. doi:10.5120/ijca2018917395

Rommers, M. K., Teepe-Twiss, I. M., and Guchelaar, H-J. (2007). Preventing adverse drug events in hospital practice: An overview. *Pharmacoepidemiol. Drug Saf.* 16 (10), 1129–1135. doi:10.1002/pds.1440

Savova, G. K., Masanz, J. J., Ogren, P. V., Zheng, J., Sohn, S., Kipper-Schuler, K. C., et al. (2010). Mayo clinical text analysis and knowledge extraction system (cTAKES): Architecture, component evaluation and applications. *J. Am. Med. Inf. Assoc.* 17 (5), 507–513. doi:10.1136/jamia.2009.001560

Sherstinsky, A., 2018. Fundamentals of recurrent neural network (RNN) and long short-term memory (LSTM) network.

Skentzos, S., Shubina, M., Plutzky, J., and Turchin, A. (2011). "Structured vs. unstructured: factors affecting adverse drug reaction documentation in an EMR repository," in *AMIA annual symposium proceedings* (American Medical Informatics Association), 2011, 1270.

Stenetorp, P., Pyysalo, S., Topić, G., Ohta, T., Ananiadou, S., and Tsujii, J. I. (2012). "BRAT: a web-based tool for NLP-assisted text annotation," in *Proceedings of the Demonstrations at the 13th Conference of the European Chapter of the Association for Computational Linguistics*, 102–107.

Sun, C., Qiu, X., Xu, Y., and Huang, X. (2019). "How to fine-tune BERT for text classification?," in *Chinese computational linguistics. CCL 2019*. Editors M. Sun, X. Huang, H. Ji, and Z. Liu (Cham: Springer), 11856. doi:10.1007/978-3-030-32381-3_16

Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., et al. (2017). "Attention is all you need," in *Advances in neural information processing systems*, 6000–6010.

Wang, P., Bai, G. R., and Stolee, K. T., 2019. Exploring regular expression evolution. Proceedings of the 2019 IEEE 26th International Conference on Software Analysis, Evolution and Reengineering (SANER), pp.502–513.

Watanabe, J. H., McInnis, T., and Hirsch, J. D. (2018). Cost of prescription drug–related morbidity and mortality. *Ann. Pharmacother.* 52 (9), 829–837.

Wei, Q., Ji, Z., Li, Z., Du, J., Wang, J., Xu, J., et al. (2020). A study of deep learning approaches for medication and adverse drug event extraction from clinical text. *J. Am. Med. Inf. Assoc.* 27 (1), 13–21. doi:10.1093/jamia/ocz063

Wishart, D. S., Knox, C., Guo, A. C., Cheng, D., Shrivastava, S., Tzur, D., et al. (2008). DrugBank: A knowledgebase for drugs, drug actions and drug targets. *Nucleic Acids Res.* 36, D901–D906. doi:10.1093/nar/gkm958

Xu, G., Meng, Y., Qiu, X., Yu, Z., and Wu, X. (2019). Sentiment analysis of comment texts based on BiLSTM. *IEEE Access* 7, 51522–51532. doi:10.1109/ACCESS.2019.2909919

frontiers | Frontiers in Pharmacology

TYPE Review
PUBLISHED 13 July 2023
DOI 10.3389/fphar.2023.1197259

Check for updates

OPEN ACCESS

EDITED BY
Ramiro E. Gilardino,
Merck and Co., Inc., United States

REVIEWED BY
Nebojsa Pavlovic,
University of Novi Sad, Serbia
Jose Diaz,
Bristol Myers Squibb, United Kingdom

*CORRESPONDENCE
Jeffrey M. Muir,
jeffrey.muir@cytel.com

RECEIVED 30 March 2023
ACCEPTED 05 July 2023
PUBLISHED 13 July 2023

CITATION
Muir JM, Radhakrishnan A, Freitag A, Ozer Stillman I and Sarri G (2023), Reconstructing the value puzzle in health technology assessment: a pragmatic review to determine which modelling methods can account for additional value elements.
Front. Pharmacol. 14:1197259.
doi: 10.3389/fphar.2023.1197259

COPYRIGHT
© 2023 Muir, Radhakrishnan, Freitag, Ozer Stillman and Sarri. This is an open-access article distributed under the terms of the Creative Commons Attribution License (CC BY). The use, distribution or reproduction in other forums is permitted, provided the original author(s) and the copyright owner(s) are credited and that the original publication in this journal is cited, in accordance with accepted academic practice. No use, distribution or reproduction is permitted which does not comply with these terms.

Reconstructing the value puzzle in health technology assessment: a pragmatic review to determine which modelling methods can account for additional value elements

Jeffrey M. Muir[1]*, Amruta Radhakrishnan[1], Andreas Freitag[2], Ipek Ozer Stillman[3] and Grammati Sarri[2]

[1]Cytel Inc., Toronto, ON, Canada, [2]Cytel Inc., London, United Kingdom, [3]Takeda Pharmaceuticals, Lexington, MA, United States

Health technology assessment (HTA) has traditionally relied on cost-effectiveness analysis (CEA) as a cornerstone of evaluation of new therapies, assessing the clinical validity and utility, the efficacy, and the cost-effectiveness of new interventions. The current format of cost-effectiveness analysis, however, does not allow for inclusion of more holistic aspects of health and, therefore, value elements for new technologies such as the impact on patients and society beyond its pure clinical and economic value. This study aimed to review the recent modelling attempts to expand the traditional cost-effectiveness analysis approach by incorporating additional elements of value in health technology assessment. A pragmatic literature review was conducted for articles published between 2012 and 2022 reporting cost-effectiveness analysis including value aspects beyond the clinical and cost-effectiveness estimates; searches identified 13 articles that were eligible for inclusion. These expanded modelling approaches mainly focused on integrating the impact of societal values and health equity in cost-effectiveness analysis, both of which were championed as important aspects of health technology assessment that should be incorporated into future technology assessments. The reviewed cost-effectiveness analysis methods included modification of the current cost-effectiveness analysis methodology (distributional cost-effectiveness analysis, augmented cost-effectiveness analysis, extended cost-effectiveness analysis) or the use of multi-criteria decision analysis. Of these approaches, augmented cost-effectiveness analysis appears to have the most potential by expanding traditional aspects of value, as it uses techniques already familiar to health technology assessment agencies but also allows space for incorporation of qualitative aspects of a product's value. This review showcases that methods to unravel additional value elements for technology assessment exist, therefore, patient access to promising technologies can be improved by moving the discussion from "if" to "how" additional value elements can inform decision-making.

KEYWORDS

cost-effectiveness analysis, value elements, health technology assessment, health policy, societal value, health equity

Introduction

Value in health technology assessment (HTA), which is the foundation upon which decision-making regarding new drugs and health technologies is made in several healthcare systems, has been primarily based on balancing the clinical benefits to patients and/or economic costs involved by introducing the new technology to the healthcare system. Expansion of the concept of value in HTAs has been the subject of recent research and debate mainly driven by patients, carers and clinicians who recognize that the value of a new technology is multidimensional (Caro et al., 2019; Reed et al., 2019). This multidimensional nature is reflected in the latest definition of HTA provided by the Professional Society for Health Economics and Outcomes Research (ISPOR) Task Force which, in part, notes that dimension of value for a health technology may be assessed "by examining the intended and unintended consequences of using a health technology" and that this evaluation should encompass a comprehensive array of factors, including ethical, social, and cultural issues (O'Rourke et al., 2020a; O'Rourke et al., 2020b).

To this end, several organizations and research groups have developed value-based frameworks as an attempt to address the limitations of current HTA decision tools (Zhang et al., 2022). The ISPOR Strategic Task Force is leading an effort to reshape the future of HTA by examining the definition of a technology's value and encouraging the integration of additional elements of value not currently included in the technology submissions. The findings regarding new concepts of value have been summarized in the ISPOR Task Force's "Value Flower" (Lakdawalla et al., 2018). Some of the proposed elements beyond the traditional clinical and cost-effectiveness analyses include the value of: the reduction in uncertainty surrounding a disease, the fear of contagion, the value of insurance, the severity of disease, the value of hope, real option value, health equity, and scientific spillovers (Lakdawalla et al., 2018). Indeed, the Second Panel on Cost-effectiveness (Sanders et al., 2016) has recommended the incorporation of reference cases in each cost-effectiveness analysis (CEA) and an "impact inventory," i.e., a cataloguing of consequences of analysis decisions both inside and outside of the healthcare sector. Previous research has also shown that even though these expanded value-based frameworks (generic or disease specific) provide the possibility of incorporating additional benefits that technologies may bring to patients and society and contextual factors to be considered through deliberative processes, there are practical limitations for their implementation in routine HTA decision-making (Willke et al., 2018; Breslau et al., 2023). One of the main barriers for the wider implementation of these value-based frameworks in decision-making, especially when CEA is the pillar of HTA, is the lack of consensus on how reliably and consistently these elements can be applied across different disease indications and technologies (Willke et al., 2018; Reed et al., 2019). Additionally, the lack of consensus regarding methods to address these concerns, the concerns of double counting of outcomes or interdependent variables raised by this lack of consensus and the historically narrow remit of HTA agencies (i.e., costs and benefits are assessed from a healthcare systems or payer perspective) represent significant barriers to widespread adoption (Fornaro et al. 2021; Hendriks and Pearson, 2021; Garrison et al., 2020). As a result, little traction has been gained thus far for their wider implementation.

Traditionally, CEA evaluates the value of an intervention from a clinical and cost perspective, determining value as a trade-off between cost and health benefit (Canadian Institute for Health Information (CIHI), 2022; Guidelines for the economic, 2017). The structured nature of CEA contributes to its appeal, as it offers a quantitative and reproducible method of analysis standardized across different disease areas and technologies for decision-makers, who are concerned largely with extracting the maximum value for treatments provided for any given condition. Herein, however, lies one of the major drawbacks of the traditional approach to CEA: its restrictive nature fails to capture the additional elements of values that do not fall precisely within these standard, well-defined parameters (Willke et al., 2018; Garrison et al., 2019; Neumann et al., 2022). The quality-adjusted life year (QALY), which is considered by many in the HTA field as the cornerstone of traditional CEA and one of the two drivers of CEA results (along with survival), is seen by others as inherently flawed and based largely on sometimes unfounded assumptions; as such, this compromises its ability to provide a meaningful calculation of an intervention's value (Nord et al., 2009; Torbica et al., 2018; Caro et al., 2019; Rand and Kesselheim, 2021).

With this background in mind, this research aimed to identify if the additional elements of value previously described for consideration in HTAs have been proposed in modified economic modelling techniques or other deliberative approaches. In theory, various methodologies have been suggested to remedy the drawbacks of the current CEA approach. These methodologies range from slight alterations to QALY modifiers that consider additional elements of value without dramatically altering the current structure to completely new methodologies that attempt to maintain the objectivity of the CEA approach while incorporating expanded concepts of value (Asaria et al., 2016; Garrison et al., 2019). A commentary by Caro et al., 2019 (Caro et al., 2019) provided a critical summary of alternative approaches to QALYs that expand the measure of benefit/value of new technologies and help further deliberations on determining aspects of technology's value. To supplement the arguments noted in this commentary, and to continue the discussion on how the new HTA era should focus on creating an equitable, efficient, and high-quality health system (O'Rourke et al., 2020a), this review aimed to identify and describe the expanded economic analyses beyond the traditional CEA approach by incorporating additional elements of a technology's value in modelling approaches.

Materials and methods

A pragmatic literature review using reproducible criteria was conducted to capture relevant peer-reviewed articles. Reporting was guided by the Preferred Reporting Items for Systematic Review and Meta-Analyses statement (Page et al., 2021). The research question followed the Sample, Phenomenon of Interest, Design, Evaluation, Research (SPIDER) format (Library UoC, 2022): how have assessments of value (beyond clinical and cost estimates) for health technologies been incorporated in recent modelling approaches and deliberative processes? A structured database search for publicly available literature published in English from

TABLE 1 Summary of current traditional and expanded modelling approaches.

Modelling approach		Description
CEA	Traditional CEA Zamora et al., 2021	Uses the incremental cost-effectiveness ratio, which measures the costs incurred by the health system per quality adjusted life year gain when a new treatment or medical technology is used
Modified CEA	Augmented CEA Zamora et al., 2021	An extension of CEA that includes novel elements of value (e.g., insurance value, option value, and the value of hope) and considers trade-offs among them. This approach attaches a monetary value to all health gains
	Distributional CEA Cookson et al., 2017; Diaby et al., 2021	Focuses on the distributions of health effects (health gains/disease burden) associated with healthcare interventions at both population (societal) and subgroup (e.g., sex, race/ethnicity) levels as well as the distribution of health opportunity costs per equity-relevant sociodemographic variables and per disease categories. Decision making considers the trade-offs between improving total population health and reducing unfair health inequality.
	Extended CEA Cookson et al., 2017	Assesses the distribution of both health benefits and financial risk protection benefits and considers financial benefits of policies considering out-of-pocket payments in certain geographies
MCDA	Traditional MCDA Baltussen et al., 2017	Involves a structured and rational decision-making approach informed by evidence on multiple criteria that uses quantitative scores to choose, rank, select options
Modified MCDA	Equitable MCDA Diaby et al., 2021	Explicitly considers multiple criteria including the impact of treatment on health equality
	Qualitative MCDA DiStefano and Levin, (2019)	Incorporates qualitative considerations into MCDA by considering decision makers' opinions on the importance of each criterion while prioritizing interventions and/or subgroups, as opposed to solely relying on quantitative scores
	Reflective MCDA Goetghebeur and Cellier, (2018)	Focuses on compassionate care and assumes that decision makers reflect on the goals of the analysis and whether those goals align with a compassionate care approach while considering both quantitative and qualitative factors
	Advance value tree Angelis and Kanavos, (2017)	A modified MCDA that uses three criteria levels to measure value across five domains (burden of disease, therapeutic impact, safety profile, innovation level, and socioeconomic impact)
DCE	Traditional DCE Ngorsuraches, (2021)	Involves participants sequentially choose between hypothetical options to make decisions on choices of treatment or healthcare service based on attributes such as efficacy, side effects, and costs
Modified DCE	Latent class DCE Ngorsuraches, (2021)	Incorporates a latent class model to derive the value of equity
	Quantum choice DCE Ngorsuraches, (2021)	Incorporates equity attributes for individual alternatives in choice tasks to derive the value of equity

Abbreviations: CEA, cost-effectiveness analysis; DCE, discrete choice experiment; MCDA: multi-criteria decision analysis.

2012 to the present was conducted in Embase and MEDLINE on 24 March 2023 (see Appendix for complete search strategy). As the HTA process is rapidly evolving across many countries and "value" may be defined differently across cultures and healthcare systems, the review was not restricted by geography. Prior to commencing screening, a calibration exercise among reviewers was conducted on a random sample of 50 articles. Screening of titles/abstracts was carried out in the DistillerSR platform (Evidence Partners Incorporated; Ottawa, Canada) by a single reviewer with a second reviewer screening 15% of excluded articles as a quality check. The same approach was used for full-text screening. Eligible studies were required to meet all the following criteria: published following a peer-review process; discussed current HTA value frameworks in the context of CEA; provided new or expanded definitions of value; and discussed new modelling approach (es) to HTA. Studies that focused on disease-specific, value-based frameworks, solely on patient experiences, or strictly on economic modelling with no reference on how additional value elements can be incorporated were excluded. No grey literature sources or commentaries/editorials were considered for inclusion. Conference abstracts were excluded given the limited information provided.

Data extraction of included studies was carried out in a pre-specified template by a single reviewer and validated by a second reviewer. Data were extracted on publication characteristics, key themes, limitations in existing CEA approaches, and new recommendations for incorporating value within CEA. Each eligible article was evaluated based on three broad criteria: did the article comment on the suitability of the current HTA methodology; did the article discuss what aspects could or should be added to the current approach; and did the article discuss new methods for evaluating therapies? Included studies were categorized based on their recommendations for HTA agencies. The three main areas of methodology were: modifications to the current CEA approach (modified CEA: mCEA), which can include variations such as multi-criteria decision analysis (MCDA); and alternate approaches, such as discrete choice experiments (DCE). Within these frameworks, several sub-methodologies exist, such as distributional CEA (DCEA), augmented CEA (ACEA) and extended CEA (ECEA) within the mCEA framework; and different variations of current MCDA methods (Table 1).

Results

Study eligibility

The database searches returned 3,614 records and after removing duplicates, 2,528 unique records were screened at

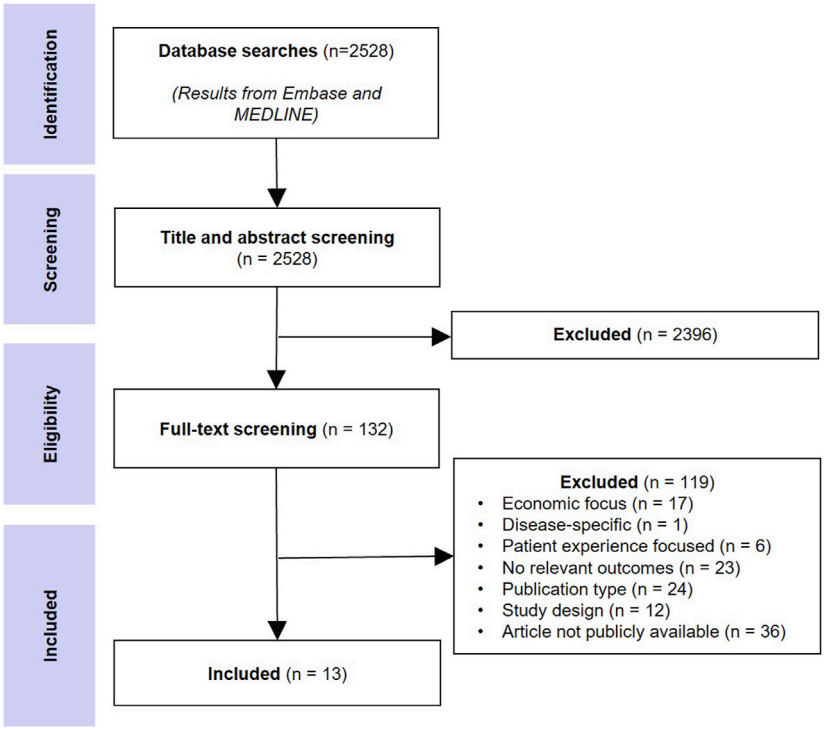

FIGURE 1
PRISMA diagram detailing literature search results and subsequent review process.

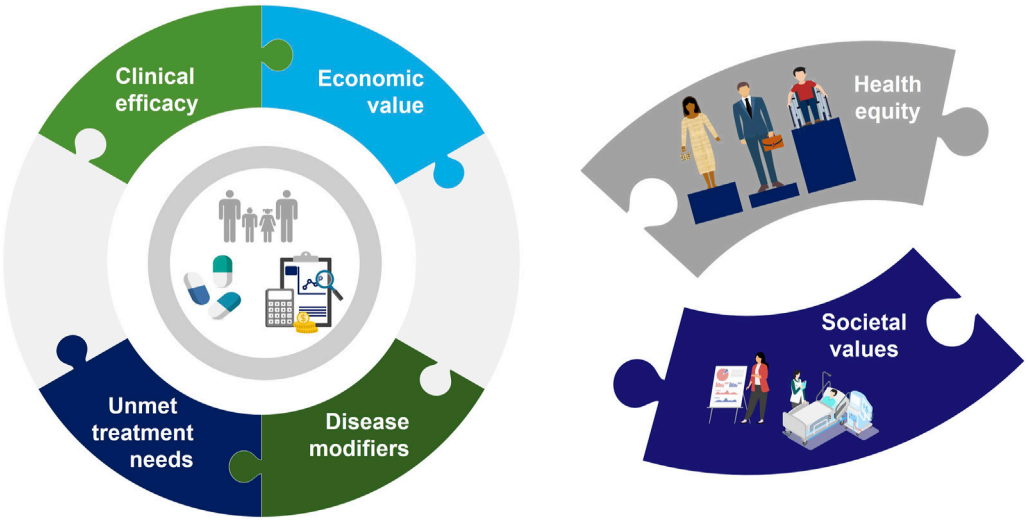

FIGURE 2
The "Value Puzzle" illustrates the existing aspects of CEA (clinical efficacy, economic value, disease modifiers and unmet treatment needs) but also highlights the missing aspects of the current system (health equity and societal values).

the title/abstract level, and 132 were selected for full-text screening. Thirteen peer-reviewed studies that provided recommendations on new approaches to HTA in the context of CEA were included. Figure 1 shows the literature selection procedure.

Description of included publications

Ten of the 13 included articles were literature review articles offering expert opinion while three were reports from multistakeholder workshops or committees. Six articles

specifically discussed at least one additional element to the current HTA (clinical and cost-effectiveness) value paradigm. Societal values and health equity were identified as the top two core pillars where the current CEA paradigm is wanting, with authors across these publications generally recommending an expansion of the definition of value within CEA to include these broad aspects (Figure 2).

Societal values were the most identified elements, mentioned in four articles (Dionne et al., 2015; Phelps and Madhavan, 2017; Pearson et al., 2019; Diaby et al., 2021). Societal aspects encompass a relatively broad spectrum of elements, but all authors agreed that the impact of disease on the patient is central to these societal considerations. For example, in the context of potentially curative treatments (Pearson et al., 2019), considerations of disease severity, its rarity, and the potential for a cure to extend life or limit the burden of illness (especially in children), as well as the value of hope and real option value offered by these treatments should be considered. The impact of productivity is also considered an important aspect to add to CEA (Dionne et al., 2015), as patients' contributions to society are directly related to their health and wellness. Above all, the perspectives of all relevant parties, labelled as the "5Ps" (patients, providers, payers, producers, and planners) are encouraged to be considered by decision-makers (Dionne et al., 2015; Phelps and Madhavan, 2017).

Health equity was identified in four articles (Dionne et al., 2015; Cookson et al., 2017; Goetghebeur and Cellier, 2018; Diaby et al., 2021) as an important factor that is largely lacking in the existing HTA frameworks. Equity is defined by the World Health Organization as: "the absence of unfair, avoidable or remediable differences among groups of people, whether those groups are defined socially, economically, demographically, or geographically or by other dimensions of inequality (e.g., sex, gender, ethnicity, disability, or sexual orientation)" (Organization WH, 2022; Sarri, 2022). Inequity is thus evident in circumstances where a deficit in one of these areas affects access to affordable care, which is limited in one or more marginalized groups. Cookson et al., 2017 (Cookson et al., 2017) made equity the core of their argument for new approaches in HTA analyses, focusing on the trade-offs required to ensure health equity and the net equity impact of HTA decisions. They argued that the tools for health equity analysis do exist (i.e., who gains and who loses in policy decisions) and that assessing the equity trade-offs should be incorporated into existing CEA methods. Similarly, Goetghebeur et al., 2018 (Goetghebeur and Cellier, 2018) framed equity as central to an approach based on the application of compassionate care concepts, where ethical considerations are contemplated by decision-makers to maximize equity and sustainability. Diaby et al., 2021 (Diaby et al., 2021) and Dionne et al., 2015 (Dionne et al., 2015) discussed equity from the patient's perspective, with patient demographics and a lack of patient heterogeneity in clinical studies mainly contributing to inequity in health assessment. The low representation of minority groups in clinical studies is suggested to under-represent the effect of therapies on these populations, thus contributing to decreased availability of treatments for these patients. Consideration of individual patient needs (i.e., patient preferences) and fairness in how health-economic decisions are made (i.e., balancing population and individual priorities while considering patient age, alternate treatments, and equity across different jurisdictions and populations) are additional dimensions of health equity domain (Dionne et al., 2015). In summary, researchers have long argued for societal values and equity considerations to be incorporated into existing HTA frameworks. In the context of societal values, it was argued that the impact of disease and its characteristics on patients and their productivity should also factor into decision-making. Similarly, patient preferences, addressing the needs of underrepresented groups, and ensuring access to affordable care are central to including equity considerations in HTA frameworks (Dionne et al., 2015; Diaby et al., 2021; Sarri, 2022).

Summary of mCEA or new modelling approaches

All eligible studies provided recommendations on new or modified approaches to HTA and CEA, which are mainly grouped as follows: mCEA; adoption of MCDA methods for CEA; and methods taking alternate approaches, such as DCE (Figure 3). The main theses and conclusions of the included peer-reviewed articles are summarized in Table 2.

mCEA

Four articles (Kristensen et al., 2017; Pearson et al., 2019; Diaby et al., 2021; Zamora et al., 2021) recommended mCEA as an expanded CEA method to incorporate additional value elements, albeit their suggestions differed considerably. Kristensen et al., 2017 (Kristensen et al., 2017) summarized the results of a decade-long analysis of HTA methods by the European Network for HTA (EUnetHTA), which recommends a tiered system that accounts for typical domains such as effectiveness, safety, and health economics but also includes domains addressing social, patient, legal, and organizational elements. EUnetHTA identified nine core elements that should be considered by HTA agencies and, as part of its tiered approach, delineated between a rapid relative effectiveness assessment (REA) for interventions requiring a short turnaround and a full, comprehensive assessment for all other interventions. The REA would focus on basic clinical elements (e.g., health problem identification, intervention description, safety, and clinical effectiveness), while the full assessment would add elements such as costs, ethical analysis, organization impact, patient/societal aspects, and legal aspects. The core elements are designed to be interconnected, such that the costs/economics domain can draw information from other domains (e.g., organizational or patients/social aspects) to expand the calculation of cost and affordability. This allows for a more comprehensive and nuanced analysis that better incorporates non-traditional elements.

Diaby et al., 2021 (Diaby et al., 2021) offered two recommendations on mCEA methods: the DCEA, and a two-part appraisal that augments incremental cost-effectiveness ratio (ICER) with a comprehensive benefits and value (CBV) score. DCEA refers to a blanket approach to technology assessment that models health distributions as a comparison of health gains vs. disease burden. This approach allows for analysis of health interventions at the population (i.e., societal) and demographic subgroup (e.g., sex, race) levels, enabling an analysis of health gains in the context of sociodemographic variables, which inevitably incorporates elements

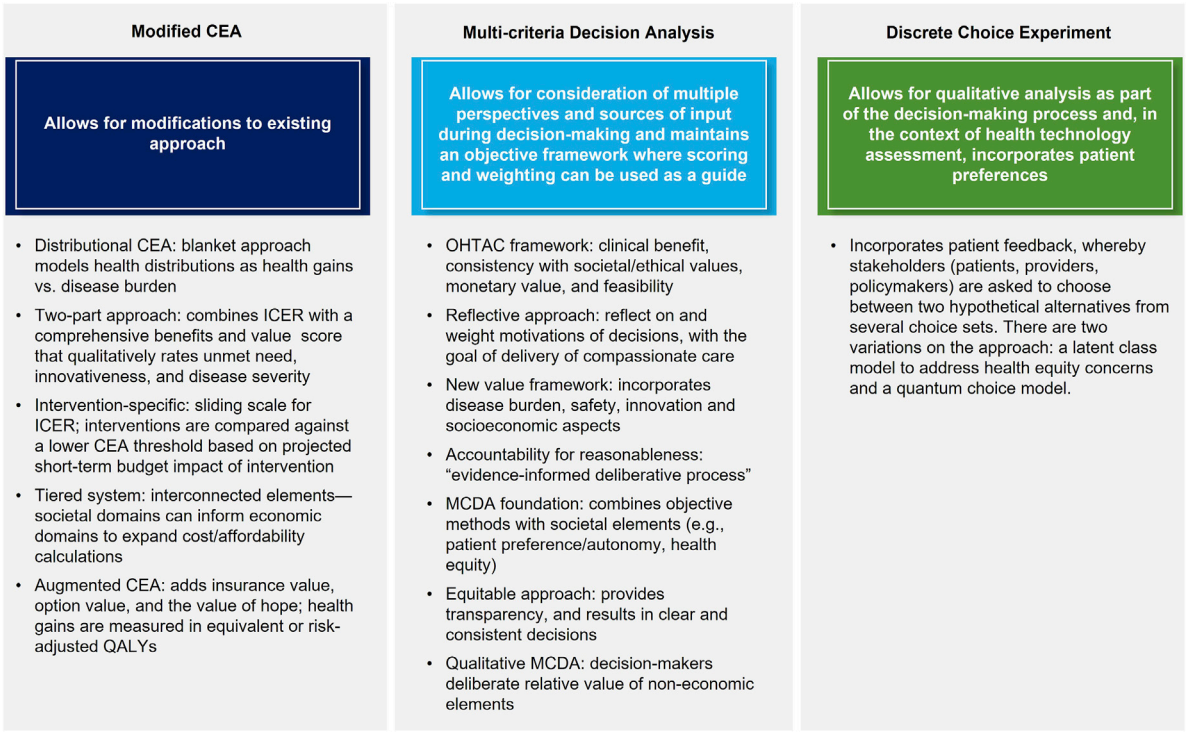

FIGURE 3
Overview of approaches recommended in the literature.

of equity as defined by these variables. Health gains and losses can thus be analyzed based on individual sociodemographic variables and/or by disease category. Under this proposed approach, decision-makers are asked to make trade-offs between decisions that would improve the overall health of the population and those that would reduce inequity in healthcare availability among specific patient subgroups. The two-part approach combines ICER threshold with a CBV score, allowing for a more robust analysis that considers quantitative and qualitative assessment factors. The CBV score is a composite, qualitative rating calculated using elements such as innovativeness, disease severity, and unmet need (Goldman et al., 2010; Diaby et al., 2021) which provides a more holistic assessment of the non-economic aspects of a given intervention.

Zamora et al., 2021 (Zamora et al., 2021) examined the potential of ACEA to incorporate additional individual value elements such as insurance value, option value, and the value of hope to the traditional ICER approach. Any health gains from new elements are measured in equivalent or risk-adjusted QALYs. A hierarchical approach is then used to calculate the final aggregate impact of an intervention, beginning with the incremental QALY and then incorporating QALY equivalents for new elements. Ultimately, final decisions on technologies' value are made through consideration of the trade-offs among the elements, as gains in one area may be associated with losses in another. The ability to quantify benefits/losses of an intervention in a common unit of measure (QALY) creates a single denominator in the calculation, which maintains an objective framework while incorporating elements that may traditionally be considered subjective. Additionally, the authors compared the ACEA and MCDA approaches and found them to be fairly similar, such that the choice between the two was largely pragmatic and thus their research question was left unresolved.

Finally, Pearson et al., 2019 (Pearson et al., 2019) discussed a more intervention-specific method that does not explicitly incorporate non-economic factors. They suggested several modifications, such as a sliding scale for the ICER, specifically for curative treatments, where interventions can be compared against a lower CEA threshold based on the projected short-term budget impact of the intervention. They further recommended adaptations including disallowing full credit for cost offsets for any interventions no longer required after a condition is cured, if that intervention itself was not cost effective; capping costs based on patient willingness to pay; and using shared savings, such that cost savings realized by curative treatments are shared between the innovator and the healthcare system. Fundamentally, their approach seeks to modify the calculations made during HTA but maintain the ability to objectively calculate costs and cost-savings, an approach not unlike that of MCDA in its desire to maintain a level of objectivity in decision-making.

MCDA

Nine articles (Dionne et al., 2015; Angelis and Kanavos, 2017; Baltussen et al., 2017; Phelps and Madhavan, 2017; Goetghebeur and Cellier, 2018; Krahn et al., 2018; DiStefano and Levin, 2019; Diaby et al., 2021; Zamora et al., 2021) recommended the adoption of some version of MCDA as a method for future CEA. MCDA methods

TABLE 2 Summary of eligible articles.

Author, publication date	Modelling approaches	Targeted value elements	Themes noted
Diaby et al., 2021	MCDA, mCEA (distributional), CBV score	Health equity	The authors examine the current landscape regarding value frameworks in HTA, with a focus on the impact of current methodologies on health equity. They highlight the lack of diversity among most patient populations in RCTs and the fact that current frameworks largely ignore patient heterogeneity, complex demographic factors, and access to care. Three approaches are proffered to address these shortcomings: a 2-part HTA incorporating traditional HTA methods with a CBV score; a distributional CEA method; and equitable MCDA methods.
Ngorsuraches, (2021)	DCE (latent, quantum)	Health equity	Drawbacks of the current HTA methodologies are discussed and two new methods of assessment, based on DCEs are provided. One model utilizes a latent class model to address health equity in value assessment; the other uses a quantum choice model.
Zamora et al., 2021	mCEA (augmented), MCDA	Societal values (value of hope, value of insurance)	A comparison between a modified CEA approach (augmented CEA, ACEA) and an MCDA approach to HTA decisions is explored including an examination of the trade-offs between financial loss and healthcare gain. A context of insurer coverage for healthcare innovations, i.e., new medical technologies, is used.
DiStefano and Levin, (2019)	MCDA (qualitative)	Health equity	The authors discuss current drug prescribing guidelines and how the addition of CEA concepts in the guidelines may help promote improved equity in health. They discuss several concepts within traditional CEA methods, including MCDA approaches, arguing that a *qualitative* MCDA approach may be preferred. A qualitative approach forgoes the aggregation of scores and allows decisions to include deliberations amongst decision-makers.
Pearson et al., 2019	mCEA	Societal values (real option value, value of hope, value of insurance, value of potential cure)	The authors outline drawbacks of current "utilitarian" HTA frameworks, including the inability of current methods to account for social values such as disease severity and rarity, burden of illness and the ability of curative treatments to extend life, especially that of children. They suggest modifications to the current CEA approach, such as adopting a "sliding scale" for the ICER and capping drug prices based on willingness-to-pay metrics.
Krahn et al., 2018	MCDA	Social values (quality, evidence, effectiveness, equity, population health, collaboration)	A summary of the OHTAC framework is presented. An audit of the existing HTA methodologies is presented and recommendations for future assessments are made, including a focus on 4 key attributes: overall clinical benefit; consistency with expected societal and ethical values; value for money; and feasibility of adoption into the healthcare system.
Goetghebeur and Cellier, (2018)	MCDA (reflective)	Compassionate care	A new approach to CEA based on MCDA methods is discussed. Central to the approach is the concept of compassionate care, the underlying impetus of healthcare. The method involves analysis of quantitative (e.g., disease severity) and qualitative (e.g., health system capacity) factors but also allows for an opportunity for reflection on the goals of the analysis and whether those goals align with a compassionate care approach.
Angelis and Kanavos, (2017)	MCDA	Disease burden, therapeutic value, safety, innovation, socioeconomic value	The authors discuss the current CEA approaches, identify drawbacks and present a new approach that includes evaluation of the burden of disease, the level of innovation of the intervention, its ease of use and its socioeconomic value. A decision-tree model is utilized to provide guidance in decision-making.
Kristensen et al., 2017	mCEA	Organizational aspects, ethical aspects, patient/social aspects	This study presents a new approach for CEA based on a 10-year effort by the European HTA agencies. It discusses the application of a model designed for both full and rapid assessments. Nine domains are outlined, all of which are part of the full assessment but only four of which are part of the rapid assessment.
Baltussen et al., 2017	MCDA	Ethical issues	The authors summarize the current landscape of CEA and presents a new approach, utilizing MCDA methods. The new method importantly includes stakeholder deliberation to

(Continued on following page)

TABLE 2 (*Continued*) Summary of eligible articles.

Author, publication date	Modelling approaches	Targeted value elements	Themes noted
			facilitate learning and combines MCDA methods with accountability for reasonableness techniques into a new approach dubbed an "evidence-informed deliberative process." The responsibilities of the various HTA parties in implementing the new approach are discussed in detail.
Phelps and Madhavan, (2017)	MCDA	Patient-centered variables	A summary of the shortcomings of the current CEA methods is presented, highlighting several areas where current methods are lacking. A new approach is recommended, based on MCDA methods, which provides scores for aspects of decisions for oncology patients such as the likelihood of hair loss or nausea with certain treatments. The authors focus on the concept of perspective and highlight the importance of perspective from multiple perspectives, including that of the patient, the provider, the payer, the producer (manufacturer) and the planner (the "5Ps").
Cookson et al., 2017	mCEA (distributional, extended)	Health equity	The authors focus on the costs and benefits of CEA in the context of health equity and equitable access to treatment. A new approach to CEA using extended CEA or distributional CEA as the preferred methods of analysis is recommended. As part of this approach, the roles of equity impact analysis and equity trade-off analysis in decision-making are explored.
Dionne et al., 2015	MCDA	Societal benefit, health equity, patient autonomy, innovation	The authors identify several areas where the current CEA methodology is lacking and discuss additional aspects that should be included in future, including factors that address societal values. Several factors are deemed important when considering new aspects, including comparative effectiveness, adoption feasibility, patient autonomy, societal benefit, equity, innovation and disease prevention. A new methodology using MCDA techniques is recommended and discussed in the context of rare diseases and end-of-life decisions.

Abbreviations: ACEA, augmented cost-effectiveness analysis; CBV, comprehensive benefits and value; CEA, cost-effectiveness analysis; DCE, discrete choice experiment; HTA, health technology assessment; ICER, incremental cost-effectiveness ratio; MCDA, multi-criteria decision analysis; mCEA, modified cost-effectiveness analysis; OHTAC, ontario health technology advisory committee; RCT, randomized controlled trial.

allow for consideration of multiple perspectives and sources of input during decision-making with the aim to maintain an objective framework where scoring and weighting can be used to guide the process (Koksalan et al., 2011). Krahn et al., 2022 (Krahn et al., 2018), summarized the findings of the Ontario Health Technology Advisory Committee which recommend a framework that includes four key attributes: consideration of overall clinical benefit, consistency with societal/ethical values, value for money, and feasibility. It is within this framework that they advocate for the use of MCDA methods, citing the Evidence and Values Impact on DEcision Making framework (EVIDEM) (Goetghebeur et al., 2008) as an approach that is being increasingly explored and should be considered in future HTAs. Goetghebeur (Goetghebeur and Cellier, 2018) extended this work to suggest the use of a reflective MCDA approach, where decision-makers can reflect on and weight the motivations of their decisions, bearing in mind that decisions should be made in a patient-centric manner, with an eye toward the delivery of compassionate care. Angelis et al., 2017 (Angelis and Kanavos, 2017) outlined a new value framework using MCDA methods as a foundation, which incorporates several key aspects such as burden of disease, therapeutic considerations, safety, innovation, and socioeconomic considerations. Their resulting decision-tree approach considers each of these elements, with subsequent downstream decisions made based on each one; the final decision is based ultimately on the cumulative impact of each element and decision. Baltussen et al., 2017 (Baltussen et al., 2017) combined MCDA methods with accountability for reasonableness concepts as part of a new approach that they refer to as an "evidence-informed deliberative process." They categorized the traditional elements assessed by HTA agencies (e.g., safety, effectiveness, budget impact) as "general criteria" and advocated for the additional consideration of "contextual criteria," which encompass more patient-centric or societal considerations. They recommended consultation with the public on what contextual elements may be important during an HTA. As such, a combination of quantifiable criteria and non-quantifiable (i.e., qualitative) criteria should be considered, with both ultimately being used as inputs into the deliberative process. Phelps, et al., 2017 (Phelps and Madhavan, 2017), Dionne et al., 2015 (Dionne et al., 2015), and Zamora et al., 2021 (Zamora et al., 2021) also advocated for the use of MCDA methods as a way to maintain objectivity in decision-making, while still taking into account societal elements such as patient preference/autonomy and health equity. Diaby et al., 2021 (Diaby et al., 2021) similarly suggested using MCDA methods as part of an "equitable MCDA" approach, one that is transparent and results in clear and consistent decisions. They suggested the importance of both the consideration of multiple criteria as well as the impact of a given treatment on health equality. Finally, DiStefano et al., 2019 (DiStefano and Levin, 2019) stressed the importance of qualitative MCDA (Baltussen et al., 2019) which, by forgoing

aggregation of scores, allows decisions to include deliberation among decision-makers regarding the relative value of non-economic elements, thus maintaining transparency and allowing for more subjective incorporation of elements such as equity, rather than integration of those elements into more traditional or mCEA. This aims to maintain the objective nature of MCDA while allowing for subjective consideration in the decision-making process.

DCE

One included publication specifically looked at the use of the DCE model for future HTA decisions (Ngorsuraches, 2021). The authors suggested that DCE allows for a qualitative analysis as part of the decision-making process and, in the context of HTA, incorporates patient preferences while all stakeholders involved (patients, providers, policymakers) are asked to choose between two hypothetical alternatives from a number of choice sets. As such, the prevalence and importance of equity to each stakeholder is determined in the initial stage, after which those equity elements can be incorporated into a choice model which, when applied, can be used to establish the value of equity. Two variations on the DCE approach include one which utilizes a latent class model to address concerns about health equity in value assessment, and another which utilizes a quantum choice model. The author did not express a preference for one method over another but noted that the use of either methods would address the inadequacies of current methodologies and help address health disparities and underrepresented patient populations.

Discussion

In most countries, HTA remains anchored by CEAs, the cornerstone analysis when considering reimbursement of new therapies. The objective nature of the traditional CEA is seen as a benefit that lends itself to impartial decision-making although this method did not entirely prevent discrepancies in decision-making; however, there is increasing sentiment that the objective approach in fact marginalizes the subjective aspects of the healthcare assessment. The definition of "value" is a prime example of the drawbacks of the current system, as there is a growing opinion that value in HTA should be viewed through more than simply an economic lens. Health gains are not straightforwardly assessed, and several approaches have been proposed to define additional elements of value beyond the clinical and cost gains. Lately, there is an increased discontent with the inability of ICERs and QALYs to sufficiently capture the benefits valued by patients and societies overall when a new health technology is introduced (Caro et al., 2019). Although mainly the discussion so far has focused on defining these additional value elements, little effort has been put on demonstrating how these additional considerations can be implemented in modelling approaches to be used in the HTA context. To address this gap, the current study examined recent modelling approaches that included expanded or new definitions of value; two main analytical approaches were identified and advocated by most authors: a modification of the current CEA and the use of advanced decision-making techniques such as MCDA, both of which have merit. To date, however, no preferred method has been established for HTA adoption although several concerns have been raised regarding the implementation of MCDA as part of HTA decision-making (Marsh et al., 2018).

Despite these efforts, consensus on the most efficient and appropriate way to incorporate expanded definitions of value (and added benefit value frameworks) into current HTA in general and CEA methods in particular has remained elusive. One consensus finding from this review was that the current approach to CEA is lacking and that there are several elements–especially relating to the current narrow definition of value in CEA–that should be added to CEA methods going forward. These aspects represent missing pieces of the "Value Puzzle" (Figure 2) and illustrate the challenges assessors face in integrating these new factors into their decision-making. These factors have been identified by several groups, including the ISPOR Task Force, which summarized these concepts in the "Value Flower" (Lakdawalla et al., 2018; Willke et al., 2018; Garrison et al., 2019; Neumann et al., 2022). Generally, these missing aspects center around an expanded definition of value, one which includes more qualitative elements such as the ability of a treatment to provide hope to the patient (the value of hope), the value associated with extending life and opening possibilities for future treatments (real option value), the value of scientific discoveries and their wide applicability (scientific spillover), and more. These new elements reflect the prevailing opinion that the current approach, with its focus on cost-effectiveness and the use of metrics such as the ICER, is inadequate. However, these additional value elements may not transport at the same degree across disease areas and populations (Shafrin et al., 2021). Indeed, many are of the opinion that the central metric in these calculations–the QALY–is an inherently flawed metric built upon assumptions (Hall, 2020) which marginalizes the sickest in a population by presenting only an aggregate calculation of health (Caro et al., 2019). Indeed, many jurisdictions have begun to move away from the QALY, which has been outright rejected in Germany and Spain (Institute for Quality and Efficiency in Healthcare IQWiG, 2022) and remains largely unused in the United States (Neumann and Weinstein, 2010), France (Rumeau-Pichon and Harousseau, 2014), and some Latin American countries (Brixner et al., 2017). It is thus important to recognize the limitations of the QALY as a final, lone decision metric and that its use represents a first albeit limited step in the process of assessing value in pharmaceutical innovation. Clearly, the lack of enthusiasm for traditional CEA methods among HTA bodies and the feedback from patients necessitates a new approach to decision-making.

The current study identified two main themes recommended to address the shortcomings of the current system: adoption of either MCDA methods or modification of the current CEA approach. The former was advocated for in the majority of articles included in the review and has been widely discussed in the CEA space as a viable alternative for some time; however, it has failed to gain traction, at least in part due to its overly mechanistic nature (Kennedy, 2009; Baltussen et al., 2019) and tendency to ignore opportunity costs (Campillo-Artero et al., 2018; Marsh et al., 2018; Baltussen et al., 2019). Quite the reverse, the use of mCEA methods has been suggested as a viable avenue for change that simultaneously addresses concerns raised by Caro et al., 2019 (Caro et al., 2019), who suggested the current CEA methods continue to be utilized by HTA bodies mainly out of convenience, and due to the lack of a viable, proven alternative. Thus, one of the draws to modifying

current methods is the fact that it does not stray far from the *status quo*. With infrastructure in place and decades of published decisions, a major change in methodology may not be palatable for key stakeholders. Some of the recommendations in the current review slightly revised the current approach but did not recommend major changes (Kristensen et al., 2017; Pearson et al., 2019; Diaby et al., 2021). As such, these recommendations perhaps do not do enough to address current concerns. Garrison et al., 2017 (Garrison et al., 2017) proposed the use of ECEA methods to incorporate the value of "knowing" into CEA, which broadly incorporates several elements identified as valuable in our review, including reducing uncertainty and incorporating insurance value, real option value and scientific spillover into CEA. Another approach that holds promise is the ACEA approach (Zamora et al., 2021) which, like ECEA, combines the known methodology while still incorporating robust definitions of value including the value of hope, real option value, and insurance value. Indeed, the summary of the ISPOR Special Task Force report (Garrison et al., 2020) advocated for the use of ACEA methods as a way to combine the known (and widely accepted) clinico-economic aspects of traditional CEA with a comprehensive list of qualitative elements reflecting the various definitions of value. That recommended method would allow for the consideration of additional value elements (insurance, disease severity, hope, and real option value) while also allowing concepts such as equity and the benefit of scientific spillover from new technologies to be incorporated into deliberations. These additional value elements can be part of the technologies scoping exercises and tailored to the patients' preferences. While more research is needed to refine the methods, this approach shows promise and may best address the documented shortcomings of the current approach.

Beyond methodology, HTA agencies face many other challenges in their efforts to fairly evaluate new therapies. As environmental awareness and concern grows worldwide, HTA agencies will be required to include an evaluation of the impact of a health technology's production, use, and disposal. Toolan et al. (Toolan et al., 2023) have recently summarized the challenges associated with this effort and identified several approaches that HTA agencies may adopt during their assessment, including republishing of data in the public domain, considering environmental data in parallel with health economic data, integrating environmental data into existing methodologies, or specific evaluation of technologies that may have minimal health benefits but claim environmental benefits with their use. From a more patient-centric perspective, patients' perspectives and preferences have been suggested as important factors that warrant more attention in the HTA process. Several authors have referred to "The 5Ps" as important contextual considerations in HTA, namely, that the perspectives of many stakeholders–patients, practitioners, payers, producers and policymakers–must be part of any CEA (Phelps and Madhavan, 2017; Slejko et al., 2019; Hall, 2020; van Overbeeke et al., 2021). Incorporating patient preference and experience, and their perception of the quality of life amidst their illness, offsets the objective nature of the traditional CEA methods and theoretically allows for a more comprehensive assessment (Sarri et al., 2021). For example, factors important to the patient regarding the impact their diagnosis will have on those around them (Vrinzen et al., 2022) or life satisfaction should be considered in any assessment (Hall, 2020). Furthermore, a patient's preference can be reflected in their willingness to pay for or undergo treatment based on whether that treatment can offer them hope for recovery (Peasgood et al., 2022). Several authors have noted that patients are more willing to pay for a "hopeful therapy", with patients with cancer identified as those who prefer a therapy that has the possibility of a large therapeutic gain, even when the average response to that therapy may be similar to other options (Lakdawalla et al., 2012). As Hall comments: a patient who adapts to illness may live longer but may be less able to fight off future illness. Do the patient's values change as they adapt to their disease? And how does the QALY account for this adaptation (Hall, 2020)? Administratively, the financial burden placed on healthcare systems will only continue to increase. Healthcare systems stretched thin by the recent COVID-19 pandemic face ongoing challenges in integrating costs for new therapies into an already strained system (Epstein et al., 2020; Information CIfH, 2022; Youn et al., 2022). However, recent trends such as the growing use of real-world evidence (RWE) in healthcare research in general and with it a concomitant uptake in the use of RWE in regulatory and HTA agency filings may provide the opportunity to unravel existing health inequalities that directly fit in the decision-making (Sarri, 2022). However, the potential of RWE to capture the direction and magnitude of impact a new health technology may have on health inequalities has not been fully explored (Goetghebeur and Cellier, 2018). Proposed checklists to guide HTA decision-makers include equity considerations in their assessment may help on this front (Benkhalti et al., 2021). The struggle for HTA staff to keep pace with evolving RWE methodological complexities adds to the challenges facing these agencies. This is especially true in cases of rare disease or where ethical concerns prevent the designing of placebo-controlled, two-armed studies (Thorlund et al., 2020; Ramagopalan et al., 2021; Popat et al., 2022). All told, the challenges facing HTA bodies are layered and complex. More case studies are needed to demonstrate how reliably these holistic value aspects can be integrated into HTA, although buy-in among assessors and researchers is also required, to facilitate the widespread use of new and expanded methodologies and the learnings from demonstration of case studies.

This study should be considered with the following limitations. The pragmatic nature of the search, while comprehensive, could have missed some relevant articles, although the broad nature of the search may mitigate this concern. Related is the decision to include only peer-reviewed articles in data/theme collection. Commentaries and/or editorials were excluded from this review, which may result in some valid recommendations regarding these modelling techniques being missed. However, any commentaries that were captured in the search were reviewed for relevant opinions and referenced in the discussion as appropriate. Finally, articles that focused on a specific disease were excluded, as the aim was to provide a broad overview of these modelling techniques. This may also result in missing some articles that may have offered valuable perspectives on this topic; however, the wider focus of the review may make the findings more broadly applicable and initiate some methodological discussion.

Conclusion

This research demonstrated that modelling methods are being expanded from the traditional CEA approach to

incorporate value elements with a more holistic view of what matters most to patients and society. Although the methods differ, a consensus exists on the need for HTA agencies to redefine "value" with a wider lens that looks at more than just the clinical and economic benefits of a new technology. Societal factors and health equity scored highly as additional value elements. Future efforts are needed to increase the confidence of stakeholders in the importance of "testing" these expanded CEAs approaches in case studies.

Author contributions

JM drafted the manuscript and all authors reviewed, contributed to revisions, and approved the final version of the manuscript. All authors contributed to the article and approved the submitted version.

Funding

Financial support for this study was provided entirely by Takeda Pharmaceuticals, Inc. The funding agreement ensured the authors' independence in designing the study, interpreting the data, writing, and publishing the report.

Acknowledgments

The authors wish to acknowledge the editorial support by Colleen Dumont.

Conflict of interest

Authors JM, GS, AR, and AF were employed by Cytel, Inc. Author IO was employed by Takeda Pharmaceuticals, Inc.

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

Angelis, A., and Kanavos, P. (2017). Multiple criteria decision analysis (MCDA) for evaluating new medicines in health technology assessment and beyond: The advance value framework. *Soc. Sci. Med.* 188, 137–156. doi:10.1016/j.socscimed.2017.06.024

Asaria, M., Griffin, S., and Cookson, R. (2016). Distributional cost-effectiveness analysis: A tutorial. *Med. Decis. Mak.* 36 (1), 8–19. doi:10.1177/0272989X15583266

Baltussen, R., Jansen, M. P. M., Bijlmakers, L., Grutters, J., Kluytmans, A., Reuzel, R. P., et al. (2017). Value assessment frameworks for HTA agencies: The organization of evidence-informed deliberative processes. *Value Health* 20 (2), 256–260. doi:10.1016/j.jval.2016.11.019

Baltussen, R., Marsh, K., Thokala, P., Diaby, V., Castro, H., Cleemput, I., et al. (2019). Multicriteria decision analysis to support health technology assessment agencies: Benefits, limitations, and the way forward. *Value Health* 22 (11), 1283–1288. doi:10.1016/j.jval.2019.06.014

Benkhalti, M., Espinoza, M., Cookson, R., Welch, V., Tugwell, P., and Dagenais, P. (2021). Development of a checklist to guide equity considerations in health technology assessment. *Int. J. Technol. Assess. Health Care* 37, e17. doi:10.1017/S0266462320002275

Breslau, R. M., Cohen, J. T., Diaz, J., Malcolm, B., and Neumann, P. J. (2023). A review of HTA guidelines on societal and novel value elements. *Int. J. Technol. Assess. Health Care* 39 (1), e31. doi:10.1017/S026646232300017X

Brixner, D., Maniadakis, N., Kaló, Z., Hu, S., Shen, J., and Wijaya, K. (2017). Applying multi-criteria decision analysis (MCDA) simple scoring as an evidence-based HTA methodology for evaluating off-patent pharmaceuticals (OPPs) in emerging markets. *Value Health Reg. Issues* 13, 1–6. doi:10.1016/j.vhri.2017.02.001

Campillo-Artero, C., Puig-Junoy, J., and Culyer, A. J. (2018). Does MCDA trump CEA? *Appl. Health Econ. Health Policy* 16 (2), 147–151. doi:10.1007/s40258-018-0373-y

Canadian Institute for Health Information (CIHI) (2022). *COVID-19's Impact on Health Care Systems* (Accessed November 18, 2022).

Caro, J. J., Brazier, J. E., Karnon, J., Kolominsky-Rabas, P., McGuire, A. J., Nord, E., et al. (2019). Determining value in health technology assessment: Stay the course or tack away? *Pharmacoeconomics* 37 (3), 293–299. doi:10.1007/s40273-018-0742-2

Cookson, R., Mirelman, A. J., Griffin, S., Asaria, M., Dawkins, B., Norheim, O. F., et al. (2017). Using cost-effectiveness analysis to address health equity concerns. *Value Health* 20 (2), 206–212. doi:10.1016/j.jval.2016.11.027

Diaby, V., Ali, A., Babcock, A., Fuhr, J., and Braithwaite, D. (2021). Incorporating health equity into value assessment: Frameworks, promising alternatives, and future directions. *J. Manag. Care Spec. Pharm.* 27 (9-a Suppl. l), S22–S29. doi:10.18553/jmcp.2021.27.9-a.s22

Dionne, F., Mitton, C., Dempster, B., and Lynd, L. D. (2015). Developing a multi-criteria approach for drug reimbursement decision making: An initial step forward. *J. Popul. Ther. Clin. Pharmacol.* 22 (1), e68–e77.

DiStefano, M. J., and Levin, J. S. (2019). Does incorporating cost-effectiveness analysis into prescribing decisions promote drug access equity? *AMA J. Ethics* 21 (8), E679–E685. doi:10.1001/amajethics.2019.679

Epstein, R. H., Dexter, F., Smaka, T. J., and Candiotti, K. A. (2020). Policy implications for the COVID-19 pandemic in light of most patients (≥72%) spending at most one night at the hospital after elective, major therapeutic procedures. *Cureus* 12 (8), e9746. doi:10.7759/cureus.9746

Fornaro, G., Federici, C., Rognoni, C., and Ciani, O. (2021). Broadening the concept of value: A scoping review on the option value of medical technologies. *Value Health* 24 (7), 1045–1058. doi:10.1016/j.jval.2020.12.018

Garrison, L. P., Jr., Kamal-Bahl, S., and Towse, A. (2017). Toward a broader concept of value: Identifying and defining elements for an expanded cost-effectiveness analysis. *Value Health* 20 (2), 213–216. doi:10.1016/j.jval.2016.12.005

Garrison, L. P., Jr., Neumann, P. J., and Willke, R. J. (2019). Reflections on the ISPOR special Task Force on U.S. Value frameworks: Implications of a health economics approach for managed care pharmacy. *J. Manag. Care Spec. Pharm.* 25 (11), 1185–1192. doi:10.18553/jmcp.2019.25.11.1185

Garrison, L. P., Jr., Zamora, B., Li, M., and Towse, A. (2020). Augmenting cost-effectiveness analysis for uncertainty: The implications for value assessment-rationale and empirical support. *J. Manag. Care Spec. Pharm.* 26 (4), 400–406. doi:10.18553/jmcp.2020.26.4.400

Goetghebeur, M. M., and Cellier, M. S. (2018). Can reflective multicriteria be the new paradigm for healthcare decision-making? The EVIDEM journey. *Cost Eff. Resour. Allocation* 16 (Suppl. 1), 54. doi:10.1186/s12962-018-0116-9

Goetghebeur, M. M., Wagner, M., Khoury, H., Levitt, R. J., Erickson, L. J., and Rindress, D. (2008). Evidence and Value: Impact on DEcisionMaking-the EVIDEM framework and potential applications. *BMC Health Serv. Res.* 8, 270. doi:10.1186/1472-6963-8-270

Goldman, D., Lakdawalla, D., Philipson, T. J., and Yin, W. (2010). Valuing health technologies at NICE: Recommendations for improved incorporation of treatment value in HTA. *Health Econ.* 19 (10), 1109–1116. doi:10.1002/hec.1654

Guidelines for the economic (2017). *Guidelines for the economic evaluation of health technologies.*

Hall, A. (2020). Quality of life and value assessment in health care. *Health Care Anal.* 28 (1), 45–61. doi:10.1007/s10728-019-00382-w

Hendriks, S., and Pearson, S. D. (2021). Assessing potential cures: Are there distinctive elements of value beyond health gain? *J. Comp. Eff. Res.* 10 (4), 255–265. doi:10.2217/cer-2020-0190

Information CIfH (2022). *Overview: COVID-19's impact on health care systems.*

Institute for Quality and Efficiency in Healthcare (IQWiG) (2022). Methods for assessment of the relation of benefits to costs in the German statutory healthcare system. Available at: http://www.iqwig.de/download/08-10-14_Methods_of_the_Relation_of_Benefits_to_Costs_v_1_1.pdf.

Kennedy, I. (2009). *Appraising the value of innovation and other benefits*, 22. New York: A short study for NICE.

Koksalan, M., Wallenius, J., and Zionts, S. (2011). *Multiple criteria decision making: From early history to the 21st century*. Singapore: World Scientific.

Krahn, M., Miller, F., Bayoumi, A., Brooker, A. S., Wagner, F., Winsor, S., et al. (2018). Development of the ontario decision framework: A values based framework for health technology assessment. *Int. J. Technol. Assess. Health Care* 34 (3), 290–299. doi:10.1017/S0266462318000235

Kristensen, F. B., Lampe, K., Wild, C., Cerbo, M., Goettsch, W., and Becla, L. (2017). The HTA core Model(®)-10 Years of developing an international framework to share multidimensional value assessment. *Value Health* 20 (2), 244–250. doi:10.1016/j.jval.2016.12.010

Lakdawalla, D. N., Doshi, J. A., Garrison, L. P., Jr., Phelps, C. E., Basu, A., and Danzon, P. M. (2018). Defining elements of value in health care-A health economics approach: An ISPOR special Task Force report [3]. *Value Health* 21 (2), 131–139. doi:10.1016/j.jval.2017.12.007

Lakdawalla, D. N., Romley, J. A., Sanchez, Y., Maclean, J. R., Penrod, J. R., and Philipson, T. (2012). How cancer patients value hope and the implications for cost-effectiveness assessments of high-cost cancer therapies. *Health Aff. (Millwood)* 31 (4), 676–682. doi:10.1377/hlthaff.2011.1300

Library UoC (2022). Literature reviews for the health sciences. Available at: https://canberra.libguides.com/c.php?g=940615&p=6808850.

Marsh, K. D., Sculpher, M., Caro, J. J., and Tervonen, T. (2018). The use of MCDA in HTA: Great potential, but more effort needed. *Value Health* 21 (4), 394–397. doi:10.1016/j.jval.2017.10.001

Neumann, P. J., Garrison, L. P., and Willke, R. J. (2022). The history and future of the "ISPOR value flower": Addressing limitations of conventional cost-effectiveness analysis. *Value Health* 25 (4), 558–565. doi:10.1016/j.jval.2022.01.010

Neumann, P. J., and Weinstein, M. C. (2010). Legislating against use of cost-effectiveness information. *N. Engl. J. Med.* 363 (16), 1495–1497. doi:10.1056/NEJMp1007168

Ngorsuraches, S. (2021). Using latent class and quantum models to value equity in health care: A tale of 2 stories. *J. Manag. Care Spec. Pharm.* 27 (9-a Suppl. l), S12–S16. doi:10.18553/jmcp.2021.27.9-a.s12

Nord, E., Daniels, N., and Kamlet, M. (2009). QALYs: Some challenges. *Value Health* 12 (Suppl. 1), S10–S15. doi:10.1111/j.1524-4733.2009.00516.x

O'Rourke, B., Oortwijn, W., and Schuller, T. (2020a). Announcing the new definition of health technology assessment. *Value Health* 23 (6), 824–825. doi:10.1016/j.jval.2020.05.001

O'Rourke, B., Oortwijn, W., and Schuller, T., and International Joint Task Group (2020b). The new definition of health technology assessment: A milestone in international collaboration. *Int. J. Technol. Assess. Health Care* 36 (3), 187–190. doi:10.1017/S0266462320000215

Organization WH (2022). Health equity. Available at: https://www.who.int/health-topics/health-equity#tab=tab_1.

Page, M. J., McKenzie, J. E., Bossuyt, P. M., Boutron, I., Hoffmann, T. C., Mulrow, C. D., et al. (2021). The PRISMA 2020 statement: An updated guideline for reporting systematic reviews. *Bmj* 372, 790–799. doi:10.1016/j.rec.2021.07.010

Pearson, S. D., Ollendorf, D. A., and Chapman, R. H. (2019). New cost-effectiveness methods to determine value-based prices for potential cures: What are the options? *Value Health* 22 (6), 656–660. doi:10.1016/j.jval.2019.01.012

Peasgood, T., Mukuria, C., Rowen, D., Tsuchiya, A., and Wailoo, A. (2022). Should we consider including a value for "hope" as an additional benefit within health technology assessment? *Value Health* 25 (9), 1619–1623. doi:10.1016/j.jval.2022.03.006

Phelps, C. E., and Madhavan, G. (2017). Using multicriteria approaches to assess the value of health care. *Value Health* 20 (2), 251–255. doi:10.1016/j.jval.2016.11.011

Popat, S., Liu, S. V., Scheuer, N., Hsu, G. G., Lockhart, A., Ramagopalan, S. V., et al. (2022). Addressing challenges with real-world synthetic control arms to demonstrate the comparative effectiveness of Pralsetinib in non-small cell lung cancer. *Nat. Commun.* 13 (1), 3500. doi:10.1038/s41467-022-30908-1

Ramagopalan, S., Gupta, A., Arora, P., Thorlund, K., Ray, J., and Subbiah, V. (2021). Comparative effectiveness of atezolizumab, nivolumab, and docetaxel in patients with previously treated non-small cell lung cancer. *JAMA Netw. Open* 4 (11), e2134299. doi:10.1001/jamanetworkopen.2021.34299

Rand, L. Z., and Kesselheim, A. S. (2021). Controversy over using quality-adjusted life-years in cost-effectiveness analyses: A systematic literature review. *Health Aff. (Millwood).* 40 (9), 1402–1410. doi:10.1377/hlthaff.2021.00343

Reed, S. D., Dubois, R. W., Johnson, F. R., Caro, J. J., and Phelps, C. E. (2019). Novel approaches to value assessment beyond the cost-effectiveness framework. *Value Health* 22 (6s), S18–s23. doi:10.1016/j.jval.2019.04.1914

Rumeau-Pichon, C., and Harousseau, J. L. (2014). Analysis of cost-effectiveness assessments in France by the French national authority for health (has). *Value Health* 17 (7), A414. doi:10.1016/j.jval.2014.08.997

Sanders, G. D., Neumann, P. J., Basu, A., Brock, D. W., Feeny, D., Krahn, M., et al. (2016). Recommendations for conduct, methodological practices, and reporting of cost-effectiveness analyses: Second Panel on cost-effectiveness in health and medicine. *Jama* 316 (10), 1093–1103. doi:10.1001/jama.2016.12195

Sarri, G. (2022). Can real-world evidence help restore decades of health inequalities by informing health care decision-making? Certainly, and here is how. *Front. Pharmacol.* 13, 905820. doi:10.3389/fphar.2022.905820

Sarri, G., Freitag, A., Szegvari, B., Mountian, I., Brixner, D., Bertelsen, N., et al. (2021). The role of patient experience in the value assessment of complex technologies - do HTA bodies need to reconsider how value is assessed? *Health Policy* 125 (5), 593–601. doi:10.1016/j.healthpol.2021.03.006

Shafrin, J., Dennen, S., Pednekar, P., Birch, K., Bhor, M., Kanter, J., et al. (2021). For which diseases do broader value elements matter most? An evaluation across 20 ICER evidence reports. *J. Manag. Care Spec. Pharm.* 27 (5), 650–659. doi:10.18553/jmcp.2021.20471

Slejko, J. F., Mattingly, T. J., 2nd, Mullins, C. D., Perfetto, E. M., and dosReis, S. (2019). Future of patients in healthcare evaluation: The patient-informed reference case. *Value Health* 22 (5), 545–548. doi:10.1016/j.jval.2019.02.003

Thorlund, K., Dron, L., Park, J. J. H., and Mills, E. J. (2020). Synthetic and external controls in clinical trials - a primer for researchers. *Clin. Epidemiol.* 12, 457–467. doi:10.2147/CLEP.S242097

Toolan, M., Walpole, S., Shah, K., Kenny, J., Jónsson, P., Crabb, N., et al. (2023). Environmental impact assessment in health technology assessment: Principles, approaches, and challenges. *Int. J. Technol. Assess. Health Care* 39 (1), e13. doi:10.1017/S0266462323000041

Torbica, A., Tarricone, R., and Drummond, M. (2018). Does the approach to economic evaluation in health care depend on culture, values, and institutional context? *Eur. J. Health Econ.* 19 (6), 769–774. doi:10.1007/s10198-017-0943-1

van Overbeeke, E., Forrester, V., Simoens, S., and Huys, I. (2021). Use of patient preferences in health technology assessment: Perspectives of Canadian, Belgian and German HTA representatives. *Patient* 14 (1), 119–128. doi:10.1007/s40271-020-00449-0

Vrinzen, C. E. J., Bloemendal, H. J., Stuart, E., Makady, A., van Agthoven, M., Koster, M., et al. (2022). Cancer treatments touch a wide range of values that count for patients and other stakeholders: What are the implications for decision-making? *Cancer Med.* 12, 6105–6116. doi:10.1002/cam4.5336

Willke, R. J., Neumann, P. J., Garrison, L. P., Jr., and Ramsey, S. D. (2018). Review of recent US value frameworks-A health economics approach: An ISPOR special Task Force report [6]. *Value Health* 21 (2), 155–160. doi:10.1016/j.jval.2017.12.011

Youn, H. M., Quan, J., Mak, I. L., Yu, E. Y. T., Lau, C. S., Ip, M. S. M., et al. (2022). Long-term spill-over impact of COVID-19 on health and healthcare of people with non-communicable diseases: A study protocol for a population-based cohort and health economic study. *BMJ Open* 12 (8), e063150. doi:10.1136/bmjopen-2022-063150

Zamora, B., Garrison, L. P., Unuigbe, A., and Towse, A. (2021). Reconciling ACEA and MCDA: Is there a way forward for measuring cost-effectiveness in the U.S. Healthcare setting? *Cost. Eff. Resour. Alloc.* 19 (1), 13. doi:10.1186/s12962-021-00266-8

Zhang, M., Bao, Y., Lang, Y., Fu, S., Kimber, M., Levine, M., et al. (2022). What is value in health and healthcare? A systematic literature review of value assessment frameworks. *Value Health* 25 (2), 302–317. doi:10.1016/j.jval.2021.07.005

frontiers | Frontiers in Pharmacology

TYPE Systematic Review
PUBLISHED 08 August 2023
DOI 10.3389/fphar.2023.1220950

Check for updates

OPEN ACCESS

EDITED BY
Mauro Tettamanti,
Mario Negri Institute for Pharmacological Research (IRCCS), Italy

REVIEWED BY
Tanja Mueller,
University of Strathclyde, United Kingdom
Erik Koffijberg,
University of Twente, Netherlands

*CORRESPONDENCE
Claire Hawksworth,
claire.hawksworth@nice.org.uk

†These authors have contributed equally to this work

RECEIVED 11 May 2023
ACCEPTED 25 July 2023
PUBLISHED 08 August 2023

CITATION
Vithlani J, Hawksworth C, Elvidge J, Ayiku L and Dawoud D (2023), Economic evaluations of artificial intelligence-based healthcare interventions: a systematic literature review of best practices in their conduct and reporting.
Front. Pharmacol. 14:1220950.
doi: 10.3389/fphar.2023.1220950

COPYRIGHT
© 2023 Vithlani, Hawksworth, Elvidge, Ayiku and Dawoud. This is an open-access article distributed under the terms of the Creative Commons Attribution License (CC BY). The use, distribution or reproduction in other forums is permitted, provided the original author(s) and the copyright owner(s) are credited and that the original publication in this journal is cited, in accordance with accepted academic practice. No use, distribution or reproduction is permitted which does not comply with these terms.

Economic evaluations of artificial intelligence-based healthcare interventions: a systematic literature review of best practices in their conduct and reporting

Jai Vithlani[1†], Claire Hawksworth[2*†], Jamie Elvidge[2], Lynda Ayiku[2] and Dalia Dawoud[1,3]

[1]National Institute for Health and Care Excellence, London, United Kingdom, [2]National Institute for Health and Care Excellence, Manchester, United Kingdom, [3]Faculty of Pharmacy, Cairo University, Cairo, Egypt

Objectives: Health economic evaluations (HEEs) help healthcare decision makers understand the value of new technologies. Artificial intelligence (AI) is increasingly being used in healthcare interventions. We sought to review the conduct and reporting of published HEEs for AI-based health interventions.

Methods: We conducted a systematic literature review with a 15-month search window (April 2021 to June 2022) on 17th June 2022 to identify HEEs of AI health interventions and update a previous review. Records were identified from 3 databases (Medline, Embase, and Cochrane Central). Two reviewers screened papers against predefined study selection criteria. Data were extracted from included studies using prespecified data extraction tables. Included studies were quality assessed using the National Institute for Health and Care Excellence (NICE) checklist. Results were synthesized narratively.

Results: A total of 21 studies were included. The most common type of AI intervention was automated image analysis (9/21, 43%) mainly used for screening or diagnosis in general medicine and oncology. Nearly all were cost-utility (10/21, 48%) or cost-effectiveness analyses (8/21, 38%) that took a healthcare system or payer perspective. Decision-analytic models were used in 16/21 (76%) studies, mostly Markov models and decision trees. Three (3/16, 19%) used a short-term decision tree followed by a longer-term Markov component. Thirteen studies (13/21, 62%) reported the AI intervention to be cost effective or dominant. Limitations tended to result from the input data, authorship conflicts of interest, and a lack of transparent reporting, especially regarding the AI nature of the intervention.

Conclusion: Published HEEs of AI-based health interventions are rapidly increasing in number. Despite the potentially innovative nature of AI, most have used traditional methods like Markov models or decision trees. Most attempted to assess the impact on quality of life to present the cost per QALY gained. However, studies have not been comprehensively reported. Specific reporting standards for the economic evaluation of AI interventions would help improve transparency and promote their usefulness for decision making. This is fundamental for reimbursement decisions, which in turn will generate the necessary data to develop flexible models better suited to capturing the potentially dynamic nature of AI interventions.

KEYWORDS

artificial intelligence, cost effectiveness, cost utility, simulation models, health economic evaluation, mixed-methods, systematic review

1 Introduction

The use of artificial intelligence (AI) has significantly grown in the healthcare sector. Exploiting its ability to streamline tasks, provide real-time analytics, and process larger quantities of data has contributed to its increased prominence (Panch et al., 2018). Additionally, it may have the potential to deliver quality care at lower costs. AI is being used to address challenges ranging from staff shortages to ageing populations and rising costs (Dall et al., 2013). The number of AI technologies approved by the US Food and Drink Administration (FDA) was nearly 350 between 2016 and mid-2021, compared to less than 30 in the preceding 19 years (Miller, 2021).

Several systematic reviews have been published that examine health economic evaluations (HEEs) for AI in healthcare. The most recent is Voets et al. (1 April 2021) (Voets et al., 2022), who searched for publications from 5 years prior and included 20 full texts, discussing the methods, reporting quality and challenges. They found that automated medical image analysis was the most common type of AI technology, just under half of studies reported a model-based HEE, and the reporting quality was moderate. Overall, Voets et al. concluded that HEEs of AI in healthcare often focus on costs rather than health impact, and insight into benefits is lagging behind the technological developments of AI.

An up-to-date representation of the economic evidence base may be insightful. Clearly, AI is a rapidly developing area in healthcare, demonstrated by the National Institute for Health and Care Excellence (NICE) recently incorporating AI technologies into its Evidence Standards Framework (Unsworth et al., 2021; National Institute for Health and Care Excellence, 2022). While some of this rise may be attributable to changes in legislation, it indicates the importance of AI in the current healthcare climate and the need to have a contemporary understanding of its economic value. Additionally, the COVID-19 pandemic has led to a rapid increase in the digitalization of data and health services including teleconsultations, online prescriptions and remote monitoring (Gunasekeran et al., 2021). Therefore, we sought to update the Voets et al. systematic review. We report updated results consistent with the original review, by disaggregating the HEEs into costs, clinical effectiveness, modelling characteristics and methodologies to understand common techniques, limitations, assumptions, and uncertainties. This update allows us to advance the discussion around whether existing modelling methods and reporting standards are suitable to appropriately assess the cost effectiveness of AI technologies compared to non-AI technologies in healthcare.

This review was undertaken to inform ongoing work within the HTx project. HTx is a Horizon 2020 project supported by the European Union lasting for 5 years from January 2019. The main aim of HTx is to create a framework for the Next-Generation Health Technology Assessment (HTA) to support patient-centred, societally oriented, real-time decision-making on access to and reimbursement for health technologies throughout Europe.

2 Data and methods

2.1 Literature search strategy

The search strategy included the period from 1 April 2021 to 17 June 2022, in order to update the original search conducted by Voets et al. (Voets et al., 2022). The original search used the PubMed and Scopus databases. For the present update, the original search strategy was translated for use in MEDLINE, EMBASE, via the Ovid platform, and Cochrane Central, via Wiley. These databases were preferred due to their accessibility, and searching all 3 was considered to provide comparable coverage to PubMed and Scopus (Ramlal et al., 2021).

The search strategy was simplified into 2 concept pathways: 1. "Artificial intelligence" and 2. "Health economic evaluations". The search queries in Supplementary Appendix SA show the strategies divided into their respective databases. Subsequent terms in the AI pathway included, "artificial intelligence", "machine learning", and "data driven". The second pathway included terms such as, "cost effectiveness", "health outcomes", "cost", "budget". An English language query was applied to the search strategy. The initial database selection and search strategies were guided by NICE information specialists. The review and search protocol were not registered.

2.2 Inclusion and exclusion criteria

Studies were included if they were a HEE of an AI intervention and a comparator, such as current standard of care or a non-AI intervention. This included trial-based economic evaluations and model-based studies. There were no exclusion criteria on types of economic evaluation, such that cost-effectiveness analyses (CEAs), cost-utility analyses (CUAs), cost-minimization analyses (CMA) and budget impact analyses (BIAs) were included. We term all of these as HEEs, which are defined as the "comparative analysis of alternative courses of action in terms of both their costs and consequences" (Rudmik and Drummond, 2013). CEAs evaluate whether an intervention provides relative value, in terms of cost and health outcomes, to a respective comparator. CUAs are a subset of CEAs where the health outcome includes a preference-based measure such as the Quality Adjusted Life Year (QALY). BIA studies evaluate the affordability of an intervention for payers to allocate resources. Included studies reported a quantitative health economic outcome such as costs, or costs in relation to effectiveness. For the exclusion criteria in the initial screening of titles and abstracts, studies that were not original research or systematic reviews such as commentaries, letters, and editorials were excluded. Overall, the inclusion and exclusion criteria were consistent with Voets et al. (Voets et al., 2022).

After duplicates were removed, 2 reviewers independently screened titles, and abstracts. The reviewers discussed any discrepancies, and where agreement could not be reached, an

independent third reviewer was consulted. The same process was followed for subsequent full-text screening.

2.3 Data extraction

The data extraction was initially completed by 1 reviewer, and then validated by a second reviewer who independently extracted and compared data from the included studies. The extraction strategy was divided into three components, the first and second components included the characteristics and the methodological details of the studies. The former included aspects such as the purpose of the AI technology, medical field, funding, care pathway phase (prevention, diagnostics, monitoring, treatment) and the type of AI (i.e., pattern recognition, risk prediction, etc.). The second table of methodological details included aspects such as the type of HEE, the comparator, and the outcome measure. The third component was relevant only for model-based HEEs, extracting parameters such as model states, time horizon, and details of sensitivity analyses.

2.4 Data analysis

The extracted data were synthesised using a narrative approach as heterogeneity between studies inhibited the utility of a quantitative synthesis. Descriptive statistics were used to summarize the characteristics of the retrieved studies, where appropriate.

2.5 Quality assessment

The quality assessment of all included studies was conducted using the NICE quality appraisal checklist for economic evaluations (National Institute for Health and Care Excellence, 2012). This checklist has been adopted in the literature of economic evaluation reviews (Elvidge et al., 2022) and is used by NICE when assessing HEE evidence for all public health guidelines. Included studies with a decision-analytic model were quality assessed independently by 2 reviewers using the methodological checklist section of the quality appraisal checklist. The checklist has 11 individual questions to create an overall assessment of whether there are minor-, potentially serious-, or very serious limitations that affects the robustness of the results. Quality assessment was not used as part of the exclusion criteria, as one of the research aims was to explore the reporting standards.

Although it is not possible to fully remove the potential of bias due to the subjective nature of the assessment, pre-set criteria were created to minimize its effects. The criteria are as follows: studies with very serious limitations included studies that had significant modelling discrepancies that could materially change the cost-effectiveness conclusion (e.g., the intervention changing from dominant to dominated). Also, very serious limitations are derived from a financial conflict of interest, where the developer of the AI technology also funded the HEE. Potentially serious limitations refer to methodological uncertainties which may change the quantitative result (e.g., an increase in the cost-effectiveness ratio), however the outcome could stay the same (e.g., the increase is not meaningful). All other limitations were considered to be minor limitations. The reviewers discussed any discrepancies in their quality assessments, and if major disagreements emerged, an independent third reviewer was consulted.

3 Results

3.1 Search results

The searches across the 3 databases yielded 4,475 records, resulting in 3,033 unique records following deduplication (Table 1). After screening titles and abstracts against the study selection criteria 2,993 were excluded due to not relating to a human health intervention, not reporting a HEE, not relating to an AI-based intervention, or being a excludable study type (e.g., commentary). Therefore, 40 studies proceeded to full-text screening. Of those, 16 were excluded based on the selection criteria, and 2 were excluded as duplicates that had already been included in the Voets et al. review (Voets et al., 2022). We excluded a further study due to unclear reporting about whether it was a primary analysis or a review of other economic models. Therefore, 21 studies remained which were suitable for data extraction. See Figure 1 for the PRISMA flowchart showing the inclusion and exclusion stages.

3.2 Overview of included studies

The general characteristics of the 21 included studies are presented in Table 2. The majority were published in 2022. There was a wide variation of AI interventions in different medical fields. The most frequent were general medicine and oncology (each 4/21, 19%), followed by ophthalmology and respiratory medicine (each 3/21, 14%), cardiology (2/21, 10%), and dermatology, mental health, radiology, sleep and analgesics (each 1/21, 5%). The interventions spanned the screening (9/21, 43%), diagnosis (8/21, 38%), treatment (1/21, 5%) and monitoring (3/21, 14%) stages of the clinical pathway. The most common type of AI evaluated was automated image analysis (9/21, 43%). Others were risk prediction (6/21, 29%), pattern recognition (2/21, 10%), personalized treatment recommendation (1/21, 5%), clinical decision support (1/21, 5%) and combined risk prediction and clinical decision support (2/21, 10%). Most studies were funded by governments and industry (each 5/21, 24%), followed by academia (3/21, 14%). Two (2/21, 10%) were jointly funded by industry and academia and one (1/21, 5%) was funded by the European Commission.

3.3 HEE characteristics

The 21 HEEs contained 10 (10/21, 48%) CUAs, 8 (8/21, 38%) CEAs and 2 (2/21, 10%) BIAs. One (1/21, 5%) HEE reported results as both a CEA and a CUA. Among the CEAs the outcomes ranged from cost saved per patient screened, cost per death averted, cost per DALY averted, cost per case prevented and cost saving per

TABLE 1 Database search results.

Databases	Date searched	Database version	Number of records retrieved
Medline (Ovid)	17th June 2022	Ovid MEDLINE(R) ALL <1946 to 16 June 2022>	1,876
Embase (Ovid)	17th June 2022	Embase <1974 to 2022 June 16>	2,529
Cochrane Central (Wiley)	17th June 2022	Issue 5 of 12, May 2022	70
			4,475

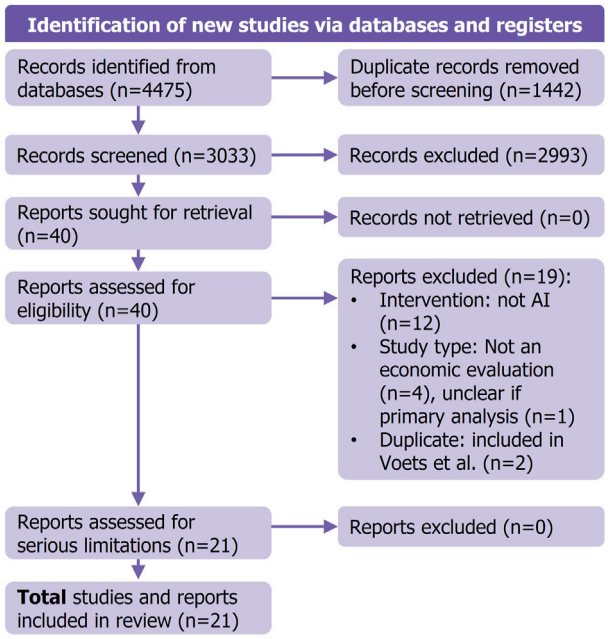

FIGURE 1
PRISMA flowchart describing study selection and reasons for exclusion during full-text screening.

additional tooth retention year. The healthcare system perspective was the most common. Of the 21, 10 (10/21, 48%) took a healthcare system perspective, 6 (6/21, 29%) payer, 4 (4/21, 19%) societal and 1 study (1/21, 5%) took both a societal and health system perspective. In some studies, the payer perspective represented insurers, both public and private.

The time horizon for the 21 studies ranged from 8 weeks to lifetime, with lifetime being the most common (5/21, 24%). One year was the second most common time horizon (3/21, 14%), followed by 6 months and 5 years with two each (2/21, 10%). Time horizons of 8 weeks, 16 months, 3 years, 15 years, 20 years, 30 years, and 35 years were all present in one study each (1/21, 5%). In two studies the time horizon was not reported (2/21, 10%). Most HEEs with a time horizon longer than 1 year used a 3% annual discount rate (7/13, 54%). Six studies discounted costs and health outcomes differentially. Of these, 2 studies (2/13, 15%) discounted costs at 4% and health outcomes at 1.5%, 2 (2/13, 15%) discounted the costs but did not report discount rates for health outcomes, 1 (1/13, 8%) used undiscounted costs but did not report discounting of health outcomes, and 1 (1/13, 8%) did not report discount rates for the costs but discounted health outcomes at 3%. Table 3 reports all the methodological details of the included HEEs.

3.4 Modelling characteristics

Of the 21 HEEs, 16 (16/21, 76%) included a decision analytic model. The modelling characteristics of these are summarized in Table 4. The most common model types were Markov models (6/16, 38%) and decision trees (4/16, 25%) with 3 (3/16, 19%) using a short-term decision tree followed by a longer-term Markov component. Of the remaining 3 studies, there was 1 cost simulation, 1 Markov chain Monte Carlo simulation, and 1 hybrid decision tree and microsimulation model. Authors typically justified their chosen model type by linking the decision to the type of AI intervention, the outcome measure, and the time horizon. Most Markov models used a cycle length of 1 year, and the rest used 1 month or 1 day. Studies that used decision tree models stated their primary reason for doing so was for their simplicity.

In terms of results, 7 (7/21, 33%) HEEs reported the AI intervention was cost effective versus the comparator relative to an appropriate threshold value, 5 (5/21, 24%) demonstrated that the AI intervention was dominant, and 2 (2/21, 10%) demonstrated equivalence. In 1 (1/21, 5%) study the AI intervention was cost effective versus one comparator and dominant versus the other. In 2 (2/21, 10%) studies the AI interventions produced savings. Three (3/21, 14%) studies did not state a preferred cost-effectiveness threshold to determine if the result was cost effective. The AI intervention was found to be cost ineffective in 1 (1/21, 5%) study.

Of the studies that reported sensitivity analysis (18/21, 86%), 17 reported one-way sensitivity analyses, though the remaining study did conduct probabilistic sensitivity analysis. Seven (7/21, 33%) studies reported both one-way and probabilistic analyses, while 4 (4/21, 19%) reported both one-way and scenario analyses. Three studies (3/21, 14%) reported one-way, probabilistic and scenario analyses.

3.5 Quality assessment

A summary of the results from the quality appraisal checklist is shown in Table 5. The assessment resulted in 6 (6/21, 29%) studies with very serious limitations, 11 (11/21, 52%) with potentially serious limitations, and 4 (4/21, 19%) with minor limitations. Initially the two reviewers disagreed on the assessment for two of the studies (Ericson et al., 2022; Mital and Nguyen, 2022). Both were upgraded for the reasons given below.

TABLE 2 Characteristics of the included studies.

Main author	Year	Population	Location	Description of AI intervention	Medical field	Care pathway phase	AI technology	Funding
Adams et al. (2021)	2021	A representative cohort of 3,197 baseline screening patients	United States	Risk score predictor	Oncology	Screening	Risk prediction	Industry
Areia et al. (2022)	2022	A hypothetical cohort of 100,000 individuals aged 50–100 years	United States	AI tools to detect precancerous polyps during colonoscopy	Oncology	Screening	Pattern recognition	EU Commission and JSPS
de Vos et al. (2022)	2022	Dutch Patients	Holland	Decision-making support tool to discharge patients from ICU	General	Diagnostic	Clinical decision support	None
Delgadillo et al. (2022)	2022	Patients with common mental health disorders	United Kingdom	Decision-support tool providing personalized treatment recommendations (stratified care)	Mental health	Treatment	Personalised treatment recommendation	Industry and Academia
Ericson et al. (2022)	2022	Adult patients who were not diagnosed with sepsis at the time of admission	Sweden	Early detection of sepsis	General	Diagnostic	Risk prediction	Industry
Fusfeld et al. (2022)	2022	Kidney transplant recipients receiving a for-cause biopsy	United States	MMDx-Kidney assesses the probability of biopsy rejection or injury	General	Diagnostic	Pattern recognition	Industry
Huang et al. (2022)	2022	Diabetes patients without retinopathy	Rural China areas	Automated retinal image analysis system for diabetic retinopathy screening	Ophthalmology	Screening	Automated image analysis	Industry and Academia
Kessler et al. (2021)	2021	High-risk Medicaid members with multiple chronic conditions	Southern California, United States	Risk score predictor and decision-support for pharmacists offering medicine management to high-risk Medicaid members	General	Monitoring	Risk prediction and decision support	Government
MacPherson et al. (2021)	2021	Adults attending acute primary services	Malawi	Computer-aided digital chest x-ray (DCXR-CAD) for HIV-TB screening	Respiratory	Screening	Automated image analysis	Academia
Mallow and Belk (2021)	2021	Hypothetical cohort undergoing elective orthopedic procedures that commonly have opioids prescribed	United States	Machine learning algorithm analyzing alleles involved in reward pathway of the brain to identify patients with a higher risk of opioid use (OUD)	Analgesics	Diagnostic	Risk prediction	Industry
Mital and Nguyen (2022)	2022	Women aged 40–49	United States	AI to read mammography images to predict breast cancer risk	Oncology	Screening	Automated image analysis	None
Morrison et al. (2022)	2022	Theoretical cohort of infants requiring ROP screening	United States	Artificial intelligence (AI)based retinopathy of prematurity (ROP) screening. Both assistive and autonomous	Ophthalmology	Screening	Automated image analysis	Academia
Nsengiyumva et al. (2021)	2021	Patients with symptoms suggestive of pulmonary TB	Pakistan	AI-based radiograph to triage persons with possible tuberculosis	Respiratory	Diagnostic	Automated image analysis	Government

(Continued on following page)

TABLE 2 (*Continued*) Characteristics of the included studies.

Main author	Year	Population	Location	Description of AI intervention	Medical field	Care pathway phase	AI technology	Funding
				symptoms and identification of those who require further testing				
Salcedo et al. (2021)	2021	Adults undergoing active TB treatment	United States	Monitors real-time medication consumption and adherence for TB treatment	Respiratory	Monitoring	Automated image analysis	Government
Schwendicke et al. (2022)	2022	31-year-olds, whose proximal surfaces were initially either good, or in an E2, D1 or D2-3 lesion	Germany	AI-based software to detect proximal caries lesions	Dentistry	Diagnostic	Automated image analysis	None
Szymanski et al. (2022)	2022	Adults aged 65 years or older registered with a GP	United Kingdom	AF risk prediction algorithm to improve AF detection	Cardiology	Screening	Risk prediction	Industry
Tseng et al. (2021)	2021	Hypothetical cohort of asymptomatic 65-year-olds	US	AI ECG algorithm to detect asymptomatic left ventricular dysfunction	Cardiology	Screening	Risk prediction	Academia
Turino et al. (2021)	2021	Adults with newly diagnosed obstructive sleep apnea	Spain	AI monitoring system for improving CPAP compliance	Sleep	Monitoring	Risk prediction and decision support	Government
van Leeuwen et al. (2021)	2021	71,840 adults aged 66 years from a stroke registry that received CTA diagnosis work up of acute stroke	United Kingdom	AI software aiding detection of intracranial LVO in stroke patients	Radiology	Diagnostic	Automated image analysis	None
Xiao et al. (2021)	2021	Asymptomatic adults aged 65 years and above for population screening	China	AI diagnosis of glaucoma	Ophthalmology	Screening	Automated image analysis	Government
Ziegelmayer et al. (2022)	2022	60-year-olds with 20 pack years of smoking history	United States	AI convolutional neural networks supported low dose CT at initial screening for lung cancer	Oncology	Diagnostic	Risk prediction	None

Atrial Fibrillation, AF; artificial intelligence, AI; continuous positive airway pressure, CPAP; CTA, computed tomography angiography; ECG, electrocardiography; European Union, EU; general practice, GP; intensive care unit, ICU; japan society for the promotion of science, JSPS; LVO, large vessel occlusions; Molecular microscope diagnostic system, MMDx; Opioid use disorder, OUD; retinopathy of prematurity, ROP; ROP; tuberculosis, TB.

Studies deemed to have very serious limitations were those where an issue in 1 or more quality criteria were highly likely to materially change the cost-effectiveness conclusion for the AI intervention. There were several key reasons which led to this assessment for 5 of the included studies. In one there was an acknowledged overestimation of cost data, representation issues between the dataset and target population, and a short 6-month horizon rather than the 12-month time horizon deemed best practice by the American College of Radiology (Rosenthal and Dudley, 2007). In another, adverse health effects were not captured, which the authors suggested would increase the cost-effectiveness estimate (Fusfeld et al., 2022). This study also had a financial conflict of interest where research was funded by the company which developed the AI intervention. This was true for another 2 studies (Ericson et al., 2022; Szymanski et al., 2022). In another study, the result changed from intervention dominant to cost ineffective when input data, arising from multiple sources and assumption, were varied during the sensitivity analyses (Ziegelmayer et al., 2022).

Studies with potentially serious limitations tended to have a paucity of appropriate input data. Instead, alternative sources, or multiple sources were used with resulting generalizability issues. It was common for studies to have assumptions for the cost and effectiveness of the AI intervention, compliance, and the impact of the AI intervention on the subsequent treatment pathway. Examples of this are 1 study that assumed all patients would consent to a test (Mallow and Belk, 2021); 1 study that used a primary outcome that was patient reported (Delgadillo et al., 2022) and 1 study that assumed the effectiveness of the AI intervention last for 10 years, despite having data for only 5 years (Mital and Nguyen, 2022). These

TABLE 3 Health economic details of included studies.

Main author	HEE type	Intervention	Comparator	Perspective	Discount rate	Time horizon	Outcome measure
Adams et al. (2021)	CEA	Combining Artificial Intelligence and Lung-RADS	Lung-RADS	Payer	NA	6 months	Cost saving of AI-informed management per patient screened
Areia et al. (2022)	CEA	AI detection of polyps	Screening without AI tools	Societal	3%	30 years	Cost saving of screening with AI per individual
de Vos et al. (2022)	CUA	AI decision support tool for ICU discharge decision-making	Standard care discharge decisions based on medical expertise	Societal	Costs 4%, Health outcomes 1.5%	1 year	ICER- cost per QALY gained
Delgadillo et al. (2022)	CEA	AI personalized treatment recommendation to provide stratified care	Standard of care- stepped care	Healthcare system	NR	NR	Incremental cost of stratified care per patient and additional case of reliable improvement
Ericson et al. (2022)	CUA/ CEA	AI detection of sepsis	Standard care for sepsis diagnosis	Healthcare system	3%	1 year	Cost savings per patient
Fusfeld et al. (2022)	BIA	Pattern recognition in gene expression in biopsy	Histology biopsy alone	Payer	Costs 0%, Health outcomes NR	5 years	Cost per patient and savings per biopsy
Huang et al. (2022)	CUA	AI based DR screening	No screening or ophthalmologist screening	Healthcare system and societal	3%	35 years	ICER- cost per QALY gained
Kessler et al. (2021)	CEA	AI risk score predictor and decision-support for medication management	The same cohort pre-AI intervention start	Payer	NR	NR	Savings per member, per month
MacPherson et al. (2021)	CUA	AI chest x-ray interpretation providing a probabilistic score for TB	Standard of care	Healthcare system	NA	8 weeks	ICER- cost per QALY gained
Mallow and Belk (2021)	CUA	AI prediction to decrease risk of OUD	Current standard of care	Payer	3%	5 years	ICER- cost per QALY gained
Mital and Nguyen (2022)	CUA	Automated mammography image analysis	Alternative screening strategies including no screening, screening guided by risk scores (PRS) and screening guided by family history	Healthcare system	3%	Lifetime	ICER- cost per QALY gained
Morrison et al. (2022)	CUA	Deep learning algorithm	Telemedicine and Ophthalmoscopy	Healthcare system	Costs NR, Health outcomes 3%	Lifetime	ICER- cost per QALY gained
Nsengiyumva et al. (2021)	CEA	AI detection of TB	No AI triage before microbiologic testing. Current standard of care- smear microscopy or GeneXpert	Payer	NA	1 year	Incremental cost per DALY averted
Salcedo et al. (2021)	CUA	AI monitoring for tuberculosis treatment adherence	Standard of care: DOT	Societal	NR	16 months	ICER- cost per QALY gained and NMB
Schwendicke et al. (2022)	CEA	AI detection for proximal caries	Caries detection without AI	Payer	Costs 3%, Health outcomes NR	Lifetime	ICER- cost per year of tooth retention gained
Szymanski et al. (2022)	BIA	AI risk score predictor to detect AF using data from baseline risk factors	Standard care (opportunistic screening and diagnosis) or combined use of standard care and AI	Healthcare system	NR	3 years	Budget impact in £
Tseng et al. (2021)	CUA	AI detection of ALVD	No screening	Healthcare system	3%	Lifetime	ICER- cost per QALY gained

(Continued on following page)

TABLE 3 (*Continued*) Health economic details of included studies.

Main author	HEE type	Intervention	Comparator	Perspective	Discount rate	Time horizon	Outcome measure
Turino et al. (2021)	CEA	AI monitoring of CPAP compliance	Standard of care	Healthcare system	NA	6 months	Cost per hour of CPAP compliance gained per day
van Leeuwen et al. (2021)	CUA	AI software aiding detection of intracranial large vessel occlusions LVO	Standard of care	Societal	Costs 4%, Health outcomes 1.5%	Lifetime	Incremental cost, incremental effects
Xiao et al. (2021)	CEA	AI detection of glaucoma	No screening	Healthcare system	Costs 5%, Health outcomes NR	15 years	Incremental cost of PACG prevented
Ziegelmayer et al. (2022)	CUA	AI-based CT scan	Stand alone low-dose CT scan	Healthcare system	3%	20 years	ICER- cost per QALY gained

Artificial Intelligence, AI; budget impact assessment, BIA; computerized tomography, CT; cost effectiveness analysis, CEA; cost utility analysis, CUA; diabetic retinopathy, DR; directly observed therapy, DOT; Disability-adjusted life years, DALY; large vessel occlusions, LVO; left ventricular systolic dysfunction, LVSD; net monetary benefit, NMB; opioid use disorder, OUD; Primary angle-closure glaucoma, PACG; reporting and data system, RADS.

studies did account for the key uncertainties in sensitivity analyses and the effect was either minor or the initial assumptions were shown to be robust. Some studies were assessed as having potentially serious limitations due to unclear reporting, which reduced transparency around key information such as whether a cost had been applied for the AI intervention, how it would integrate with clinical care, and who the anticipated user of the AI intervention was.

4 Discussion

This paper systematically reviewed 21 HEEs of AI interventions. The studies mainly evaluated AI-based automated image analysis interventions for diagnosis and screening in general medicine, oncology and ophthalmology. Nearly all were CUAs and CEAs that took a healthcare system or payer perspective, and most had lifetime time horizons. Some of the HEEs were trial-based analyses, but the large majority were model-based which mostly used Markov models. In terms of the HEE results, the AI interventions were cost effective or dominant in just over half and all the studies performed sensitivity analyses.

This study reports an updated search to the review conducted by Voets et al. (Voets et al., 2022), providing a contemporary snapshot of the HEE evidence base for AI health technologies Our update captures an additional 15-month period in a time where AI health based technologies are on the exponential rise, evidenced by the near quadruple number of initial unique search results since April 2021 (Voets et al., 2022). It appears there has been no change in the most commonly evaluated purpose of AI being used as a healthcare intervention, as Voets et al. also found the most common to be automated image analysis (Voets et al., 2022). Ophthalmology and screening were the dominant specialty and phase of the care pathway at which the AI intervention was used, and these were also prevalent in this updated review. The prevailing type of HEE in the original review was cost minimization with the preferred outcome measure of cost saved per case identified. This was common among our included studies, although we termed it CEA, but CUA was the most common study type in this update. There was a difference between the two reviews in how many of the technologies were found to be cost saving. Voets et al. found the majority were whilst this was true for only 2 studies in this review. This could be due to differences in applying the terms 'cost-saving' and 'cost-effective' as a large proportion of studies in this updated review were cost-effective.

Another difference was the fact that the large majority of HEEs in our review were model-based, compared to 45% of those in Voets (Voets et al., 2022). This could suggest a shift towards using models to estimate future costs and benefits of AI technologies, permitting longer time horizons than trial-based evaluations (the most common time horizon is our review was lifetime, compared to 1 year in Voets). Furthermore, the increasing use of model-based evaluations may suggest AI interventions are moving towards traditional value assessment frameworks that are commonplace in the health technology assessment of medicines. This increase in model-based technologies may also explain the differences in results regarding cost saving versus cost effective. Perhaps it is easier or more expected to generate cost-effectiveness estimates when using a model compared to non-model HEEs where it may be more common to focus on costs.

Voets et al. (2022) found that the evidence supporting the chosen analytical methods, assessment of uncertainty, and model structures was underreported. Our quality assessment determined that most studies had potentially serious limitations tending to arise from the sources and assumptions regarding the input data. These findings are consistent, which suggests that despite an increase in the use of more sophisticated economic evaluation techniques, the evidence supporting them remains limited. In some cases, the uncertainty and lack of clarity for the reader were due to the reporting of the HEE rather than the data quality. In numerous studies it was hard to determine fundamentals such as whether a cost had been applied for the AI intervention, how it would integrate with clinical care and who the anticipated user of the AI intervention was. As mentioned, not all of the studies we identified clearly stated how the AI intervention would integrate with clinical care. Studies did not typically thoroughly or transparently estimate subsequent care

TABLE 4 Summary of economic evaluation parameters and outcomes.

Main author	Model type	Model states/tree summary	Time horizon, cycle length	Sensitivity analysis	Outcome	Result
Adams et al. (2021)	Cost simulation	NR	6 months	NR	USD 72 to USD 242 saved per patient screened	Intervention cost-effective
Areia et al. (2022)	Markov model	No colorectal neoplasia; low risk adenomas, high risk adenomas, localized, regional, or distant CRC; and CRC-related death	30 years, 1 year	One-way and probabilistic analysis	0.1% absolute (6.9% relative) reduction in colorectal mortality vs. screening without AI, USD 57 saving per individual screened	Intervention cost-effective
de Vos et al. (2022)	Markov model	ICU ineligible, ICU eligible, General ward, Readmission ICU ineligible, Readmission ICU eligible, Discharged, Death	1 year, 1 day	One-way, probabilistic and scenario analysis	EUR 18,507 per QALY gained vs. standard care	Intervention cost-effective
Delgadillo et al. (2022)	Within trial analysis	NR	NR	NR	Incremental cost of stratified care was £104.50 per patient	Intervention potentially cost-effective. Threshold NR
Ericson et al. (2022)	Decision tree	True- and false-positive and true negative detections for sepsis	1 year	One-way and probabilistic analysis	CEA: 356 ICU deaths averted, EUR 2.8m saved/CUA: negative ICER, higher effect, lower cost	Intervention dominant
Fusfeld et al. (2022)	Decision tree	Functioning initial transplant, graft failure + re-transplant, graft failure + dialysis, death with functioning graft, death after graft failure	5 years	One-way and scenario analysis	Savings of USD 19,721 per biopsy over a 5 year period	Produces savings to commercial payers within 2 years
Huang et al. (2022)	Markov model	DR, Mild DR, Moderate DR, VTDR, Stable DR, Blindness and death	35 years, 1 year	One-way and probabilistic analysis	Using health system perspective: USD 1,107.63/QALY vs. no screening, Dominant vs. ophthalmologist screening. Using societal perspective: USD 10,347.12/QALY vs. no screening, Dominant vs. ophthalmologist screening	Intervention cost-effective using both perspectives
Kessler et al. (2021)	Regression analysis	NR	Mean of 20.5 weeks	NR	Saving of USD 554 per member per month	Produces savings
MacPherson et al. (2021)	Within trial analysis	NR	8 weeks	One-way sensitivity analysis	USD 4,520.47 per QALY gained vs. standard of care	Intervention not cost-effective
Mallow and Belk (2021)	Markov chain Monte Carlo simulation model	Alive and Dead. For those who developed OUD: OUD, treatment, remission, dead	5 years, 1 month	One-way, probabilistic and scenario analysis	USD 2,510 saving per patient, 0.02 QALY gain (private insurers), USD 2,682 saving per patient, 0.02 QALY gain (self-insured employers)	Intervention dominant using both perspectives
Mital and Nguyen (2022)	Hybrid decision tree/microsimulation model	No screening, Annual screening for all, AI + no screening for low risk, AI + biennial screening for low risk, PRS + no screening for low risk, PRS + biennial screening for low risk, Family history + no screening for low risk, Family history + biennial screening for low risk. For all interventions any deemed high risk moved to annual screening	Lifetime, 1 year	One-way and probabilistic analysis	AI + no screening for low risk dominated PRS + no screening for low risk, family history + biennial screening for low risk, PRS + biennial screening for low risk, AI + biennial screening for low risk and annual screening for all and extendedly dominated family history + no screening for low risk. USD 23,755 per QALY gained vs. no screening	Intervention cost-effective vs. no screening and dominant vs. other comparators
Morrison et al. (2022)	Decision tree	Ophthalmoscopy, Telemedicine, Assistive AI, Autonomous AI	Lifetime	One-way and probabilistic analysis	Autonomous AI less costly and as effective as telemedicine and ophthalmoscopy. Assistive AI USD 83,350 vs. telemedicine and dominated ophthalmoscopy	Intervention cost-effective

(Continued on following page)

TABLE 4 (*Continued*) Summary of economic evaluation parameters and outcomes.

Main author	Model type	Model states/tree summary	Time horizon, cycle length	Sensitivity analysis	Outcome	Result
Nsengiyumva et al. (2021)	Decision tree	Triage with AI-based CXR followed by standard of care with upfront smear or GeneXpert	1 year	One-way and scenario analysis	USD 43/DALY averted vs. smear as microbiologic test. Dominant vs. GeneXpert as microbiologic test	Intervention cost-effective
Salcedo et al. (2021)	Markov model	On treatment, Completed treatment, Defaulted	16 months, 1 month	One-way, probabilistic and scenario analysis	AI dominated DOT NMB: USD 3,142, 4,057 and 4,973 at WTP thresholds of USD 50, 100 and 150K respectively	Intervention dominant
Schwendicke et al. (2022)	Markov model	Sound|E1-2|D1|D2-3, True or false negative, No treatment, Development or progression, Restorative Treatment; True or false positive, Treatment, According to dentists' decision making in each group, Arrested, Restorative treatment	Lifetime, 1 year	One-way sensitivity analysis	AI and no AI showed identical effectiveness and nearly identical costs	Equivalence
Szymanski et al. (2022)	Budget impact model	Opportunistic screening or AI screening, ECG assessment	3 years	One-way and scenario analysis	Standard care + AI generated savings of £71,345,158 and improved clinical outcomes vs. standard care. AI alone generated savings of £80,441,386 but had worse clinical outcomes vs. standard care	Intervention potentially cost-effective. Threshold NR
Tseng et al. (2021)	Decision tree and Markov model	No Screen, Screen with AI algorithm; Treated ALVD, Untreated ALVD, Symptomatic, Untreated no ALVD, Dead	Lifetime, NR	One-way and probabilistic analysis	USD 43,351/QALY vs. no screening	Intervention cost-effective
Turino et al. (2021)	Within trial analysis	NR	6 months	Probabilistic sensitivity analysis	Mean increase of 1.14 h in daily compliance with AI intervention. Non-significant difference in cost between interventions	Intervention cost-effective
van Leeuwen et al. (2021)	Decision tree and Markov model	Patients suspected of stroke receiving CTA, Large vessel occlusion, No or other vessel inclusion; No IAT eligible, IAT eligible; Occlusion detected, Occlusion not detected; mRS 0–5, Death	Lifetime, 1 year	One-way and scenario analysis	AI cost saving of USD 156,000 and gain of 0.01 QALY	Intervention dominant
Xiao et al. (2021)	Markov model	Primary angle closure suspect, primary angle closure, primary angle closure glaucoma, PACG-related unilateral blindness and PACG- related bilateral blindness	15 years, 1 year	One-way sensitivity analysis	USD 1,464 per PACG case prevented over 15 years. Additional healthcare costs from screening were not offset by decreased disease progression over 15 years	Intervention potentially cost-effective. Threshold NR
Ziegelmayer et al. (2022)	Decision tree and Markov model	Decision; CT, CT + AI; Markov; No BC true negative, No BC false positive, BC undetected false negative, BC after resection, BC palliative, Death	20 years, 1 year	One-way and probabilistic analysis	AI CT cost saving USD 67.62 vs. CT screening. AI CT incremental QALY 0.01 vs. CT screening	Intervention dominant

*Self-reported as a simulation model. Artificial Intelligence, AI; asymptomatic left ventricular dysfunction, ALVD; bronchial cancer, BC; chest radiograph, CXR; colorectal cancer, CRC; CTA, computed tomography angiography; Diabetic retinopathy, DR; intensive care unit, ICU; molecular microscope diagnostic system, MMDx; Net monetary benefit, NMB; not applicable, NA; not reported, NR; opioid use disorder, OUD; Primary angle-closure glaucoma, PACG; polygenic risk scores, PRS; standard of care, SOC.

TABLE 5 Summary of quality assessment of included studies.

Study	Notable limitations identified	Assessment
Adams et al. (2021)	Strict assumptions regarding underlying parameters, such as an overestimation of costs, which directly determine the intervention outcome. The 6-month time horizon was short of 12 months deemed best practice by the American College of Radiology, also potentially impacting cost-effectiveness. Finally, the dataset used was not representative of the target populations, notably "overrepresenting white persons and underrepresenting racial minorities"	Very serious limitations
Areia et al. (2022)	Misrepresentation of population data from clinical trials to clinical practice. The overall death rate modelled was lower than the actual. Assumption of compliance of tests and the linear relationship between cancer prevention effect and increased ADR were made, however impact on cost-effectiveness is not severe	Potentially serious limitations
de Vos et al. (2022)	Short time horizon due to literature available for input parameters. Made assumptions from non-Dutch sources which was controlled for with sensitivity analysis, but limits generalisability of results	Potentially serious limitations
Delgadillo et al. (2022)	There were weaknesses regarding the internal validity. The primary outcome was patient reported, and used a general measure rather than disorder specific measures. The majority of patients were white which has generalizability implications	Potentially serious limitations
Ericson et al. (2022)	Limitations arise from patients who should have been included for Sepsis, not included. The model base case was purposely set to be conservative to not exaggerate the positive effects, however the assumptions made limits the validity of the outcomes. Finally, the research and funding were funded by the company who developed the intervention, creating potential for bias	Very serious limitations
Fusfeld et al. (2022)	The model does not capture adverse events due to antirejection medication which they suspect MMDx would increase leading to uncertainty in the result. There is also a potential conflict of interest where the research was funded by the company which developed the AI technology	Very serious limitations
Huang et al. (2022)	Limited data available from study population led to values derived from other countries which were accounted for in sensitivity analysis. Data regarding sensitivity and specificity of the AI screening derived from one paper, but did not greatly affect cost effectiveness in the sensitivity analyses	Minor limitations
Kessler et al. (2021)	Retrospective observational study limits conclusions on causality. Clinical outcomes were not analyzed	Potentially serious limitations
MacPherson et al. (2021)	Trial-based analysis with small number of events and short follow up resulted in less precise treatment estimates. Study presence in the clinic may have modified health worker behaviour for standard of care. Alternative diagnoses to TB were not investigated	Minor limitations
Mallow and Belk (2021)	The model assumed all patients would consent to the test which excludes the costs and effects if patients refused. The model also did not exhaust all features of the treatment pathways due to the high number of possibilities	Potentially serious limitations
Mital and Nguyen (2022)	Main limitation is the cost of using AI for breast cancer prediction is not yet known in clinical practice which led to data retrieved from the European Society of Radiology. This was accounted for with one-way sensitivity analysis with all results holding. Data for efficacy of AI intervention extrapolated beyond studied period	Potentially serious limitations
Morrison et al. (2022)	Speculative assumptions and imprecision in model inputs. However, the authors used conservative estimates and performed sensitivity analyses. Model time horizon was lifetime despite the life expectancy in the population (very premature babies) being unknown	Potentially serious limitations
Nsengiyumva et al. (2021)	The analysis examines the intervention in low HIV prevalence, the accuracy of results may vary in high prevalence	Minor limitations
Salcedo et al. (2021)	The model did not consider possible side effects or delays in appropriate care due to less nurse contact. Relatively short time horizon that assumes equal quality of life post-treatment between arms	Potentially serious limitations
Schwendicke et al. (2022)	Range of sources for input data which will lead to a degree of bias, although accounted for in sensitivity analyses. Lacked validity as in practice treatment decision would not be based on image analysis only	Potentially serious limitations
Szymanski et al. (2022)	Used an unvalidated threshold to determine AF risk and assumed 100% adherence to ECG assessment which lacks external validity. Did not include cost of implementation. The study was funded by the AI developer	Very serious limitations
Tseng et al. (2021)	The data estimates for the baseline (SOLVD) probabilities and effects were based on a study published 30 years ago from the last RCT. The model was calibrated to use a prespecified threshold which was not varied in the sensitivity analyses. There is also a conflict of interest where the research was funded by the organization which developed the AI technology	Very serious limitations
Turino et al. (2021)	Patients with severe chronic pathologies were excluded which could limit the generalizability of results and the follow-up period is relatively short. The study collected EQ-5D data but did not report utility data	Potentially serious limitations
van Leeuwen et al. (2021)	Model relied on two key inputs that were assumptions: percentage of missed LVOs in practice, and the capability of the AI to reduce missed LVOs. These were both varied in the sensitivity analyses and result did not change. The model only included early presenters but IAT would also include late presenters which limits generalizability. The authors also assumed that false positives would be neutralized by the reader and would not lead to unnecessary care	Minor limitations

(Continued on following page)

TABLE 5 (*Continued*) Summary of quality assessment of included studies.

Study	Notable limitations identified	Assessment
Xiao et al. (2021)	The predictive accuracy of the intervention came from the literature and may not be generalizable to the setting. Any varying of this was not reported. There was a lack of robust data on the efficacy of treatment that followed a positive screening result which was accounted for in the sensitivity analysis	Potentially serious limitations
Ziegelmayer et al. (2022)	Input parameters came from multiple sources including assumptions and numerous published studies, leading to a degree of bias. Varying the specificity of the AI or CT and cost of AI greatly increased the ICER changing the result from intervention dominant to not cost-effective	Very serious limitations

and downstream health outcomes resulting from the use of an AI intervention. Our findings from this literature review suggest this is an area that needs to be better considered and reported.

AI-based interventions have the potential to be distinct from traditional medical interventions if they can learn (from data) over time. Theoretically, this means the relationship between the intervention and outcome may not be fixed; an AI intervention could get *more* effective over time, unlike the typical effect *waning* assumption associated with medicines. This has implications when considering future benefits and how to extrapolate this over the time horizon of the HEE. The prevailing model structures used in HEEs of AI interventions to date—Markov models, decision trees, and hybrids of the 2—may limit the extent to which studies have been able to capture and examine the dynamic nature of AI interventions. Therefore, there is the possibility that the existing HEE evidence base has not captured the true potential value of many AI interventions due to limitations imposed by their model structures, and only a third of our included studies explored the impact of structural uncertainty in sensitivity analysis. Furthermore, traditional, 'simple' models may not facilitate easy modelling of downstream costs and benefits, by quickly becoming slow or unwieldy. This, potentially, fails to show the full benefit of the AI intervention, inhibiting implementation. Guo et al. (Guo et al., 2020) acknowledge this through a paradox of "no evidence, no implementation—no implementation, no evidence". More sophisticated types of model, that are less restricted by the structural limitations that affect simple decision tree and Markov models may be better placed to capture full pathway effects in addition to potential time-dependent effectiveness of AI-based interventions.

Simulation-based modelling presents the opportunity to build flexible, sophisticated models that can overcome several limitations of Markov models and decision trees. They can easily incorporate the history of past events, model factors that can vary between patients and have a non-linear relationship with outcomes, and do not use discrete time intervals (Davis et al., 2014). They can also track the path of each person over time and estimate individual-level effects or mean group-level effects for a population (Davis et al., 2014). These possibilities may lead to models capable of addressing the potential dynamic nature of AI interventions learning over time and the impact on linked decision points and subsequent care in a clinical pathway. As data on AI-based interventions continues to be collected and reported, the ability to develop these models should improve. One thing to note, however, is that for these models to underpin reimbursement decisions HTA agencies would need to be able to critique and utilize them. This may require new skills, knowledge and experience and present other challenges. Utilizing these sorts of models also leads to the debate of whether HTA should be more 'living'. This refers to regular and scheduled updates of recommendations instead of the more traditional 'one-off' decisions. Living HTA presents opportunities as well as challenges (Thokala et al., 2023) and is not yet common practice.

The usefulness of a published HEE for decision making depends on how well it is conducted and reported. Reporting guidelines play an important role in improving transparency and completeness and as new technologies emerge, can help drive best practice. A prominent reporting standard within the field of HEEs is the Consolidated Health Economic Evaluation Reporting Standards (CHEERS) (Husereau et al., 2022). This outlines minimum reporting standards and was recently updated in 2022. It includes a 28-item checklist covering methodological approach, data identification, model inputs, assumptions, uncertainty analysis, and conflicts of interest. It does not include any reporting items that are specific to any AI components of the intervention, but the authors did recognize that CHEERS could be more specific for certain situations and welcomed opportunities to create additional reporting guidance. An extension to CHEERS covering AI specific items could improve the reporting, transparency and ultimately decision making for AI interventions. This could also help mitigate the paradox of poor reporting inhibiting adoption of AI interventions.

The system-wide need and motivation for improving best practice around data collection and transparency for AI health interventions is evident. Extensions for AI technologies have already been developed for other checklists. CONSORT-AI (Liu et al., 2020) contains AI-specific items for the reporting of RCTs, and it was done in collaboration with the SPIRIT-AI extension for trial protocols (Rivera et al., 2020). Including AI-specific items in the reporting of HEEs may be a logical step to contribute to this standard setting and help to ensure that all relevant information is available to decision makers.

4.1 Limitations

This study has some limitations. We updated the Voets et al. systematic literature review, but searched different databases. It is possible there may have been relevant studies within our search window that we missed by not searching the same databases; however, we believe the databases we searched should give at least equivalent, and probably superior, sensitivity to the original review. Indeed, the sensitivity of our search strategy is

evidenced by the large number of studies excluded at primary screening (2,993) relative to the total number of unique records (3,033). The sensitivity of HEE search filters is well known (Hubbard et al., 2022). While this means our review is highly likely to have identified all relevant published studies, it does mean further updates may be labor intensive with lots of records to screen to identify a relatively small number of relevant studies.

Our review specifically focused on economic evaluations and whilst out of scope, some studies, such as those only reporting patient reported outcome measures, may have been of interest to readers. Additionally, a potential limitation is that our search only covered the period from 1 April 2021 to 17 June 2022. This relatively short search period remains informative due to the rapid advent of AI in healthcare, but it also means that it is likely that relevant economic evaluations have been published since our review.

Another limitation relates to the subjective nature of the NICE quality appraisal checklist. Although the checklist allowed for a further level of analysis regarding the quality of the economic evaluation, it should be used as a broad interpretation rather than a critique of any given study. Despite negating any potential bias by having 2 reviewers, it is possible that different reviewers may have implemented the checklist differently and produced different results. Additionally, other, similar checklists exist (Philips et al., 2004; Drummond, 2015; Adarkwah et al., 2016), and although they broadly serve a similar purpose of understanding the methodological limitations of HEEs, they may have resulted in different or more nuanced quality assessments.

5 Conclusion

This updated review, while covering just a 15-month window, found more economic evaluations of AI health interventions since the last comprehensive systematic literature review which covered the preceding 5 years. Many of the included studies were model-based evaluations and the most common AI intervention was automated image analysis used for screening or diagnosis in the areas of general medicine and oncology. Most evaluations reported the cost per QALY gained.

Overall, the reporting of the studies exhibited limitations. Only a small number of studies were judged to have just minor limitations, according to application of the NICE quality assessment checklist. The majority had potentially serious or very serious limitations resulting from conflicts between research funding and authorship, uncertainty in input data changing the outcome of the evaluation, and lack of transparent reporting of key elements, such as the cost of the technology and how it will be implemented into clinical practice. Specific reporting standards for the economic evaluation of AI interventions would help to improve transparency, reproducibility and trust, and promote their usefulness for decision making. This is fundamental for implementation and coverage decisions which in turn will generate the necessary data to develop flexible models better suited to capture the potentially dynamic nature of the AI intervention.

Data availability statement

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

Author contributions

JV designed and conducted the systematic literature review with support from LA for the database search and management of results. JV drafted the initial manuscript, extracted data and conducted quality assessment. CH extracted data as a second reviewer, conducted quality assessment, and developed the manuscript. JE provided comments and feedback throughout JV's project and on the manuscript development. DD oversaw the work. All authors contributed to the article and approved the submitted version.

Funding

CH, JE, and DD are funded through the HTx project. The HTx project has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement no. 825162. This dissemination reflects only the authors' view and the Commission is not responsible for any use that may be made of the information it contains.

Acknowledgments

The authors would like to thank Sarosh Nagar for his participation as an independent screener.

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.

Supplementary material

The Supplementary Material for this article can be found online at: https://www.frontiersin.org/articles/10.3389/fphar.2023.1220950/full#supplementary-material

References

Adams, S. J., Mondal, P., Penz, E., Tyan, C. C., Lim, H., and Babyn, P. (2021). Development and cost analysis of a lung nodule management strategy combining artificial intelligence and lung-RADS for baseline lung cancer screening. *J. Am. Coll. Radiology* 18 (5), 741–751. doi:10.1016/j.jacr.2020.11.014

Adarkwah, C. C., van Gils, P. F., Hiligsmann, M., and Evers, S. M. A. A. (2016). Risk of bias in model-based economic evaluations: The ECOBIAS checklist. *Expert Rev. Pharmacoeconomics Outcomes Res.* 16 (4), 513–523. doi:10.1586/14737167.2015.1103185

Areia, M., Mori, Y., Correale, L., Repici, A., Bretthauer, M., Sharma, P., et al. (2022). Cost-effectiveness of artificial intelligence for screening colonoscopy: A modelling study. *Lancet Digital Health* 4 (6), 436–444. doi:10.1016/S2589-7500(22)00042-5

Dall, T. M., Gallo, P. D., Chakrabarti, R., West, T., Semilla, A. P., and Storm, M. V. (2013). An aging population and growing disease burden will require ALarge and specialized health care workforce by 2025. *Health Aff.* 32 (11), 2013–2020. doi:10.1377/hlthaff.2013.0714

Davis, S., Stevenson, M., Tappenden, P., and Wailoo, A. (2014) Nice dsu technical support document 15: Cost-effectiveness modelling using patient-level simulation.

de Vos, J., Visser, L. A., de Beer, A. A., Fornasa, M., Thoral, P. J., Elbers, P. W. G., et al. (2022). The potential cost-effectiveness of a machine learning tool that can prevent untimely intensive care unit discharge. *Value Health* 25 (3), 359–367. doi:10.1016/j.jval.2021.06.018

Delgadillo, J., Ali, S., Fleck, K., Agnew, C., Southgate, A., Parkhouse, L., et al. (2022). Stratified care vs stepped care for depression. A cluster randomized clinical trial. *JAMA Psychiatry* 79 (2), 101–108. doi:10.1001/jamapsychiatry.2021.3539

Drummond, M. (2015). *Methods for the economic evaluation of health care programmes*. Fourth: Oxford University Press.

Elvidge, J., Summerfield, A., Nicholls, D., and Dawoud, D. (2022). Diagnostics and treatments of COVID-19: A living systematic review of economic evaluations. *Value Health* 25 (5), 773–784. doi:10.1016/j.jval.2022.01.001

Ericson, O., Hjelmgren, J., Sjövall, F., Söderberg, J., and Persson, I. (2022). The potential cost and cost-effectiveness impact of using a machine learning algorithm for early detection of sepsis in intensive care units in Sweden. *J. Health Econ. Outcomes Res.* 9 (1), 101–110. doi:10.36469/jheor.2022.33951

Fusfeld, L., Menon, S., Gupta, G., Lawrence, C., Masud, S. F., and Goss, T. F. (2022). US payer budget impact of a microarray assay with machine learning to evaluate kidney transplant rejection in for-cause biopsies. *J. Med. Econ.* 25 (1), 515–523. doi:10.1080/13696998.2022.2059221

Gunasekeran, D. V., Tseng, R. M. W. W., Tham, Y. C., and Wong, T. Y. (2021). Applications of digital health for public health responses to COVID-19: A systematic scoping review of artificial intelligence, telehealth and related technologies. *NPJ Digit. Med.* 4 (1), 40–41. doi:10.1038/s41746-021-00412-9

Guo, C., Ashrafian, H., Ghafur, S., Fontana, G., Gardner, C., and Prime, M. (2020). Challenges for the evaluation of digital health solutions—a call for innovative evidence generation approaches. *NPJ Digit. Med.* 3 (1), 110–114. doi:10.1038/s41746-020-00314-2

Huang, X.-M., Yang, B. F., Zheng, W. L., Liu, Q., Xiao, F., Ouyang, P. W., et al. (2022). Cost-effectiveness of artificial intelligence screening for diabetic retinopathy in rural China. *BMC Health Serv. Res.* 22 (260), 260. doi:10.1186/s12913-022-07655-6

Hubbard, W., Walsh, N., Hudson, T., Heath, A., Dietz, J., and Rogers, G. (2022). Development and validation of paired MEDLINE and Embase search filters for cost-utility studies. *BMC Med. Res. Methodol.* 22 (1), 310–319. doi:10.1186/s12874-022-01796-2

Husereau, D., Drummond, M., Augustovski, F., de Bekker-Grob, E., Briggs, A. H., Carswell, C., et al. (2022). Consolidated health economic evaluation reporting standards 2022 (CHEERS 2022) statement: Updated reporting guidance for health economic evaluations. *BMJ* 376 (Cheers), e067975–e067977. doi:10.1136/bmj-2021-067975

Kessler, S., Desai, M., McConnell, W., Jai, E. M., Mebine, P., Nguyen, J., et al. (2021). Economic and utilization outcomes of medication management at a large medicaid plan with disease management pharmacists using a novel artificial intelligence platform from 2018 to 2019: A retrospective observational study using regression methods. *J. Manag. Care Specialty Pharm.* 27 (9), 1186–1196. doi:10.18553/jmcp.2021.21036

Liu, X., Cruz Rivera, S., Moher, D., Calvert, M. J., and Denniston, A. K.SPIRIT-AI and CONSORT-AI Working Group (2020). Reporting guidelines for clinical trial reports for interventions involving artificial intelligence: The CONSORT-AI extension. *Nat. Med.* 26 (9), e537–e548. doi:10.1016/S2589-7500(20)30218-1

MacPherson, P., Webb, E. L., Kamchedzera, W., Joekes, E., Mjoli, G., Lalloo, D. G., et al. (2021). Computer-aided X-ray screening for tuberculosis and HIV testing among adults with cough in Malawi (the prospect study): A randomised trial and cost-effectiveness analysis. *PLoS Med.* 18 (9), 10037522–e1003817. doi:10.1371/journal.pmed.1003752

Mallow, P. J., and Belk, K. W. (2021). Cost-utility analysis of single nucleotide polymorphism panel-based machine learning algorithm to predict risk of opioid use disorder. *J. Comp. Eff. Res.* 10 (18), 1349–1361. doi:10.2217/cer-2021-0115

Miller, M. (2021). FDA publishes approved list of AI/ML-enabled medical devices, IQVIA blog. Available at: https://www.iqvia.com/locations/united-states/blogs/2021/10/fda-publishes-approved-list-of-ai-ml-enabled-medical-devices (Accessed: May 9, 2023).

Mital, S., and Nguyen, H. V ((2022). Cost-effectiveness of using artificial intelligence versus polygenic risk score to guide breast cancer screening. *BMC Cancer* 22 (1), 501. doi:10.1186/s12885-022-09613-1

Morrison, S. L., Dukhovny, D., Chan, R. V. P., Chiang, M. F., and Campbell, J. P. (2022). Cost-effectiveness of artificial intelligence-based retinopathy of prematurity screening. *JAMA Ophthalmol.* 140 (4), 401–409. doi:10.1001/jamaophthalmol.2022.0223

National Institute for Health and Care Excellence (2012). *Methods for the development of NICE public health guidance*.Appendix I quality appraisal checklist-economic evaluations

National Institute for Health and Care Excellence (2022). *Evidence standards framework (ESF) for digital health technologies*. National Institute for Health and Care Excellence. Available at: https://www.nice.org.uk/corporate/ecd7 (Accessed: May 9, 2023).

Nsengiyumva, N. P., Hussain, H., Oxlade, O., Majidulla, A., Nazish, A., Khan, A. J., et al. (2021). Triage of persons with tuberculosis symptoms using artificial intelligence-based chest radiograph interpretation: A cost-effectiveness analysis. *Open Forum Infect. Dis.* 8 (12), 567. doi:10.1093/ofid/ofab567

Panch, T., Szolovits, P., and Atun, R. (2018). Artificial intelligence, machine learning and health systems. *J. Glob. Health* 8 (2), 020303–020308. doi:10.7189/jogh.08.020303

Philips, Z., Ginnelly, L., Sculpher, M., Claxton, K., Golder, S., Riemsma, R., et al. (2004). Review of guidelines for good practice in decision-analytic modelling in health technology assessment. *Health Technol. Assess.* 8 (36)–iv, ix-xi, 1-158. doi:10.3310/hta8360

Ramlal, A., Ahmad, S., Kumar, L., Khan, F., and Chongtham, R. (2021). "From molecules to patients: The clinicalapplications of biological databases andelectronic health records," in *Translational bioinformatics in healthcare and medicine*. Editors K. Raza and N. Dey (First Edit. Academic Press), 107–125. doi:10.1016/B978-0-323-89824-9.00009-4

Rivera, S. C., Liu, X., Chan, A. W., Denniston, A. K., and Calvert, M. J.SPIRIT-AI and CONSORT-AI Working Group (2020). Guidelines for clinical trial protocols for interventions involving artificial intelligence: The SPIRIT-AI Extension. *BMJ* 370, m3210–m3214. doi:10.1136/bmj.m3210

Rosenthal, M. B., and Dudley, R. A. (2007). Pay-for-Performance. *JAMA* 297 (7), 740–744. doi:10.1001/jama.297.7.740

Rudmik, L., and Drummond, M. (2013). Health economic evaluation: Important principles and methodology. *Laryngoscope* 123 (6), 1341–1347. doi:10.1002/lary.23943

Salcedo, J., Rosales, M., Kim, J. S., Nuno, D., Suen, S. C., and Chang, A. H. (2021). Cost-effectiveness of artificial intelligence monitoring for active tuberculosis treatment: A modeling study. *PloS one* 16 (7), e0254950. doi:10.1371/journal.pone.0254950

Schwendicke, F., Mertens, S., Cantu, A. G., Chaurasia, A., Meyer-Lueckel, H., and Krois, J. (2022). Cost-effectiveness of AI for caries detection: Randomized trial. *J. Dent.* 119, 104080. doi:10.1016/j.jdent.2022.104080

Szymanski, T., Ashton, R., Sekelj, S., Petrungaro, B., Pollock, K. G., Sandler, B., et al. (2022). Budget impact analysis of a machine learning algorithm to predict high risk of atrial fibrillation among primary care patients. *Eur. Eur. pacing, Arrhythm. cardiac Electrophysiol. J. Work. groups cardiac pacing, Arrhythm. cardiac Cell. Electrophysiol. Eur. Soc. Cardiol.* 24 (8), 1240–1247. doi:10.1093/europace/euac016

Thokala, P., Srivastava, T., Smith, R., Ren, S., Whittington, M. D., Elvidge, J., et al. (2023). Living health technology assessment: Issues, challenges and opportunities. *PharmacoEconomics* 41 (3), 227–237. doi:10.1007/s40273-022-01229-4

Tseng, A. S., Thao, V., Borah, B. J., Attia, I. Z., Medina Inojosa, J., Kapa, S., et al. (2021). Cost effectiveness of an electrocardiographic deep learning algorithm to detect asymptomatic Left ventricular dysfunction. *Mayo Clin. Proc.* 96 (7), 1835–1844. doi:10.1016/j.mayocp.2020.11.032

Turino, C., Benítez, I. D., Rafael-Palou, X., Mayoral, A., Lopera, A., Pascual, L., et al. (2021). Management and treatment of patients with obstructive sleep apnea using an intelligent monitoring system based on machine learning aiming to improve continuous positive airway pressure treatment compliance: Randomized controlled trial. *J. Med. Internet Res.* 23 (10), 240722–e24112. doi:10.2196/24072

Unsworth, H., Dillon, B., Collinson, L., Powell, H., Salmon, M., Oladapo, T., et al. (2021). The NICE Evidence Standards Framework for digital health and care technologies – developing and maintaining an innovative evidence framework with global impact. *Digit. Health* 7, 20552076211018617–20552076211018620. doi:10.1177/20552076211018617

van Leeuwen, K. G., Meijer, F. J. A., Schalekamp, S., Rutten, M. J. C. M., van Dijk, E. J., van Ginneken, B., et al. (2021). Cost-effectiveness of artificial intelligence aided vessel occlusion detection in acute stroke: An early health technology assessment. *Insights into Imaging* 12 (133), 133. doi:10.1186/s13244-021-01077-4

Voets, M. M., Veltman, J., Slump, C. H., Siesling, S., and Koffijberg, H. (2022). Systematic review of health economic evaluations focused on artificial intelligence in healthcare: The tortoise and the cheetah. *Value Health* 25 (3), 340–349. doi:10.1016/j.jval.2021.11.1362

Xiao, X., Xue, L., Ye, L., Li, H., and He, Y. (2021). Health care cost and benefits of artificial intelligence-assisted population-based glaucoma screening for the elderly in remote areas of China: A cost-offset analysis. *BMC Public Health* 21 (1), 1065–1112. doi:10.1186/s12889-021-11097-w

Ziegelmayer, S., Graf, M., Makowski, M., Gawlitza, J., and Gassert, F. (2022). Cost-effectiveness of artificial intelligence support in computed tomography-based lung cancer screening. *Cancers* 14 (7), 1729. doi:10.3390/cancers14071729

frontiers | Frontiers in Pharmacology

TYPE Perspective
PUBLISHED 24 August 2023
DOI 10.3389/fphar.2023.1234414

Check for updates

OPEN ACCESS

EDITED BY
Dalia M. Dawoud,
National Institute for Health and Care Excellence, United Kingdom

REVIEWED BY
Milou Hogervorst,
Utrecht University, Netherlands

*CORRESPONDENCE
Sarah Norris,
sarah.norris@sydney.edu.au

RECEIVED 04 June 2023
ACCEPTED 31 July 2023
PUBLISHED 24 August 2023

CITATION
Cheyne S, Chakraborty S, Lewis S, Campbell S, Turner T and Norris S (2023), What could health technology assessment learn from living clinical practice guidelines?
Front. Pharmacol. 14:1234414.
doi: 10.3389/fphar.2023.1234414

COPYRIGHT
© 2023 Cheyne, Chakraborty, Lewis, Campbell, Turner and Norris. This is an open-access article distributed under the terms of the Creative Commons Attribution License (CC BY). The use, distribution or reproduction in other forums is permitted, provided the original author(s) and the copyright owner(s) are credited and that the original publication in this journal is cited, in accordance with accepted academic practice. No use, distribution or reproduction is permitted which does not comply with these terms.

What could health technology assessment learn from living clinical practice guidelines?

Saskia Cheyne[1,2], Samantha Chakraborty[2], Samara Lewis[3], Sue Campbell[3], Tari Turner[2] and Sarah Norris[1,3]*

[1]Faculty of Medicine and Health, University of Sydney, Sydney, NSW, Australia, [2]Australian Living Evidence Consortium, Cochrane Australia, Monash University, Melbourne, VIC, Australia, [3]Hereco, Sydney, NSW, Australia

A "living" approach to clinical practice guidelines is when the identification, appraisal and synthesis of evidence is maintained and repeated at an agreed frequency, with a clear process for when and how new evidence is to be incorporated. The value of a living approach to guidelines was emphasised during the COVID-19 pandemic when health professionals and policymakers needed to make decisions regarding patient care in the context of a nascent but rapidly evolving evidence base. In this perspective, we draw on our recent experience developing Australian and international living guidelines and reflect on the feasibility of applying living guideline methods and processes to a lifecycle approach to health technology assessment (HTA). We believe the opportunities and challenges of adopting a living approach in HTA fall into five key themes: identification, appraisal and synthesis of evidence; optimising the frequency of updates; embedding ongoing multi-stakeholder engagement; linking the emergence of new evidence to reimbursement; and system capacity to support a living approach. We acknowledge that the suitability of specific living approaches to HTA will be heavily influenced by the type of health technology, its intended use in the health system, local reimbursement pathways, and other policy settings. But we believe that the methods and processes applied successfully to guideline development to manage evidentiary uncertainty could be applied in the context of HTA and reimbursement decision-making to help manage similar sources of uncertainty.

KEYWORDS

clinical practical guidelines, health technolgy assessment, living systematic review (LSR), regulatory policies and structures, reimbursement pathways, lifecycle HTA

1 Introduction

Health Technology Assessment (HTA) is a multidisciplinary process that uses explicit methods to determine the value of a health technology at different points in its lifecycle, for the purpose of informing decision-making that promotes an equitable, efficient, and high-quality health system (O'Rourke et al., 2020) It is a formal, systematic process for translating evidence into health policy. A full HTA typically includes the following domains: a description of the health problem and its current standard of care; a description of the proposed health technology or service; the comparative safety and effectiveness of the proposed health technology or service (with these elements typically framed using the PICO criteria—Population, Intervention, Comparator, Outcomes); an economic evaluation; a

budget impact analysis; consideration of relevant organisational or implementation aspects; and consideration of relevant ethical, legal, and social aspects (EUnetHTA, 2016).

HTA is often reactive, occurring at a single point in time following initial regulatory approval or in response to regulatory changes (e.g., the expansion of approved indications) (CADTH, 2011; PBS Scheme, 2022). Full HTAs can take several months to years to complete. A lifecycle approach to HTA, whereby evidence is frequently incorporated and the HTA is dynamically updated, was first proposed in 2016 in order to more fully realise the benefits of innovations in healthcare (Husereau et al., 2016; Grammati et al., 2023). Since then a number of initiatives around the world have been exploring how a lifecycle approach to HTA can be implemented, for example, reassessments are performed by HAS and NICE, and conditional approvals exist in multiple countries such as the United Kingdom, the Netherlands and France (Ibargoyen-Roteta et al., 2022).

A lifecycle approach is even more relevant as agencies around the world are faced with assessing new, rapidl evolving classes of health technology, such as cell and gene therapies (Husereau et al., 2016). In this article, we share our recent experience developing and implementing methods and processes for Australian and international living guidelines and reflect on the opportunities and challenges of applying a living guideline approach to lifecycle HTA (Cheyne et al., 2023a).

2 Static *versus* living guidelines

The core methods for literature searching, evidence appraisal and synthesis are similar for living and partial updating of traditional (static) guidelines, but living guidelines involve a frequent and explicit approach to keeping the guidelines up-to-date. This approach includes frequent surveillance for newly published clinical studies, the prospective, ongoing incorporation of those studies into the evidence base, and the use of pre-agreed triggers for updating the corresponding evidence-based recommendations (Akl et al., 2017; Cheyne et al., 2023a; Cheyne et al., 2023b; Fraile Navarro et al., 2023; McDonald et al., 2023; Synnot et al., 2023) The criteria for selecting living topics are: clinical or policy priority of the question, important uncertainty in the existing evidence, and high likelihood of emergence of new evidence where the clinical/policy context is likely to change (Akl et al., 2017; Cheyne et al., 2023a). The frequency of updating a living topic is determined by the nature of the health problem, the flow of emerging evidence, the capacity of the evidence review team to search, screen and appraise new evidence, and the capacity of the Guideline Development Panel to meet and determine the implications of the new evidence (Cheyne et al., 2023b; McDonald et al., 2023) For example, searches for living COVID-19 guidelines were conducted on a daily basis during the height of the pandemic, whereas searches for living stroke guidelines are conducted every 3 months (Tendal et al., 2021; Hill et al., 2022).

The most tangible benefit of a living approach to guidelines is that evidence-based recommendations for clinical care retain their trustworthiness by remaining up-to-date. A less tangible (but no less important) benefit of a living approach is the way it changes the context for decision making: the knowledge that a decision can be revisited soon (typically in weeks or months) means that Guideline Development Panel members are more likely to make a decision on a recommendation in the face of uncertain evidence, rather than make no decision.

3 Similarities and differences between HTA and guidelines

Though intended for different purposes and audiences, HTA and clinical practice guidelines share core components, particularly those related to methods for the surveillance, appraisal, synthesis, and contextualisation of clinical and patient evidence (Guyatt et al., 2011; Higgins et al., 2022). Best practice in HTA and guideline development places an emphasis on early and ongoing multi-stakeholder involvement (Ibargoyen-Roteta et al., 2022). HTA and guideline development both rely on deliberative processes to translate evidence into recommendations for policy and practice.

However, there are important differences between HTA and guidelines. These differences arise from the fact that HTA has a broader scope than guidelines, is undertaken by industry as well as by government and non-profit organisations, is less transparent because of the inclusion of unpublished clinical data and commercially sensitive pricing information, and needs to comply with local regulatory and reimbursement pathways. This means that it is more straight-forward to change a guideline recommendation than it is to change an HTA decision. It also means that it cannot be assumed that the methods and processes applied in living guidelines are directly transferable to all HTA in all settings.

Despite the differences, HTA and guideline development are interdependent activities that draw from the same knowledge base: HTA often relies on guidelines to define current treatment pathways and comparators; and guidelines need to be cognisant of the regulatory and reimbursement status of treatments they recommend. The need for harmonisation of HTA and guidelines (e.g., as undertaken by NICE in the United Kingdom) is an important area of health services research and has been described by others, but is not the focus of the current article (Schünemann et al., 2022). Early multi-stakeholder dialogue frameworks allow for health technology developers to incorporate advice from HTA agencies in their health technology planning and to directly address uncertainty during technology development (Ibargoyen-Roteta et al., 2022; Hogervorst et al., 2023).

4 Opportunities and challenges in adopting a living guideline approach for HTA

We see a number of opportunities and challenges for adopting a living guideline approach in HTA (Table 1). The living guideline approaches most obviously suited to HTA relate to the methods of evidence assessment. The tools to support standard and living systematic reviews are advancing rapidly, and the potential for these to be incorporated within HTA methods have been described by others (Grammati et al., 2023; Thokala et al., 2023). To date, most evidence review within living guidelines has been limited to randomised controlled trials (RCTs) of interventions. By contrast, HTAs often include diagnostic, prognostic, economic and epidemiological questions, in addition to intervention questions, and the inclusion of non-randomised controlled data such as longer term safety evidence from observational studies or registry data. HTA is now often reliant on single-arm trials and "Real World Evidence" and a number of organisations are exploring the use of such

TABLE 1 Opportunities and challenges for adopting a living guideline approach for HTA.

Opportunities	Challenges
1. Evidence identification, appraisal and synthesis	
• Preparing clinical evidence syntheses in standardised and shareable formats to minimise duplication of effort across agencies. (e.g., the use of GRADE and MAGIC for living guidelines has enabled the sharing of Evidence Profile tables between countries)	• How and when to include unpublished clinical evidence • How to include evidence for diagnostic, prognostic, economic, and epidemiological questions • How to store data securely whilst enabling sharing • Copyright restrictions around data extracted from published evidence
2. Optimising the frequency of updates	
• More frequent updates of the evidence could resolve uncertainty regarding the technology, care pathways, patient group, uptake, market share, or economic modelling, especially where conventional evidentiary standards have not been met.	• Reimbursement and procurement systems may not be designed for frequent changes in pricing for a health technology
3. Embedding multi-stakeholder engagement	
• Early identification and ongoing dialogue with all relevant stakeholders (as occurs with a living Guideline Development Panel) would support planning and scoping for HTA	• How to facilitate effective engagement and communication between stakeholders with different perspectives or priorities (payers/government, industry, regulatory bodies, healthcare providers, healthcare professionals, patients) • How to share commercially sensitive information amongst this wider group of stakeholders
4. Linking the emergence of new evidence to reimbursement	
• Re-evaluation and value-based renegotiation in response to new evidence (especially where conditional funding decisions have been made) • Decision-makers may be more inclined to provide conditional reimbursement for technologies if they are confident that decisions can be reversed if no definitive evidence of effectiveness emerges	• Pricing negotiation and/or the implementation of new pricing agreements can be protracted and may negate any reductions in time to market access • The framework for renegotiation of pricing needs to allow for price increases as well as price decreases or disinvestment (either complete de-adoption of technologies that are not clinically effective or restrictions to ensure cost-effective use)
5. System capacity to support a living approach	
• More certainty in the timing and scope of HTA which enables better workforce planning for those undertaking the HTA.	• Fixed schedules for reimbursement decision-making • Regulatory or legislative changes may be required to compel technology developers to provide the required data • Having sufficient methodological capacity on hand to ensure the timely inclusion of new evidence as it emerges

data in HTA (HAS Sante, 2021; NICE, 2022; Bakker et al., 2023). It should be feasible, though, for a living approach to be adopted across all types of evidence searching that occur within an HTA. For example, living guidelines for COVID-19 diagnostics for antigen, serology and molecular testing (Hanson et al., 2021) and living systematic reviews are frequently conducted on these types of questions (Wynants et al., 2020).

Similar and additional factors are likely to determine the frequency with which HTA literature searches can be updated, including a combination of the capacity of HTA teams to undertake more frequent searching, and the frequency with which the respective decision-making entities can meet to adjudicate on the new evidence. One issue to be mindful of is that the frequency of updating decisions does not outpace the ability of the health system to respond. The frequency of guideline recommendation revisions is effectively limited by the ability of healthcare providers to modify local protocols and standards for care. However, the frequency of reimbursement revisions will be limited by the frequency at which decision-makers can consider updates, and the frequency with which pricing and supply contracts between industry and payers can be varied.

As is for guidelines, it is unlikely that all HTA questions will be suitable for a living approach. Given the organisational changes that would be required to support a living approach to HTA, agencies may wish to focus on technologies that promise a high benefit-to-risk ratio, where the usual levels of RCT evidence are not available and where the cost implications are significant (e.g., cell and gene therapies), or where the pace of technological innovation is very high (e.g., digital health technologies) or the policy context is changing rapidly (e.g., the use of AI in diagnostics). In these situations, it should be possible to adopt a concept known as early multi-stakeholder dialogue, which is a prospective or intentional approach to HTA where manufacturers, healthcare providers, clinicians and payers pre-agree i) the measures of most relevance for the technology and the population(s) of interest, and ii) how the pricing of the technology will vary based on those measures (Schünemann et al., 2022). Any non-RCT data informing the decision-making will need to be considered trustworthy by HTA agencies and the payer (NICE, 2022). An illustration of this approach is shown in Figure 1.

Conditional marketing authorization pathways or lifecycle approaches to HTA have been introduced for cancer drugs and for digital health technologies (Sabry-Grant et al., 2019). These pathways incorporate some elements of a living approach by allowing the flexibility to provide temporary access to treatments in limited circumstances as more evidence accumulates (Hoekman et al., 2015; Regier et al., 2022). The use of a living approach here may provide the necessary flexibility in a more robust way, with an intention from the outset to continue updating the HTA with new evidence until a higher degree of certainty is reached, or to revise or rescind an access decision if reliable evidence of a net positive effect is not eventually obtained.

A living approach to HTA could decrease research waste and duplication of effort. The sharing of evidence summaries already happens in clinical practice guidelines (NICE, 2021), and there are

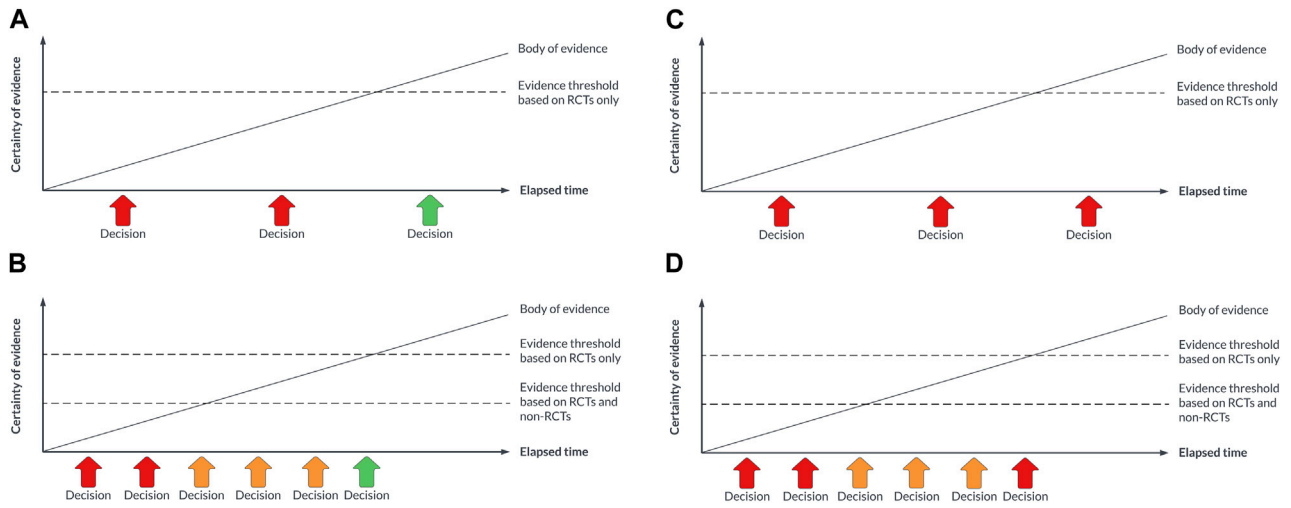

FIGURE 1
(A) Decision-making with a traditional approach to HTA for a technology that is eventually demonstrated to be safe, effective, and cost-effective. **(B)** Decision-making with a living approach to HTA for a technology that is eventually demonstrated to be safe, effective, and cost-effective. **(C)** Decision-making with a traditional approach to HTA for a technology that is eventually not demonstrated to be safe, effective, and cost-effective. **(D)** Decision-making with a living approach to HTA for a technology that is eventually demonstrated to not be safe, effective, and cost-effective.

steps towards this happening between HTA agencies in Canada, Europe, and Australia, (PBS, 2022; Hogervorst et al., 2023), however in reality the confidentiality of pricing arrangements, and the potential for price lowering or disinvestment at future reassessments, will limit the extent of such sharing (Thokala et al., 2023).

5 Discussion

In this perspective we have discussed the aspects of HTA that are most amenable to a living approach and where living guideline evidence translation methods or processes can be transferred to HTA. We also highlight what, in our view, is needed to support a transition to living HTA (Box 1).

Box 1 What is needed to optimise the impact and reliability of a living approach to HTA.

1. Development and/or testing of methods for the continuous updating of non-RCT evidence.
2. Pilot studies for different technologies for different clinical purposes, to understand what works, what does not work, and why, and the importance of context (i.e., the local health system, approaches to HTA, and health system financing).
3. Agreement on the HTA scenarios where a living approach is likely to optimise market access, defined as a combination of shorter time to market, with acceptable mitigation of safety risk to patients, and acceptable cost and cost-effectiveness.
4. Agreement on the policy levers that will be required to support partial or full disinvestment if technologies do not live up to their promise.
5. Practical guidance on the organisational and resourcing requirements for living HTA, and how to transition from reactive HTA at a single point in time to responsive HTA throughout the life-cycle of a technology.

The iterative nature of a living process allows for more nuance in the face of uncertainty, and a willingness to support innovation at early stages, knowing that decisions will be revisited and revised as new evidence emerges. It could give decision-makers comfort in making early conditional decisions for a technology/service, instead of what might otherwise be a "no" decision in the face of uncertainty. The "secret sauce" of a living guideline approach is the organisational infrastructure and collaborative culture that needs to be put in place to support it. It requires a commitment on the part of the guideline developer to provide ongoing funding to resource continuous evidence review activities, and a standing Guideline Development Panel to deliberate on new evidence as and when it emerges. Although a lot of HTA activity is undertaken as 'one off' evidence reviews, it should be possible for industry and HTA agencies to re-orient some (if not all) of their resources to a framework that supports the ongoing incorporation of new data (e.g., from health administrative systems or clinical quality registries). There is also additional efficiency to be gained by aligning the methods and timing for living guidelines and lifecycle HTA.

HTA agencies are under increased pressure to provide patients with early access to promising health technologies, while accounting for the often-incomplete picture of clinical and economic impact of a new treatment during its initial technology assessment. Often, the evidence available at the time of the first HTA is limited, and decision uncertainty may be reduced with longer term data from trials, observational and registry data. At the level of evidence review methods, further innovation and testing of living methods is required for study designs other than RCTs and for non-intervention questions, particularly given the drive for HTA to rely more on innovative clinical trial designs (e.g., platform and adaptive trials). Living HTA could expand the approaches employed by living guidelines in two key

ways: By 1) including pricing/cost considerations in the prioritisation criteria for living topics, and 2) exploring how living searches for economic and epidemiological data could feed in to economic evaluations and budget impact analyses. The policy challenges of adopting a living approach in HTA are more significant than for a living approach to guidelines: the benefits of earlier patient access to treatments need to be balanced against the potential for making "wrong" decisions—reimbursing technologies that do not end up being as safe, effective and/or cost-effective as anticipated. This highlights the importance of developing trust between stakeholders *before* living approaches are implemented, and finding a balance between policy levers that "push" (e.g., requiring developers to provide data on their technology) and "pull" (e.g., earlier market access) towards a living approach.

The introduction of the living approach may result in the ability to create a more harmonious and streamlined process between both HTA and guidelines. In this perspective we have illustrated the HTA domains where living guideline evidence translation methods or processes are directly transferable to HTA, additional aspects of HTA where a living approach is likely to be suitable (but where methods and processes still need to be developed); and aspects of HTA that are unlikely to be suitable for a living approach. However, our experience is limited by primarily conducting living guidelines and HTAs in an Australian context. Pilot case studies are needed that 1) describe the experience of introducing different living methods or processes within different HTA scenarios, 2) determine benefits and challenges of these approaches, 3) further develop methods for those areas of living methods that are specific to HTA, such as economic analysis, and 4) place these experiences within the local policy context so that broader themes can be identified regarding the suitability of living methods and processes for HTA in different countries.

Data availability statement

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

Author contributions

SC: Conceptualization, writing—original draft. SC: Writing—review and editing. SL: Writing—review and editing. SC: Writing—review and editing. TT: Writing—review and editing, supervision. SN: Conceptualization, writing—original draft, supervision. All authors contributed to the article and approved the submitted version.

Funding

In the last five years the following authors have received funding to develop the following clinical practice guidelines (with funding sources): Cheyne, Chakraborty, Norris, Campbell, Turner for the Australian COVID-19 living guidelines (Australian Living Evidence Consortium for the development of the COVID-19 Guidelines, Walter Cottman Endowment Fund, managed by Equity Trustees for the development of the MPX Guidelines, Australian Government DHAC, Victorian Government Department of Health and Human Services, The Ian Potter Foundation, Walter Cottman Endowment Fund, managed by Equity Trustees, Lord Mayors' Charitable Foundation); Turner for the MPX guidelines (Walter Cottman Endowment Fund, managed by Equity Trustees), Turner for guidelines supported by the Australian Living Evidence Consortium (Australian Government Department of Health and Aged Care, Victorian Government Department of Health and Human Services, The Ian Potter Foundation, Lord Mayors' Charitable Foundation, Gandel Foundation); Campbell, Norris for the Australian Perinatal Mental Health Guidelines (Australian Government Department of Health and Aged Care); Campbell, Lewis, Norris for the Australian endometriosis guidelines (Australian government Department of Health and Aged Care); Campbell for the mild traumatic brain injury guidelines (Australian Government research grant, MRF2008070); Chakraborty for the Clinical Guidelines for the Diagnosis and Management of Work-related Mental Health Guidelines in General Practice (Australian Government Department of Jobs and Small Business and Comcare, Office of Industrial Relations – Queensland Government, State Insurance Regulatory Authority, ReturntoWorkSA and WorkCover WA).

Conflicts 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.

References

Akl, E. A., Meerpohl, J. J., Elliott, J., Kahale, L. A., Schünemann, H. J., Agoritsas, T., et al. (2017). Living systematic reviews: 4. Living guideline recommendations. *J. Clin. Epidemiol.* 91, 47–53. doi:10.1016/j.jclinepi.2017.08.009

Bakker, E., Plueschke, K., Jonker, C. J., Kurz, X., Starokozhko, V., and Mol, P. G. M. (2023). Contribution of real-world evidence in European medicines agency's regulatory decision making. *Clin. Pharmacol. Ther.* 113 (1), 135–151. doi:10.1002/cpt.2766

CADTH (2011). *Optimal use reports*. Ottawa (ON): Canadian Agency for Drugs and Technologies in Health.

Cheyne, S., Fraile Navarro, D., Buttery, A. K., Chakraborty, S., Crane, O., Hill, K., et al. (2023a). Methods for living guidelines: early guidance based on practical experience. Paper 3: selecting and prioritizing questions for living guidelines. *J. Clin. Epidemiol.* 155, 73–83. doi:10.1016/j.jclinepi.2022.12.021

Cheyne, S., Fraile Navarro, D., Hill, K., McDonald, S., Tunnicliffe, D., White, H., et al. (2023b). Methods for living guidelines: early guidance based on practical experience. Paper 1: introduction. *J. Clin. Epidemiol.* 155, 84–96. doi:10.1016/j.jclinepi.2022.12.024

EUnetHTA (2016). "Joint action 2, work package 8," in *HTA core model ® version 3.0*. Finland. Available at: www.htacoremodel.info/BrowseModel.aspx.

Fraile Navarro, D., Cheyne, S., Hill, K., McFarlane, E., Morgan, R. L., Murad, M. H., et al. (2023). Methods for living guidelines: early guidance based on practical experience. Article 5: decisions on methods for evidence synthesis and recommendation development for living guidelines. *J. Clin. Epidemiol.* 155, 118–128. doi:10.1016/j.jclinepi.2022.12.022

Grammati, S., Anna, F., Jamie, E., and Dalia, D. (2023). Living health technology assessments: how close to living reality? *BMJ Evidence-Based Med*. doi:10.1136/bmjebm-2022-112152

Guyatt, G., Oxman, A. D., Akl, E. A., Kunz, R., Vist, G., Brozek, J., et al. (2011). GRADE guidelines: 1. Introduction: GRADE evidence profiles and summary of findings tables. *J. Clin. Epidemiol.* 64 (4), 383–394. doi:10.1016/j.jclinepi.2010.04.026

Hanson, K. E., Altayar, O., Caliendo, A. M., Arias, C. A., Englund, J. A., Hayden, M. K., et al. (2021). Infectious diseases society of America guidelines on the diagnosis of coronavirus disease 2019 (COVID-19): serologic testing. *Clin. Infect. Dis.*, ciaa1343. doi:10.1093/cid/ciaa1343

J. P. T. T. J. Higgins, J. Chandler, M. Cumpston, T. Li, M. J. Page, and V. A. Welch (Editors) (2022). *Cochrane handbook for systematic reviews of interventions* (London, UK: Cochrane).

Hill, K., English, C., Campbell, B. C. V., McDonald, S., Pattuwage, L., Bates, P., et al. (2022). Feasibility of national living guideline methods: the Australian Stroke Guidelines. *J. Clin. Epidemiol.* 142, 184–193. doi:10.1016/j.jclinepi.2021.11.020

Hoekman, J., Boon, W. P., Bouvy, J. C., Ebbers, H. C., de Jong, J. P., and De Bruin, M. L. (2015). Use of the conditional marketing authorization pathway for oncology medicines in Europe. *Clin. Pharmacol. Ther.* 98 (5), 534–541. doi:10.1002/cpt.174

Hogervorst, M., Vreman, R., Heikkinen, I., Bagchi, I., Gutierrez-Ibarluzea, I., Ryll, B., et al. (2023a). Uncertainty management in regulatory and health technology assessment decision-making on drugs: GUIDANCE OF THE HTAi-DIA WORKING GROUP. *Int. J. Technol. Assess. Health Care* 39, e40–e25. doi:10.1017/S0266462323000375

Hogervorst, M. A., Vreman, R. A., Zawada, A., Zielińska, M., Dawoud, D. M., de Jong, B. A., et al. (2023b). Synergy between health technology assessments and clinical guidelines for multiple sclerosis. *Clin. Transl. Sci.* 16 (5), 835–849. doi:10.1111/cts.13492

Husereau, D., Henshall, C., Sampietro-Colom, L., and Thomas, S. (2016). Changing health technology assessment paradigms? *Int. J. Technol. Assess. Health Care* 32 (4), 191–199. doi:10.1017/S0266462316000386

Ibargoyen-Roteta, N., Galnares-Cordero, L., Benguria-Arrate, G., Chacón-Acevedo, K. R., Gutiérrez-Sepulveda, M. P., Low-Padilla, E., et al. (2022). A systematic review of the early dialogue frameworks used within health technology assessment and their actual adoption from HTA agencies. *Front. Public Health* 10, 942230. doi:10.3389/fpubh.2022.942230

McDonald, S., Sharp, S., Morgan, R. L., Murad, M. H., and Fraile Navarro, D.Australian Living Evidence Consortium Methods and Processes Working Group and Collaborators (2023). Methods for living guidelines: early guidance based on practical experience. Paper 4: search methods and approaches for living guidelines. *J. Clin. Epidemiol.* 155, 108–117. doi:10.1016/j.jclinepi.2022.12.023

NICE (2021). The NICE strategy 2021 to 2026. Available from: https://www.nice.org.uk/about/who-we-are/corporate-publications/the-nice-strategy-2021-to-2026.

NICE. NICE (2022). *real-world evidence framework*. Manchester, United Kingdom: National Institute of Health and Care Excellence.

O'Rourke, B., Oortwijn, W., and Schuller, T.International Joint Task Group (2020). The new definition of health technology assessment: a milestone in international collaboration. *Int. J. Technol. Assess. Health Care* 36 (3), 187–190. doi:10.1017/S0266462320000215

PBS (2022). Arrangement between the department of health and aged care and health technology assessment bodies in the United Kingdom and Canada. Available from: https://www.pbs.gov.au/info/news/2022/09/collaboration-arrangement-between-the-department-of-health-and-aged-care.

Regier, D. A., Pollard, S., McPhail, M., Bubela, T., Hanna, T. P., Ho, C., et al. (2022). A perspective on life-cycle health technology assessment and real-world evidence for precision oncology in Canada. *npj Precis. Oncol.* 6 (1), 76. doi:10.1038/s41698-022-00316-1

Sabry-Grant, C., Malottki, K., and Diamantopoulos, A. (2019). The cancer drugs fund in practice and under the new framework. *Pharmacoeconomics* 37 (7), 953–962. doi:10.1007/s40273-019-00793-6

HAS (2021). "Real-world studies for the assessment of medicinal products and medical devices," in *French National Authority for Health*. Editor H. A. D. Sante (Paris, France: Haute Autorité de Santé).

PBS (2022). "*Pharmaceutical benefits Scheme post-market reviews*," in Canberra: Australian government. Editor P. B. Scheme .

Schünemann, H. J., Reinap, M., Piggott, T., Laidmäe, E., Köhler, K., Põld, M., et al. (2022). The ecosystem of health decision making: from fragmentation to synergy. *Lancet Public Health* 7 (4), e378–e390. doi:10.1016/S2468-2667(22)00057-3

Synnot, A., Hill, K., Davey, J., English, K., Whittle, S. L., Buchbinder, R., et al. (2023). Methods for living guidelines: early guidance based on practical experience. Paper 2: consumer engagement in living guidelines. *J. Clin. Epidemiol.* 155, 97–107. doi:10.1016/j.jclinepi.2022.12.020

Tendal, B., Vogel, J. P., McDonald, S., Norris, S., Cumpston, M., White, H., et al. (2021). Weekly updates of national living evidence-based guidelines: methods for the Australian living guidelines for care of people with COVID-19. *J. Clin. Epidemiol.* 131, 11–21. doi:10.1016/j.jclinepi.2020.11.005

Thokala, P., Srivastava, T., Smith, R., Ren, S., Whittington, M. D., Elvidge, J., et al. (2023). Living health technology assessment: issues, challenges and opportunities. *PharmacoEconomics* 41 (3), 227–237. doi:10.1007/s40273-022-01229-4

Wynants, L., Van Calster, B., Collins, G. S., Riley, R. D., Heinze, G., Schuit, E., et al. (2020). Prediction models for diagnosis and prognosis of Covid-19 infection: systematic review and critical appraisal. *BMJ* 369, m1328. doi:10.1136/bmj.m1328

frontiers | Frontiers in Pharmacology

TYPE Hypothesis and Theory
PUBLISHED 15 September 2023
DOI 10.3389/fphar.2023.1180962

Check for updates

OPEN ACCESS

EDITED BY
Sheyu Li,
Sichuan University, China

REVIEWED BY
Koen Degeling,
The University of Melbourne, Australia
Federico Motta,
University of Modena and Reggio Emilia, Italy

*CORRESPONDENCE
Blythe Adamson,
badamson@flatiron.com

RECEIVED 06 March 2023
ACCEPTED 25 August 2023
PUBLISHED 15 September 2023

CITATION
Adamson B, Waskom M, Blarre A, Kelly J, Krismer K, Nemeth S, Gippetti J, Ritten J, Harrison K, Ho G, Linzmayer R, Bansal T, Wilkinson S, Amster G, Estola E, Benedum CM, Fidyk E, Estévez M, Shapiro W and Cohen AB (2023), Approach to machine learning for extraction of real-world data variables from electronic health records. *Front. Pharmacol.* 14:1180962. doi: 10.3389/fphar.2023.1180962

COPYRIGHT
© 2023 Adamson, Waskom, Blarre, Kelly, Krismer, Nemeth, Gippetti, Ritten, Harrison, Ho, Linzmayer, Bansal, Wilkinson, Amster, Estola, Benedum, Fidyk, Estévez, Shapiro and Cohen. This is an open-access article distributed under the terms of the Creative Commons Attribution License (CC BY). The use, distribution or reproduction in other forums is permitted, provided the original author(s) and the copyright owner(s) are credited and that the original publication in this journal is cited, in accordance with accepted academic practice. No use, distribution or reproduction is permitted which does not comply with these terms.

Approach to machine learning for extraction of real-world data variables from electronic health records

Blythe Adamson[1,2]*, Michael Waskom[1], Auriane Blarre[1], Jonathan Kelly[1], Konstantin Krismer[1], Sheila Nemeth[1], James Gippetti[1], John Ritten[1], Katherine Harrison[1], George Ho[1], Robin Linzmayer[1], Tarun Bansal[1], Samuel Wilkinson[1], Guy Amster[1], Evan Estola[1], Corey M. Benedum[1], Erin Fidyk[1], Melissa Estévez[1], Will Shapiro[1] and Aaron B. Cohen[1,3]

[1]Flatiron Health, Inc., New York, NY, United States, [2]The Comparative Health Outcomes, Policy, and Economics (CHOICE) Institute, Department of Pharmacy, University of Washington, Seattle, WA, United States, [3]Department of Medicine, NYU Grossman School of Medicine, New York, NY, United States

Background: As artificial intelligence (AI) continues to advance with breakthroughs in natural language processing (NLP) and machine learning (ML), such as the development of models like OpenAI's ChatGPT, new opportunities are emerging for efficient curation of electronic health records (EHR) into real-world data (RWD) for evidence generation in oncology. Our objective is to describe the research and development of industry methods to promote transparency and explainability.

Methods: We applied NLP with ML techniques to train, validate, and test the extraction of information from unstructured documents (e.g., clinician notes, radiology reports, lab reports, etc.) to output a set of structured variables required for RWD analysis. This research used a nationwide electronic health record (EHR)-derived database. Models were selected based on performance. Variables curated with an approach using ML extraction are those where the value is determined solely based on an ML model (i.e. not confirmed by abstraction), which identifies key information from visit notes and documents. These models do not predict future events or infer missing information.

Results: We developed an approach using NLP and ML for extraction of clinically meaningful information from unstructured EHR documents and found high performance of output variables compared with variables curated by manually abstracted data. These extraction methods resulted in research-ready variables including initial cancer diagnosis with date, advanced/metastatic diagnosis with date, disease stage, histology, smoking status, surgery status with date, biomarker test results with dates, and oral treatments with dates.

Abbreviations: AI, artificial intelligence; BERT, bidirectional encoder representations from transformers; EHR, electronic health records; LSTM, long term short memory; ML, machine learning; NPV, negative predictive value; NSCLC, non–small cell lung cancer; P&Ps, Policies and Procedures; PPV, positive predictive value; RWD, real-world data; RWE, real-world evidence.

Conclusion: NLP and ML enable the extraction of retrospective clinical data in EHR with speed and scalability to help researchers learn from the experience of every person with cancer.

KEYWORDS

electronic health records, cancer, oncology, real-world data, machine learning, natural language processing, artificial intelligence

Introduction

A barrier to generating robust real-world evidence (RWE) is access to research-ready datasets that demonstrate sufficient recency, clinical depth, provenance, completeness, representativeness and usability. Health outcomes must be appropriately defined and consistently measured. For studies using routinely collected electronic health record (EHR)-derived data, a considerable amount of data preprocessing and labor-intensive curation is required to create a dataset with clinically meaningful variables and outcomes needed for analysis (Figure 1).

The challenge is that so much valuable information is trapped within unstructured documents like clinician notes or scanned faxes of lab reports, where extracting the relevant data is far from trivial. The traditional approach to having clinical experts manually review patient charts to abstract data is time consuming and resource intensive (Birnbaum et al., 2020). This approach limits the number of patients available for research purposes. Learnings can quickly become outdated—for example as new biomarkers and treatments emerge, the standards of care change, or new indicators for social determinants of health are prioritized. In other instances, answers to important research questions remain infeasible due to limited sample sizes.

Artificial intelligence (AI) advances in the areas of natural language processing (NLP) and machine learning (ML) have created new opportunities to improve the scale, flexibility, and efficiency of curating high-quality real-world data (RWD) in oncology (Bhardwaj et al., 2017; Bera et al., 2019; Datta et al., 2019; Koleck et al., 2019; Shah et al., 2019; Wang et al., 2019; Bertsimas and Wiberg, 2020; Karimi et al., 2021; Subbiah, 2023). The definitions of foundational AI/ML terminology are provided in Table 1. When using ML and NLP for RWE, current guidance emphasizes transparency (NICE, 2022; Norgeot et al., 2020; Center for Drug Evaluation and Research Center for Biologics Evaluation and Research Oncology Center of Excellence; Padula et al., 2022; Blueprint for trustworthy AI implementation guidance and assurance for healthcare, 2022). The United Kingdom National Institute for Health and Care Excellence instructs that "where human abstraction or artificial intelligence tools are used to construct variables from unstructured data, the methods and processes used should be clearly described." (NICE, 2022).

In response to guidance, the objective of this paper is to describe the general approach for applied NLP and ML methods that are used by Flatiron Health to extract data from unstructured documents stored in oncology care EHR. A key distinction in our terminology is the use of "abstraction" meaning performed by humans and "extraction" meaning performed by models. Out of scope for this paper are other AI, ML, and NLP innovations and contributions from Flatiron Health, such as: model-assisted cohort selection (Birnbaum et al., 2019; Birnbaum et al., 2020); continuous bias monitoring software (Birnbaum et al., 2023); automated mapping of laboratory data (Kelly et al., 2022); prediction of future health events (Chen et al., 2019); and point-of-care products to improve patient care and clinical trials (Lakhanpal et al., 2021; Coombs et al., 2022).

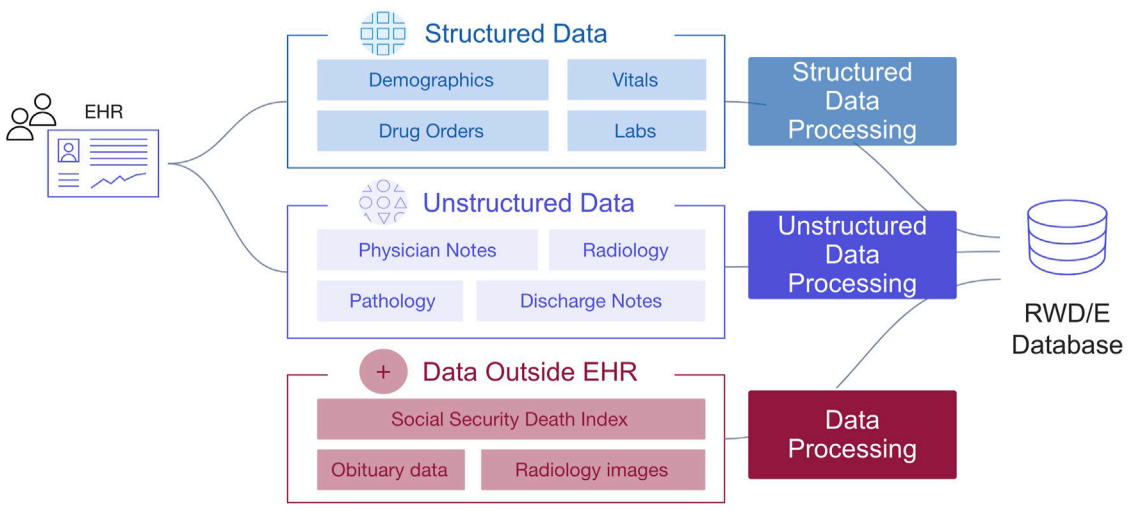

FIGURE 1
Overview of data variables defined by structured and unstructured information in EHR.

TABLE 1 Key terms in machine learning.

Foundational machine learning (ML) definitions
• **Class:** One of the possible values that a binary or categorical variable can take
• **Labels:** The known classes associated with data used to train or evaluate an ML model
• **ML-Extracted:** Algorithmic extraction of data from documented evidence in the patient chart (either structured or unstructured) at the time of running the model. Techniques include ML and natural language processing (NLP), in contrast to other data processing methods such as abstraction or derivation
• **Model:** An ML algorithm with a specific architecture and learned parameters that takes inputs (e.g., text) and produces outputs (e.g., extracted diagnosis)
• **NLP:** A field of computational systems (including but not limited to ML algorithms) that enable computers to analyze, understand, derive meaning from, and make use of human language
• **Score:** A continuous output from a model that can be interpreted as the model-assigned probability that a data point belongs to a specific class
• **Threshold:** A cutoff value that defines classes when applied to continuous scores. Binary variables (e.g., whether a patient has had surgery) have a natural default threshold of 0.5, but different thresholds might be leveraged depending on the relative tolerance for false positives vs false negatives required
Performance metric definitions
• **Sensitivity (Recall):** The proportion of patients abstracted as having a value of a variable (e.g., group stage = IV) that are also ML-extracted as having the same value
• **Positive predictive value (PPV) (Precision):** The proportion of patients ML-extracted as having a value of a variable (e.g., group stage = IV) that are also human abstracted as having the same value
• **Specificity:** The proportion of patients abstracted as not having a value of a variable (e.g., group stage does not = IV) that are also ML-extracted as not having the same value
• **Negative predictive value (NPV):** The proportion of patients ML-extracted as not having a value of a variable (e.g., group stage does not = IV) that are also abstracted as not having the same value
• **Accuracy:** The proportion of patients where the ML-extracted and abstracted values are identical. For variables with more than 2 unique values (e.g., group stage), accuracy within each class is calculated by binarizing the predictions (e.g., for Accuracy of group_stage = IV, all abstracted and ML-extracted values would be defined as either "IV" or "not IV"
• **F1 Score:** Computed as the harmonic mean of sensitivity and PPV. For a binary classifier, the threshold that maximizes F1 can be considered the optimal balance of sensitivity and PPV.

Materials and methods

Overview

We developed a set of research analysis variables using information from the documents available in patient charts. Variables were selected for exploration of ML extraction if commonly required for retrospective observational studies in oncology, but not consistently available in claims data or structured EHR data, and high-quality training data were available that had been manually curated by experts to produce a large amount of abstracted data available for training models (Haimson et al.).

The variables curated through our ML extraction approach are those where the values are solely derived from the identification of clinical details in the EHR documents by an ML model in combination of NLP techniques and rules-based logic. It is important to note that these values are not predictions or inferences, but rather a direct extraction of information that is clearly documented in the EHR.

EHR-derived data source

This study used the nationwide Flatiron Health EHR-derived de-identified database. The Flatiron Health database is a longitudinal database, comprising de-identified patient-level structured and unstructured data (Birnbaum et al., 2020; Ma et al., 2023). At the time of this research, the database included de-identified data from approximately 280 US cancer practices (~800 distinct sites of care).

Structured and unstructured data modalities are available in the database. EHR structured data elements include, but are not limited to, documented demographics (e.g., year of birth, sex, race/ethnicity, etc.), vitals (e.g., height, weight, temperature, etc.), visits, labs, practice information, diagnosis codes, medication orders, medication administrations, ECOG performance status, health insurance coverage, and telemedicine (Figure 1). EHR unstructured data and documents include, but are not limited to, paragraphs of clinic visit notes, PDF scans of lab results, radiology images with reports, pathology reports, and communications between the patient and care team (Figure 2). For the purpose of this paper, all the figures contain fictional representations of documents, sentences, dates and patient IDs.

Patient population

The large general cross-tumor cohort includes all patients with at least one International Classification of Diseases (ICD)-9 or ICD-10 cancer code and at least one unique-date clinic encounter

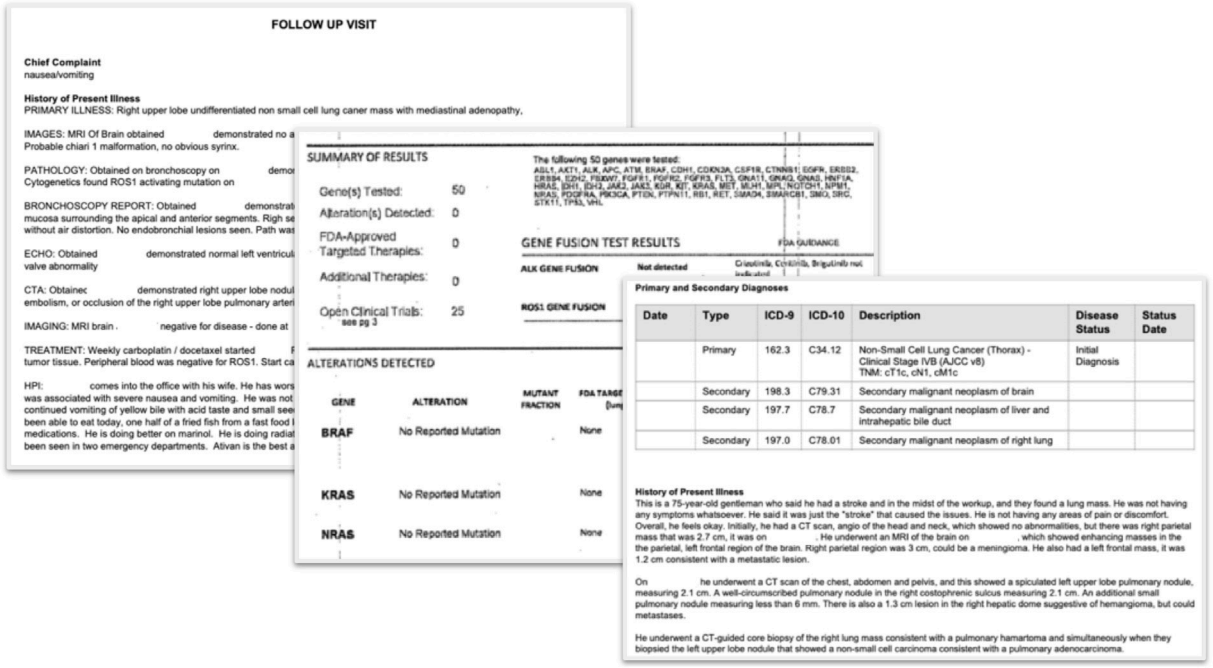

FIGURE 2
Examples of unstructured documents from EHR that are used as inputs for ML-extraction of information (all dates and patient IDs are fictitious).

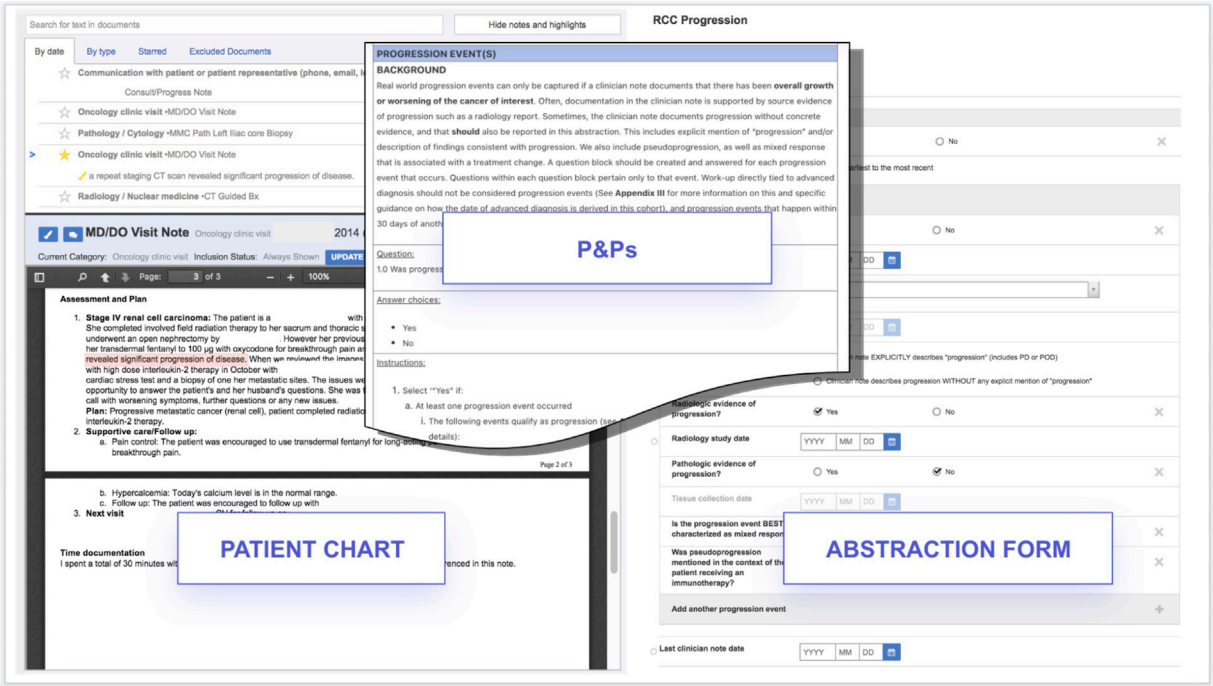

FIGURE 3
Technology enabled expert abstraction. Abbreviations: P&Ps, Policies and Procedures. All dates and patient IDs are fictitious.

documented in the EHR (reflected by records of vital signs, treatment administration, and/or laboratory tests) on or after 1 January 2011. The distribution of patients across community and academic practices largely reflects patterns of care in the US, where most patients are treated in community clinics, but can vary between cancer types.

Clinical expert abstraction of variables for model development

Critical information in patient charts has been manually abstracted by trained clinical experts (i.e., clinical oncology nurses or tumor registrars), following a set of standardized policies and procedures. To abstract data from patient charts, we use a foundational technology (Shklarski et al., 2020) that enables clinical experts to more easily review hundreds of pages of documents to determine patient characteristics, treatments, and outcomes documented in the EHR (Figure 3).

Years of manual abstraction by a workforce of thousands of abstractors at Flatiron Health have created a large and high-quality corpus of labeled oncology EHR data. Clinically-relevant details specific to each cancer type are abstracted from every form of clinical documentation available in the EHR, including clinic visit notes, radiology reports, pathology reports, etc. Abstractors are trained to locate and document relevant information by following policies and procedures tested and optimized for reliability and reproducibility through iterative processes, and oversight is provided by medical oncologists.

The abstraction process undergoes continuous auditing to monitor abstractor performance, while proprietary technology links each curated data point to its source documentation within the EHR, enabling subsequent review. At the individual patient level, this approach provides a recent and robust longitudinal view into the clinical course, capturing new clinical information as it is documented within the EHR.

Flatiron Health has abstracted sets of clinically meaningful variables from more than 300,000 people with cancer to develop disease-specific de-identified research-ready databases (Ma et al., 2023). Limited by the capacity of human abstractors, there had remained millions of patients with cancer in the Flatiron Health database for whom no unstructured data had yet been curated to create variables with the clinical depth needed to generate meaningful insights. If a hypothetical variable required 30 min of chart review by a clinical expert to abstract the information of interest for 1 patient, then it would take a team of 100 full-time abstractors more than 7 years to finish defining 1 variable for a population of 3 million patients.

Overview of machine learning extraction approach

The objective of this application of NLP and ML methods was to replicate the expert abstraction process described in the previous section. When developing ML models for extracting information, all of the clinical abstractor expertise that was incorporated into the manual abstraction of variables is available to learn from through training. Once iterated upon and placed in production, ML models can automate information extraction from unstructured clinical data sources in a way that mimics expert clinical abstractors. The models expand on previously established technology infrastructure that includes deep learning architectures (Rich et al., 2023), text snippet-based modeling approaches (Birnbaum and Ambwani), and extraction of patient events and dates (Gippetti et al.; Ballre et al., 2022; Rich et al., 2022).

Alongside the manually-abstracted labels, we use NLP to pull relevant textual information from charts to use as inputs to train built-for-purpose ML models and model architectures for a given extraction task. Through this process we can make our end variables appropriate for disease-specific or pan-tumor (i.e., histology-independent) applications. For example, by deciding whether or not to use model training data sourced from curated disease-specific cohorts or any-cancer cohorts, we can make our model's output variables built-for-purpose to be used in an analysis that generates meaningful RWE for a specific research question.

A range of model architectures were evaluated and considered for the purpose of information extraction for variables of interest. The model output of variable classes ranged, including:

- binary (e.g., metastatic diagnosis Yes/No)
- categorical unordered (e.g., never smoker, history of smoking, current smoker)
- categorical ordered (e.g., cancer stage I-IV)
- date (e.g., 02/05/2019 start of oral treatment X)

Date and classification can come from the same model, separate models, or connected models.

Natural language processing to generate model inputs

For each variable of interest, we begin with clinical experts constructing a list of clinical terms and phrases related to the variable. Since models are trying to extract explicit information from charts, rather than infer it, only terms that are directly relevant to a specific variable are included (e.g., when extracting a patient's histology, terms could include "histology," "squamous," and/or "adenocarcinoma," but do not include treatment or testing terms from which the histology might be indirectly inferred).

Next, we use NLP techniques to identify sentences in relevant unstructured EHR documents (e.g., oncology visit notes, lab reports, etc.) that contain a match to one of the clinical terms or phrases. The approach uses optical character recognition (OCR) systems to extract text from PDFs, faxes, or scans containing images; the text is then searched for relevant clinical terms. The contextual information surrounding the clinical term is critical because the words at the beginning of a sentence may change the interpretation of a key word at the end of a sentence. ML models can understand if the clinical concept appears and under what context—such as, if negativity, speculation, or affirmation exists in the surrounding clinical terms (i.e., snippets). Where applicable, any associated dates within these sentences are also identified. These sentences are then transformed into a mathematical representation that the model can interpret. The output of this document processing is a broad set of features aimed at fully capturing document structure, chronology, and clinical terms or phrases.

Machine learning model development

Features and labels

The features defined by NLP become the inputs provided to the model to score the likelihood that a given patient document is

associated with each class of a particular categorical variable (e.g., histology categories of non–squamous cell carcinoma, squamous cell carcinoma, non–small cell lung cancer [NSCLC] histology not otherwise specified). The final model output is the variable value for each patient. The labeled dataset is commonly split into three subsets: a training set, a validation set, and a test set. The training and validation sets are used to build the model, which often involves an iterative development process, while the test set is used to evaluate the performance of the final ML model.

Model development

The training set comprises labeled data points that are used to optimize the model's parameter values. In an iterative process, training examples are provided to the model, its outputs are compared to the labels, and the parameter values are adjusted in response to errors. By using manually-abstracted values as labels, the objective of this process is for the model to learn what answer a human abstractor would give when reading a specific clinical text.

The validation set is used to assess how well the model has learned these associations. Because the model does not see any data from patients in the validation examples during training, they can be used to estimate how it will perform on new, unlabeled examples once it is put into production. Validation performance is commonly assessed using metrics such as precision, recall, and F1 score (See Table 1 Key Terms in Machine Learning). These aggregate metrics, combined with review of individual errors, inform decisions about search terms, text preprocessing steps, and model architectures. Experimentation continues until a final "best" model is identified.

When a ML model is trained to perform a classification task, it outputs scores for each possible class for each data point. These scores are between 0 and 1 and show the probability that a patient belongs to each class, based on information in their electronic health record. However, the scores may vary if the wording in the records is unusual or if there is conflicting information. For example, if a patient's cancer stage is being restaged, there may be multiple mentions of different stages in the record, and the model may assign moderate scores to each stage if the restaging event is unclear.

To produce a discrete class value, the class with the highest score is often chosen, but other approaches may optimize performance. In particular, a probability threshold may be chosen such that a patient will be classified into one class if and only if their score exceeds the threshold. The optimal threshold depends on factors such as class balance and is typically chosen empirically (Lipton et al., 2014). When no class receives a sufficiently high score, another option is to defer to abstraction to resolve uncertainty (Waskom et al., 2023).

We explored and experimented with a range of ML models and architectures for the purpose of extracting specific variable information from the EHR. Deep learning architectures included long short-term memory (LSTM), Gated recurrent units (GRU), and bidirectional encoder representations from transformers (BERT) (Hochreiter and Schmidhuber, 1997; Shickel et al., 2018; Devlin et al., 2018). These models can learn thousands or millions of parameters, which enable them to capture subtleties in the text. They read sentences as a whole and use the words around a clinical term to incorporate surrounding context when determining the extracted class. When they receive very large texts as inputs, they can figure out where the relevant information is and focus on this section and its context.

For example, in LSTMs, words are passed into the model sequentially; during each step through a sentence, the model has access to memory (i.e., an internal state) that is impacted by the previous word, in effect allowing the model to "remember" the previous word (Figure 4). The LSTM block combines the new word with the information that came before to derive a more contextually rich representation of the word. For instance, when the LSTM reads the word "Advanced," it remembers (via the model's internal state) that it was preceded by the word "not" and is more likely to classify the patient as "not advanced."

Model evaluation and performance assessment

Once iteration on the ML model is complete, final model performance is measured on a test set that uses manually-abstracted labels as the source of truth. Test sets are designed to be large enough to power both top-level metrics and sub-group stratifications on a "held out" set, that is, on data not used to train the ML model or validate performance during prototyping. This allows the test set to assess the model's ability to correctly classify data points that it has never seen before, which is typically referred to as the "generalization" of the model.

Measuring performance is a complex challenge because even a model with good overall performance might systematically underperform on a particular subcohort of interest, and because while conventional metrics apply to individual models, dozens of ML extracted variables may be combined to answer a specific research question. We use a research-centric evaluation framework (Estévez et al., 2022) to assess the quality of variables curated with ML. Evaluations include one or more of the following strategies: 1) overall performance assessment, 2) stratified performance assessment, and 3) quantitative error analysis, and 4) replication analysis. As variables curated with NLP and ML are expected to be incorporated into the evidence generated that will guide downstream decision-making, variable evaluation can also include replication of analyses originally performed using abstracted data. Replication analyses allow us to determine whether ML-extracted data—either individual variables or entire datasets—are fit-for-purpose in specific use cases by assessing whether they would lead to similar conclusions.

Specific variable-level performance metrics are only interpretable for cohorts with characteristics that are similar to the test set, depending on inclusion criteria such as the type and stage of cancer. As a result, we do not report them here.

Python was the primary coding language used in the development of ML models described here. Institutional Review Board approval of the study protocol was obtained before study conduct, and included a waiver of informed consent.

Results

We successfully extracted key information from unstructured documents in the EHR using the developed proprietary ML models trained on large quantities of data labeled by expert abstractors. For this paper, we are focusing the results on examples within NSCLC as

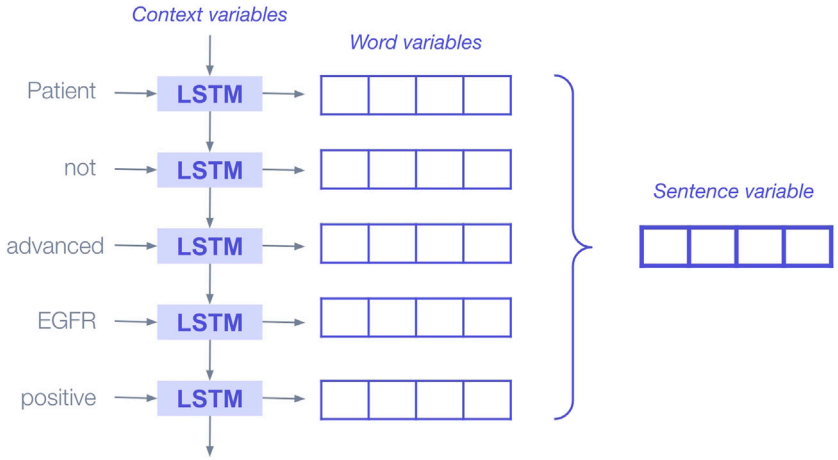

FIGURE 4
Illustration of deep learning bidirectional LSTM blocks applied sequentially to produce representations (aka, embeddings or encodings) that encapsulate the information added to the sentence by each new word. Abbreviations: LSTM, long short-term memory.

they were the first applications we developed. A set of 10 ML models output 20 distinct RWD variables for analysis, including initial cancer diagnosis with date, advanced/metastatic diagnosis with date, disease stage, histology, smoking status, surgery details, biomarker test results, and oral treatments with dates. Language snippets were the inputs for these models to produce a data point for each patient for each variable as outputs, illustrated in Figure 5.

Datatables containing variables curated by an approach using ML had the same appearance and functionality as variables curated with an approach using technology-enabled expert human abstraction (Figure 6).

Models had high performance when trained for disease-specific applications as well as histology-independent (i.e., tumor agnostic) patient cohorts. For example, the NSCLC specific *Histology Type* model had a sensitivity of 96% and a PPV of 94% for extracting non–squamous histology for patients with NSCLC. Detailed performance metrics are out of scope for this paper. Beyond satisfactory ML metrics, we found that in some cases ML-extraction can achieve similar error rates as manual abstraction by clinical experts (Waskom et al., 2023), and replication studies suggest that research analysis relying on multiple variables can reach similar results and conclusions when using variables curated by ML-extraction compared with human experts (Benedum et al., 2022; Sondhi et al., 2022; Benedum et al., 2023).

Approaches and learnings related to specific variables are described below.

Application examples

We have developed ML models for a number of different variables and use cases. A few of the more prominent models and their associated use cases are described below.

Cancer diagnosis and dates

We successfully developed deep learning models focused on the task of extracting initial, advanced, and metastatic cancer diagnosis and the corresponding diagnosis dates. Historically, ICD codes have been used as a proxy for diagnosis, as they are well captured in structured EHR data due to their use in billing. However, we have seen that the precision of ICD codes varies by disease, is not strongly correlated with disease prevalence in the larger population, and can be lower than 50%. With that in mind, extracting accurate diagnosis information is imperative to understanding patient populations, as errors at the diagnosis level propagate to all other variables. These models build on prior foundational research on extracting information from longitudinal clinic notes (Zhao et al., 2021; Agrawal et al., 2018). The initial, advanced, and metastatic variables are generated using multiple, distinct ML models. A conceptual diagram of this approach used by the metastatic variable is presented in Figure 7. We have found success chaining the models together—providing the output of one model as the input to the next—to prevent conflicting predictions and improve overall accuracy. An early investigation into model performance has been presented previously (Rich et al., 2021).

Additional complexity exists when trying to identify patients with rare cancers, primarily due to the low number of labels. We have demonstrated that techniques such as generic token replacement and leave-one-out validation can be effective in combating these complexities—allowing our models to successfully generalize to rare diseases, with few or no labels provided during training from the target disease(s).

Disease stage and histology

We successfully developed a deep learning model to extract cancer stage information and a second ML model to extract the histology of the tumor. One example of how we used this approach for a disease-specific application was training on patients with NSCLC. This model was designed to extract main stage (I-IV) and substage (letters A-C) granularity. Histology was extracted as a non-ordered categorical variable with the possible variable values of non–squamous cell carcinoma, squamous cell carcinoma, or NSCLC histology not otherwise specified.

Deep Learning Model Name	Language in Source EHR as Illustrative Snippet (Model Input)	Extracted Variables (Model Output)
Initial Diagnosis	"Mr. Smith was initially diagnosed with stage IIa NSCLC on 03-31-2017"	*IsDisease, DiseaseDiagnosisDate*
Advanced Diagnosis	"Unfortunately, Mr. Smith developed recurrence of his NSCLC on 09-01-2018"	*IsAdvanced, AdvancedDiagnosisDate*
Metastatic Diagnosis	"Mr. Smith was diagnosed with metastatic NSCLC on 03-31-2017"	*IsMetastatic, MetastaticDiagnosisDate*
Stage	"Mr. Smith was diagnosed with Stage IV NSCLC on 03-31-2017"	*GroupStage*
Histology	"Mr. Smith's biopsy showed a diagnosis of adenocarcinoma of the lung."	*HistologyType*
Smoking Status	"Mr. Smith is an 80-year non-smoker…"	*SmokingStatus*
Surgery	"Mr. Smith underwent wedge resection of his biopsy proven NSCLC."	*HasSurgery, SurgeryDate*
Biomarkers	"Mr. Smith received NGS test results on 2/15/2020 for EGFR, ALK, and ROS1 and was found to have an ALK rearrangement."	*BiomarkerName, BiomarkerStatus, ResultReturnedDate*
PD-L1	"Mr. Smith was diagnosed with adenocarcinoma of the lung, PD-L1 <1% on 2/20/2021."	*BiomarkerName, BiomarkerStatus, ResultReturnedDate*
Orals	"...she has received erlotinib since May 15th 2017 but stopped on Sept 15th 2017 for progression. She was then started on osimertinib on 9/25/2017 and remains on it currently."	*DrugName, StartDate, EndDate*

FIGURE 5
Sentences (fictional examples here) from EHR are inputs to deep learning models that produce a data variable value for each patient as an output. Language snippets are only extracted around key terms from which a variable might be extracted, and not around terms from which it could be indirectly inferred. Abbreviations: EHR, electronic health record; PD-L1, programmed death ligand 1. All dates and patient IDs are fictitious.

As cancer stage is documented similarly across solid tumor diseases, we were able to scale our approach to extract disease stage in a tumor-agnostic cohort with a similar deep learning architecture but training data composed of patients with multiple cancer types. While hematologic cancers have some important differences from solid organ cancers when it comes to assigning stage (risk stratification scores, no concept of metastatic disease, etc.), we found success using a deep learning model to extract this information for a number of hematologic cancers. Tumor histology is not as straightforward to scale across cancer types, as different cancers originate from different possible cell types (and therefore have different histologies). This means that to date, we use distinct histology models for each type of cancer. Performance evaluations for disease stage and histology are conducted at each category level and by cancer type as appropriate for use cases.

Smoking status

We successfully developed a deep learning model to extract information in the patient chart that indicates whether or not the patient has any lifetime history of smoking. The categorical variable output has the possible values as history of smoking, no history of smoking, or unknown. The most relevant sentences for this model were most often found in social history paragraphs of text that is a standard section in clinical encounter notes. Critical document categories that enabled high accuracy of this model included access to oncology clinic visit notes, radiology reports, surgery reports, lab reports, and pulmonary test result reports. The smoking status model was trained on a broad dataset of patients that included many cancer types for whom we have abstracted smoking status.

Surgery and surgery date

We successfully developed a deep learning model to extract information about whether the patient had a primary surgical procedure where the intent was to resect the primary tumor. As these types of surgeries often happen in outpatient facilities or hospitals, this valuable documentation lives in unstructured text formats in the oncology EHR. We have abstracted surgery data in certain disease cohorts but, because of the similarity in documentation approaches across cancer types, we were able to

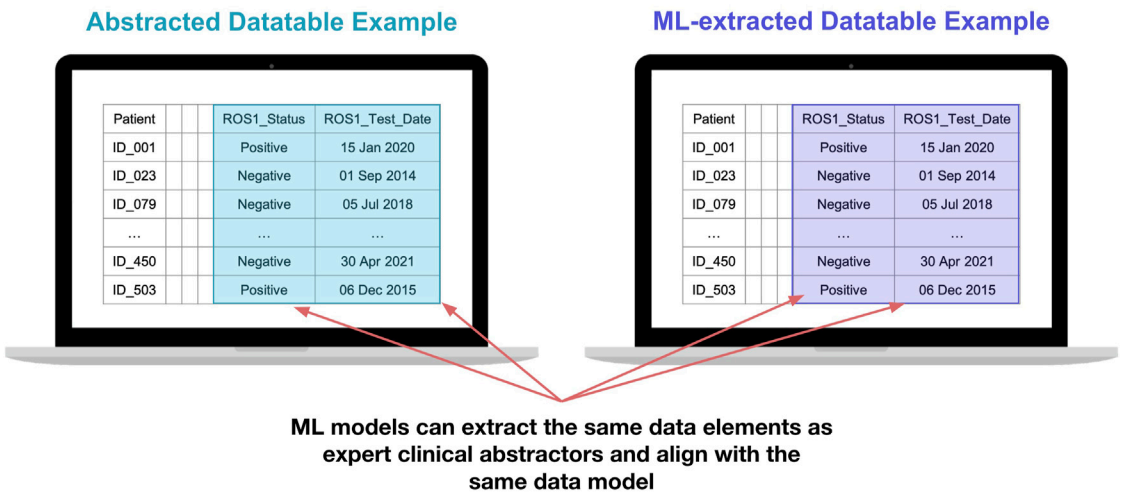

FIGURE 6
Illustration of a data table with variables curated by an approach using expert human abstractors (right) alongside a data table with variables curated by an approach using deep learning models (left) shows opportunity for exchangeable utility in real-world data analysis. All dates and patient IDs are fictitious.

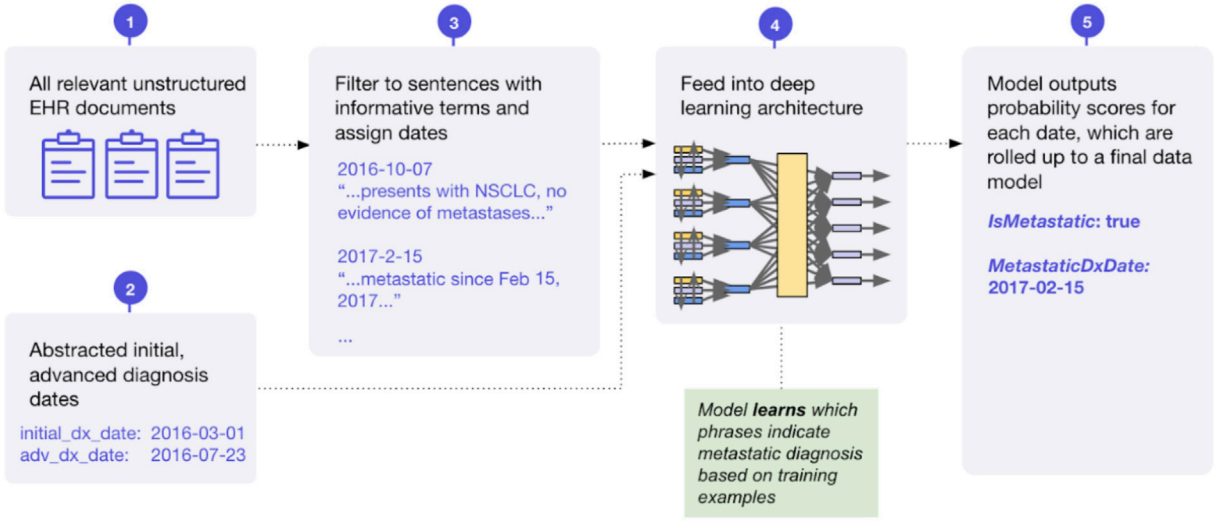

FIGURE 7
Conceptual diagram of machine learning model for extraction of metastatic diagnosis and date.

train a model that is tumor agnostic. This allowed us to scale surgery status and date in larger patient populations and in new disease types.

Biomarker testing results and result date

We successfully developed and deployed models to generate variables for biomarker testing, including extraction of the dates that the patient had results returned (Figure 6). One part of the model is able to identify whether or not a given document for a patient contains a biomarker test result. A separate part of the model is able to extract from the document the date a result was returned and the biomarker result. Early efforts with a regularized logistic regression model were presented previously (Ambwani et al., 2019) and more sophisticated models have been developed since.

A model first cycles through every EHR document for a given patient to understand whether or not the document contains biomarker testing results. These models rely on access to lab reports, including those saved in the EHR as a PDF or image of a scanned fax. The models are able to process report documents produced by different labs (e.g., Foundation Medicine, Caris, Tempus, etc.,) in addition to the clinician interpretations in visit notes.

A separate model then extracts the biomarker (e.g., included but not limited to *ALK, BRAF, EGFR, KRAS, MET, NTRK, RET, ROS1*,

or PD-L1) and test result (e.g., positive, negative, or unknown). This approach gives our ML models flexibility to extract biomarkers that the model may not have seen before in training. For PD-L1, where results are quantitatively reported, a separate ML model was developed to extract percent staining, with classes of <1%, 1%–49%, ≥49%, and unknown.

Since patients can receive biomarker testing multiple times throughout the treatment journey and at multiple facilities, it is possible that a given patient has more than one biomarker test result and date. For each patient, this allows us to determine biomarker status at different clinical milestones (e.g., advanced diagnosis date, start of second-line treatment, etc).

Oral treatments and treatment dates

We successfully developed a deep learning model to extract oral treatment information, including the treatment name, and the span for which the treatment was administered. In contrast to intravenous therapies such as chemotherapy or immunotherapy in which each dose is ordered and administered to be given in the clinic or infusion room, oral therapies are prescribed to patients to be filled by an outpatient pharmacy, which is frequently outside the clinic site. To have a complete understanding of all cancer treatments received or delayed (e.g., postponed during a hospitalization), it is necessary to enumerate the use of oral treatments through review of unstructured clinician visit notes, prescriptions, and communications with the patient or patient representative. Important information to select within the paragraphs of text include the treatment name, start date, and end date. We previously published an initial framework (Agrawal et al., 2018) for extracting drug intervals from longitudinal clinic notes, prescriptions, and patient communication documents and have developed more sophisticated and accurate methods since then. We found the visit notes contained key pieces of information about treatments being held or started when patients were hospitalized.

The model is trained to select mentions of a specific list of drug names used for oral treatment in the specific cancer type, along with the start date and end date. These oral treatment variables are generated using three distinct ML models. The list of oral treatments of interest were specific to each disease and defined by oncology clinicians. Expert abstraction from charts includes policies and procedures for collection of treatment start dates and discontinuation dates as both are needed to execute many common RWE study designs. To be fit for purpose, ML models were trained to extract both start and end dates of treatments.

Discussion

This paper described one approach to curating real-world oncology data variables from unstructured information in EHR using NLP and ML methods. Model development was possible with access to a large and high-quality corpus of labeled oncology EHR data produced via manual abstraction by a workforce of thousands of clinical expert abstractors over the course of several years. We now have models that are able to meet or even exceed human abstraction performance on certain tasks (Waskom et al., 2023). Using a performance evaluation framework (Devlin et al., 2018) for variables curated using the approach of ML extraction we affirmed high quality and fitness-for-use in RWE generation. We have shown that validations using the combination of multiple ML-extracted variables in one RWD analysis demonstrated no meaningful difference in RWE findings based on replications with the Flatiron Health variables curated by ML extraction compared with expert human abstraction (Forsyth et al., 2018; Zeng et al., 2018; Jorge et al., 2019; Karimi et al., 2021; Maarseveen et al., 2021; Benedum et al., 2022; Sondhi et al., 2022; Yang et al., 2022; Benedum et al., 2023).

Crucial information about clinical details may be recorded only within free-text notes or summaries in unstructured EHR documents. Our models primarily rely on deep learning architectures, because curating data from such sources usually requires techniques that capture the nuances of natural language. We select model architectures on a case-by-case basis depending on what works best for each variable, but we have found that the quality of the training data and labels can be just as, if not more, important to success than the architecture used. Despite this, we do expect that advances in generative AI and advancing LLM architectures will make deeper and more nuanced clinical concepts accessible to ML extraction, as LLMs are able to take into account a fuller context of the patient data and rely less on having high quality labels for training. The impressive generative abilities of models like gpt3 and its ChatGPT application have demonstrated this, although the generative framework itself may remain more suited for tasks such as summarization (Adams et al., 2021) than for scalable curation of structured real-world datasets.

The mission to improve and extend lives by learning from the experience of every person with cancer is more important than ever. With increasingly specific combinations of patient characteristics, disease, and therapy, we need to learn from as many relevant examples as possible to have statistically meaningful results. ML expands the opportunity to learn from patients who have been oppressed or historically marginalized in oncology clinical trials (Adamson et al., 2019; Hooley et al., 2019). As oncology care rapidly evolves, and the treatment landscape becomes more personalized—targeting new biomarkers, finely tuned to increasingly particular patient profiles—transparent fit-for-purpose applications of ML will have increasing importance. This will be valuable to gain trust with decision-makers in applications such as postmarket safety surveillance. With high performance models, we can truly learn from every patient, not just a sample. It also creates an opportunity to improve the completeness of RWD variables that were previously defined by only structured data elements, reducing potential bias in evidence.

There are strengths and limitations to the EHR curation approaches described here. Strengths include the large size, representativeness, and quality of training data used; success across a multitude of cancer types; and the explainability of approach to finding clinical details in documents. Massive volumes of high-quality expert abstracted data were a unique advantage for training high-quality ML models. Researchers at Stanford have confirmed similar capabilities with a different EHR dataset—detecting the timeline of metastatic recurrence of breast cancer (Banerjee et al., 2019). An example of a variable that would be challenging for ML extraction could be

microsatellite instability (MSI), where results are reported in a wide range of formats. One of the formats is a graphic where the result is reported visually on a sliding scale rather than in text format. This would be difficult for a model that relies on interpretation of text. The ML models described here were trained for and applied only in a US population (Ma et al., 2023). While the most suitable model architectures for each variable may be transferable across country borders, a limitation of this approach is that models must be re-trained with local data for highest performance.

The capability to build ML models that can extract RWD variables accurately for a large number of patients further enables the possible breadth and depth of timely evidence generation to answer key policy questions and understand the effects of new treatment on health outcomes.

Data availability statement

The data analyzed in this study is subject to the following licenses/restrictions: The data that support the findings of this study have been originated by Flatiron Health, Inc. Requests for data sharing by license or by permission for the specific purpose of replicating results in this manuscript can be submitted to dataaccess@flatiron.com.

Author contributions

BA, JK, SN, and EE contributed to the conception of this review paper. AB, GH, GA, JK, JG, JR, KH, KK, MW, RL, TB, and SW developed the ML models. CB, ME, EF, AC, and BA conducted performance evaluations and validations. BA wrote the first draft of the manuscript. SN, GA, WS, MW, ME, AB, AC, EF, and RL wrote sections of the manuscript. All authors contributed to the article and approved the submitted version.

Funding

This study received funding from Flatiron Health. The funder was involved in the study design, collection, analysis, interpretation of data, the writing of this article or the decision to submit it for publication.

Acknowledgments

The authors would like to thank Flatiron Health's Selen Bozkurt, Sharang Phadke, Shreyas Lakhtakia, Nick Altieri, Qianyu Yuan, Geetu Ambwani, Lauren Dyson, Chengsheng Jiang, Somnath Sarkar, Javier Jimenez, Arjun Sondhi, Alexander Rich, Benjamin Birnbaum, Andrej Rosic, Barry Leybovich, Jamie Irvine, Nisha Singh, Sankeerth Garapati, Hannah Gilham, and Jennifer Swanson. Flatiron Health's Catherine Au-Yeung and Tanya Elshahawi contributed to illustration design. A version of the manuscript is currently under consideration as a preprint at medRxiv.org.

Conflict of interest

Authors BA, MW, AB, JK, KK, SN, JG, JR, KH, GH, RL, TB, SW, GA, EE, CB, EF, ME, WS, and AB are employees of Flatiron Health, Inc., which is an independent member of the Roche group, and own stock in Roche.

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

Adams, G., Alsentzer, E., Ketenci, M., Zucker, J., and Elhadad, N. (2021). What's in a summary? Laying the groundwork for advances in hospital-course summarization. *Proc. Conf.* 2021, 4794–4811. doi:10.18653/v1/2021.naacl-main.382

Adamson, B. J., Cohen, A. B., Cheever, M. A., et al. (2019). "Cancer immunotherapy use and effectiveness in real-world patients living with HIV," Presented at the Abstract Presented at: 17th International Conference on Malignancies in HIV/AIDS. Bethesda, Maryland. October 21-22.

Agrawal, M., Adams, G., Nussbaum, N., et al. (2018). Tifti: A framework for extracting drug intervals from longitudinal clinic notes. arXiv:Preprint posted online Nov 30, 2018

Ambwani, G., Cohen, A., Estévez, M., Singh, N., Adamson, B., Nussbaum, N., et al. (2019). PPM8 A machine learning model for cancer biomarker identification in electronic health records. *Value Health* 22, S334. doi:10.1016/j.jval.2019.04.1631

Ballre, A., Baruah, P., and Amster, G. (2022). *Systems and methods for predicting biomarker status and testing dates*. United States.

Banerjee, I., Bozkurt, S., Caswell-Jin, J., Kurian, A. W., and Rubin, D. L. (2019). Natural language processing approaches to detect the timeline of metastatic recurrence of breast cancer. *JCO Clin. Cancer Inf.* 3, 1–12. doi:10.1200/CCI.19.00034

Benedum, C., Adamson, B., Cohen, A. B., Estevez, M., Sondhi, A., Fidyk, E., et al. (2022). P57 machine learning-accelerated outcomes research: A real-world case study of biomarker-associated overall survival in oncology. *Value Health* 25, S13–S14. doi:10.1016/j.jval.2022.09.069

Benedum, C. M., Sondhi, A., Fidyk, E., Cohen, A. B., Nemeth, S., Adamson, B., et al. (2023). Replication of real-world evidence in oncology using electronic health record data extracted by machine learning. *Cancers (Basel)* 15, 1853. doi:10.3390/cancers15061853

Bera, K., Schalper, K. A., Rimm, D. L., Velcheti, V., Madabhushi, A., et al. (2019). Artificial intelligence in digital pathology - new tools for diagnosis and precision oncology. *Nat. Rev. Clin. Oncol.* 16, 703–715. doi:10.1038/s41571-019-0252-y

Bertsimas, D., and Wiberg, H. (2020). Machine learning in oncology: Methods, applications, and challenges. *JCO Clin. Cancer Inf.* 4, 885–894. doi:10.1200/CCI.20.00072

Bhardwaj, R., Nambiar, A. R., and Dutta, D. (2017). A study of machine learning in healthcare. Presented at the Abstract Presented at: 2017 IEEE 41st Annual Computer Software and Applications Conference (COMPSAC). Turin, Italy. July 4-8.

Birnbaum, B., Nussbaum, N., Seidl-Rathkopf, K., et al. (2020). Model-assisted cohort selection with bias analysis for generating large-scale cohorts from the EHR for oncology research. arXiv:Preprint posted online January 13.

Birnbaum, B. E., and Ambwani, G. (2023). *Generalized biomarker model*. United States.

Birnbaum, B. E., Haimson, J. D., and He, L. D. (2023). *Systems and methods for automatic bias monitoring of cohort models and un-deployment of biased models*. United States.

Birnbaum, B. E., Haimson, J. D., and He, L. D. (2019). *Systems and methods for model-assisted cohort selection*. United States.

Blueprint for trustworthy AI implementation guidance and assurance for healthcare (2022). December 7, update https://www.coalitionforhealthai.org/insights.

Center for Drug Evaluation and Research Center for Biologics Evaluation and Research Oncology Center of Excellence (2022). *Real-world data: Assessing electronic health records and medical claims data to support regulatory decision-making for drug and biological products; draft guidance for industryhttps*. Available at: https://www.fda.gov/regulatory-information/search-fda-guidance-documents/real-world-data-assessing-electronic-health-records-and-medical-claims-data-support-regulatory.

Chen, R., Garapati, S., Wu, D., Ko, S., Falk, S., Dierov, D., et al. (2019). Machine learning based predictive model of 5-year survival in multiple myeloma autologous transplant patients. *Blood* 134, 2156. doi:10.1182/blood-2019-129432

Coombs, L., Orlando, A., Wang, X., Shaw, P., Rich, A. S., Lakhtakia, S., et al. (2022). A machine learning framework supporting prospective clinical decisions applied to risk prediction in oncology. *NPJ Digit. Med.* 5, 117. doi:10.1038/s41746-022-00660-3

Datta, S., Bernstam, E. V., and Roberts, K. (2019). A frame semantic overview of NLP-based information extraction for cancer-related EHR notes. *J. Biomed. Inf.* 100, 103301. doi:10.1016/j.jbi.2019.103301

Devlin, J., Chang, M., and Lee, K. (2018). Bert: Pre-training of deep bidirectional transformers for language understanding. arXiv:Preprint posted online October 11, 2018.

Estévez, M., Benedum, C. M., Jiang, C., Cohen, A. B., Phadke, S., Sarkar, S., et al. (2022). Considerations for the use of machine learning extracted real-world data to support evidence generation: A research-centric evaluation framework. *Cancers (Basel)* 14, 3063. doi:10.3390/cancers14133063

Forsyth, A. W., Barzilay, R., Hughes, K. S., Lui, D., Lorenz, K. A., Enzinger, A., et al. (2018). Machine learning methods to extract documentation of breast cancer symptoms from electronic health records. *J. Pain Symptom Manage* 55, 1492–1499. doi:10.1016/j.jpainsymman.2018.02.016

Gippetti, J., Phadke, S., and Amster, G. (2023). *Systems and methods for extracting dates associated with a patient condition*. United States.

Haimson, J. D., Baxi, S., and Meropol, N. (1997). *Prognostic score based on health information*. United States.

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

Hooley, I., Chen, R., Long, L., Cohen, A., and Adamson, B. (2019). PCN166 optimization of natural language processing-supported comorbidity classification algorithms in electronic health records. *Value Health* 22, S87. doi:10.1016/j.jval.2019.04.290

Jorge, A., Castro, V. M., Barnado, A., Gainer, V., Hong, C., Cai, T., et al. (2019). Identifying lupus patients in electronic health records: Development and validation of machine learning algorithms and application of rule-based algorithms. *Semin. Arthritis Rheum.* 49, 84–90. doi:10.1016/j.semarthrit.2019.01.002

Karimi, Y. H., Blayney, D. W., Kurian, A. W., Shen, J., Yamashita, R., Rubin, D., et al. (2021). Development and use of natural language processing for identification of distant cancer recurrence and sites of distant recurrence using unstructured electronic health record data. *JCO Clin. Cancer Inf.* 5, 469–478. doi:10.1200/CCI.20.00165

Kelly, J., Wang, C., Zhang, J., Das, S., Ren, A., and Warnekar, P. (2022). Automated mapping of real-world oncology laboratory data to LOINC. *AMIA Annu. Symp. Proc.* 2021, 611–620.

Koleck, T. A., Dreisbach, C., Bourne, P. E., Bakken, S., et al. (2019). Natural language processing of symptoms documented in free-text narratives of electronic health records: A systematic review. *J. Am. Med. Inf. Assoc.* 26, 364–379. doi:10.1093/jamia/ocy173

Lakhanpal, S., Hawkins, K., Dunder, S. G., Donahue, K., Richey, M., Liu, E., et al. (2021). An automated EHR-based tool to facilitate patient identification for biomarker-driven trials. *JCO* 39, 1539. doi:10.1200/jco.2021.39.15_suppl.1539

Lipton, Z. C., Elkan, C., and Naryanaswamy, B. (2014). Optimal thresholding of classifiers to maximize F1 measure. *Mach. Learn Knowl. Discov. Databases* 8725, 225–239. doi:10.1007/978-3-662-44851-9_15

Ma, X., Long, L., and Moon, S. (2023). Comparison of population characteristics in real-world clinical oncology databases in the US: Flatiron health, SEER, and NPCR. medRxiv. Preprint posted online June 07.

Maarseveen, T. D., Maurits, M. P., Niemantsverdriet, E., van der Helm-van Mil, A. H. M., Huizinga, T. W. J., and Knevel, R. (2021). Handwork vs machine: A comparison of rheumatoid arthritis patient populations as identified from EHR free-text by diagnosis extraction through machine-learning or traditional criteria-based chart review. *Arthritis Res. Ther.* 23 (1), 174. doi:10.1186/s13075-021-02553-4

NICE (2022). *NICE real-world evidence framework*. Available at: https://www.nice.org.uk/corporate/ecd9/chapter/overview.

Norgeot, B., Quer, G., Beaulieu-Jones, B. K., Torkamani, A., Dias, R., Gianfrancesco, M., et al. (2020). Minimum information about clinical artificial intelligence modeling: The MI-CLAIM checklist. *Nat. Med.* 26, 1320–1324. doi:10.1038/s41591-020-1041-y

Padula, W. V., Kreif, N., Vanness, D. J., Adamson, B., Rueda, J. D., Felizzi, F., et al. (2022). Machine learning methods in health economics and outcomes research—the PALISADE checklist: A good practices report of an ISPOR task force. *Value Health* 25, 1063–1080. doi:10.1016/j.jval.2022.03.022

Rich, A., Amster, G., and Adams, G. (2023). *Deep learning architecture for analyzing unstructured data*. United States.

Rich, A., Leybovich, B., and Irvine, B. (2022). *Machine learning model for extracting diagnoses, treatments, and key dates*. United States.

Rich, A. S., Leybovich, B., Estevez, M., Irvine, J., Singh, N., Cohen, A. B., et al. (2021). Extracting non-small cell lung cancer (NSCLC) diagnosis and diagnosis dates from electronic health record (EHR) text using a deep learning algorithm. *J. Clin. Oncol.* 39, 1556. doi:10.1200/jco.2021.39.15_suppl.1556

Shah, P., Kendall, F., Khozin, S., Goosen, R., Hu, J., Laramie, J., et al. (2019). Artificial intelligence and machine learning in clinical development: A translational perspective. *NPJ Digit. Med.* 2, 69. doi:10.1038/s41746-019-0148-3

Shickel, B., Tighe, P. J., Bihorac, A., Deep, E. H. R., et al. (2018). Deep EHR: A survey of recent advances in deep learning techniques for electronic health record (EHR) analysis. *IEEE J. Biomed. Health Inf.* 22, 1589–1604. doi:10.1109/JBHI.2017.2767063

Shklarski, G., Abernethy, A., and Birnbaum, B. (2020). *Extracting facts from unstructured data*. United States.

Sondhi, A., Benedum, C., Cohen, A. B., Nemeth, S., Bozkurt, S., et al. (2022). RWD112 can ML-extracted variables reproduce real world comparative effectiveness results from expert-abstracted data? A case study in metastatic non-small cell lung cancer treatment. *Value Health* 25, S470. doi:10.1016/j.jval.2022.09.2337

Subbiah, V. (2023). The next generation of evidence-based medicine. *Nat. Med.* 29, 49–58. doi:10.1038/s41591-022-02160-z

Wang, L., Wampfler, J., Dispenzieri, A., Xu, H., Yang, P., Liu, H., et al. (2019). Achievability to extract specific date information for cancer research. *AMIA Annu. Symp. Proc.* 2019, 893–902. Published 2020 Mar 4, 2020.

Waskom, M. L., Tan, K., Wiberg, H., et al. (2023). A hybrid approach to scalable real-world data curation by machine learning and human experts. *medRxiv:Preprint posted online March* 8. doi:10.1101/2023.03.06.23286770

Yang, R., Zhu, D., Howard, L. E., De Hoedt, A., Williams, S. B., Freedland, S. J., et al. (2022). Identification of patients with metastatic prostate cancer with natural language processing and machine learning. *JCO Clin. Cancer Inf.* 6, 2022, e2100071. doi:10.1200/CCI.21.00071

Zeng, Z., Espino, S., Roy, A., Li, X., Khan, S. A., Clare, S. E., et al. (2018). Using natural language processing and machine learning to identify breast cancer local recurrence. *BMC Bioinforma.* 19, 498–x. doi:10.1186/s12859-018-2466-x

Zhao, J., Agrawal, M., Razavi, P., et al. (2021). Directing human attention in event localization for clinical timeline creation. *PMLR* 149, 80–102.

frontiers | Frontiers in Pharmacology

TYPE Perspective
PUBLISHED 20 September 2023
DOI 10.3389/fphar.2023.1249611

Check for updates

OPEN ACCESS

EDITED BY
Dalia M. Dawoud,
National Institute for Health and Care Excellence, United Kingdom

REVIEWED BY
Anton Avanceña,
The University of Texas at Austin, United States

*CORRESPONDENCE
Emma K. Mackay,
emma.mackay@cytel.com

RECEIVED 29 June 2023
ACCEPTED 11 September 2023
PUBLISHED 20 September 2023

CITATION
Mackay EK and Springford A (2023), Evaluating treatments in rare indications warrants a Bayesian approach. *Front. Pharmacol.* 14:1249611. doi: 10.3389/fphar.2023.1249611

COPYRIGHT
© 2023 Mackay and Springford. This is an open-access article distributed under the terms of the Creative Commons Attribution License (CC BY). The use, distribution or reproduction in other forums is permitted, provided the original author(s) and the copyright owner(s) are credited and that the original publication in this journal is cited, in accordance with accepted academic practice. No use, distribution or reproduction is permitted which does not comply with these terms.

Evaluating treatments in rare indications warrants a Bayesian approach

Emma K. Mackay* and Aaron Springford

Cytel, Toronto, ON, Canada

Evaluating efficacy and real-world effectiveness for novel therapies targeting rare mutations or patient subpopulations with unmet needs is a growing challenge in health economics and outcomes research (HEOR). In these settings it may be difficult to recruit enough patients to run adequately powered randomized clinical trials, resulting in greater reliance on single-arm trials or basket trial designs. Additionally, evidence networks for performing network meta-analysis may be sparse or disconnected when comparing available treatments in narrower patient populations. These challenges create an increased need for use of appropriate methods for handling small sample sizes, structural modelling assumptions and more nuanced decision rules to arrive at "best-available evidence" on comparative and non-comparative efficacy/effectiveness. We advocate for greater use of Bayesian methods to address these challenges as they can facilitate efficient and transparent borrowing of information across varied data sources under flexible modelling assumptions, probabilistic sensitivity analysis to assess model assumptions, and more nuanced decision-making where limited power reduces the utility of classical frequentist hypothesis testing. We illustrate how Bayesian methods have been recently used to overcome several challenges of rare indications in HEOR, including approaches to borrowing information from external data sources, evaluation of efficacy in basket trials, and incorporating non-randomized studies into network meta-analysis. Lastly, we provide several recommendations for HEOR practitioners on appropriate use of Bayesian methods to address challenges in the rare disease setting.

KEYWORDS

Bayesian, statistical methods, health economics and outcomes research (HEOR), evidence synthesis, real-world evidence (RWE), rare diseases, comparative effectiveness

Introduction

A core task of health economics and outcomes research (HEOR) is to compare the effectiveness of two or more competing treatments. Over the past several decades, researchers in HEOR have been working to realize the promise of a "big data" revolution in which an excess of evidence can be brought to bear on any given decision problem (Berger and Doban, 2014). However, due to advances in health technologies which target smaller populations and/or very rare diseases we continue to see challenges of limited data and small sample sizes. In response to these trends, we advocate for modern Bayesian approaches which can incorporate all available information in a principled and transparent way. In our view, Bayesian approaches are particularly valuable if primary data sources are insufficient to establish reliable and statistically conclusive superiority of a novel treatment compared to the *status quo*. In these cases, the novel treatments that are urgently needed by patients may be passed over if typical large-sample,

dichotomous statistical significance thresholds are treated as an unquestioned default by decision-makers.

While Bayesian methods have seen substantial uptake in the area of meta-analysis–for example, in guidance from the UK National Institute for Health and Care Excellence's (NICE) Decision Support Unit (DSU) (Dias et al., 2011)–, we suggest that significant gains can also be made in rare disease settings where sample sizes and available evidence bases are more limited. A goal of this paper is to provide examples and guidance on how practitioners can incorporate external information using Bayesian modelling to address some of the challenges of evaluating efficacy/effectiveness that arise in health technology assessments (HTA) of newly developed therapies for rare indications. We point to some key applications in which we believe important gains can be made: borrowing from external sources to augment a concurrent control arm or to estimate a historical control rate for rare diseases; incorporating disparate data sources (such as randomized controlled trial (RCT) and non-randomized study (NRS) data) into a meta-analysis; and applying Bayesian hierarchical models (BHM) to partially pool information across heterogeneous data sources. In each of these applications, common questions emerge: 1) What relevant information can we draw on to improve existing analyses and estimates? 2) When existing data are limited, what assumptions might enable incorporation of external information, and are these plausible? Or, when very strong assumptions are needed, what would constitute "best-available evidence"? And 3) how can we characterize the limitations of the analysis and assess sensitivity to violations of key assumptions?

What are Bayesian methods and why use them?

Bayesian inference defines a probability model for data which is a function of parameters (the likelihood), and a probability model for parameters before any data are observed (the prior). After data are observed, the prior and the likelihood are used to calculate a probability distribution for the parameters given the data (the posterior). If there is information available which is related directly to the model parameters, it can be included in the prior. If there is information available in the form of additional data from another source, it can be included in the likelihood. The posterior distribution contains all available information about the model parameters, and in practice is a very useful mathematical object. For example, functions of the posterior such as the probability that one treatment is superior to another, or the expected benefit of selecting one treatment over another, or the distribution of predicted patient outcomes in a given population, can all be obtained without much additional computational effort. In the frequentist approach, many of these derived quantities are not available, and even if they are available their calculation is considerably more burdensome.

Because Bayesian methods lead to probability statements about model parameters, they are vital to formal decision analysis and thus HTA (see Spiegelhalter et al. (1999) and examples therein). Bayesian inference leads to statements like: "there is a 95% probability that the hazard ratio is between 0.6 and 0.84"; whereas frequentist inference leads to statements such as: "if the trial were repeated many times, and a 95% confidence interval constructed for each, the true hazard ratio would be within 95% of the intervals." In an HTA context, we argue that the former is not only more interpretable, but also more directly useful for decision making. A more extensive comparison of Bayesian and frequentist methods can be found in Spiegelhalter et al. (1999).

How can Bayesian borrowing help bolster limited sample sizes in HEOR analyses?

Bayesian borrowing methods can incorporate information about model parameters (e.g., the control arm response rate) from external data in a transparent manner. These methods allow for down-weighting of the external data to mitigate potential bias arising from different parameter values in the current population compared to the external populations. One established approach is to borrow information by means of a power prior (Ibrahim and Chen, 2000; Ibrahim et al., 2015). The power prior is formed by taking a prior for the parameter and combining it with a discounted likelihood for the parameter on the external data. The external data parameter likelihood is discounted by raising it to the power of a discount parameter between 0 and 1. A discount parameter value of 0 corresponds to no borrowing and a value of 1 yields complete pooling of the datasets. Due to the challenge of selecting a value for the discount parameter, one option for practitioners is to vary the discount parameter and assess how much borrowing from the external data is required before a specified decision threshold—or "tipping point"—is reached (e.g., for concluding that a treatment is effective). This sort of tipping point analysis has precedent in a regulatory context when using Bayesian borrowing (US Food and Drug Administration, 2018). Another option is to use dynamic borrowing, in which the discount parameter is treated as a random quantity with its own prior distribution. In theory, this approach allows the amount of borrowing to depend on the degree of agreement in observed outcomes between the current and external data sources. In practice, there may not be much information available in the data about the discount parameter, and results can be sensitive to the choice of prior. Regardless, proper implementation of dynamic borrowing using a power prior requires a normalization step (Neuenschwander et al., 2009) which can be computationally challenging to implement. Ibrahim et al. (2015) provide a more detailed overview of power priors, including extensions such as commensurate priors, for interested readers. Additionally, Viele et al. (2014) compare several approaches to borrowing from historical data, including the use of power priors.

Another prior-based approach to Bayesian borrowing is to formulate a meta-analytic predictive (MAP) prior for the parameter of interest (Neuenschwander et al., 2010). As an example, suppose we want to borrow information on the response rate for the control treatment. We conduct a Bayesian meta-analysis (typically a random effects meta-analysis) to obtain the posterior predictive distribution for the control treatment response rate. This posterior then becomes the prior for the response rate in our concurrent control arm—if there is one—or represents the entirety of information available for this parameter if there is no concurrent control arm. The posterior predictive distribution is preferred because it incorporates heterogeneity in response rates across trial populations, and we seek to generalize from the external populations to our current population. A narrower/more precise MAP prior in effect represents a larger sample size being borrowed from the external data. In cases where the generalization from external to current is insufficiently conservative, robust MAP priors have been used (Schmidli et al., 2014). Robust MAP priors are defined as a weighted mixture of the MAP prior

and a vague prior. This approach is analogous to the power prior in that placing more weight on the vague component in the mixture results in a more diffuse prior distribution which imparts less information, down-weighting the contribution of the external data.

Power prior and robust MAP prior methods have different strengths and weaknesses in practice. Power prior methods can be more challenging to implement (especially when the discount parameter is a random quantity), but they have a simple form and can easily be adapted to incorporate disparate sources of external information. MAP prior methods will be more familiar to those experienced with Bayesian meta-analysis, and may be easier to explain and justify in many HEOR settings. Both approaches can incorporate aggregate data and/or individual patient data (IPD) from multiple sources, and both can be used for tipping point analysis if desired (US Food and Drug Administration, 2018; Best et al., 2021). In one prospective RCT using robust MAP to reduce control group allocation, variance of the robust MAP prior was inflated to achieve a target effective sample size (Richeldi et al., 2022)—a practical approach to borrowing which could also be applied to a power prior with fixed discount parameter.

How can we model structural relationships between data sources while also accounting for potential heterogeneity?

In cases where a structural relationship among data sources can be assumed, Bayesian hierarchical models (BHMs) are another option for partial pooling of information in which hierarchical dependencies of key parameters are modeled explicitly (Gelman et al., 2013). BHMs assume that some model parameters are related by virtue of being drawn from a common distribution—i.e., that they are exchangeable—but that the parameters of the distribution are themselves random quantities. For example, response rates for a specific control treatment are often assumed to be heterogeneous across data sources but nonetheless may be interrelated. Under a BHM approach, information on the control treatment response rate can be partially pooled across data sources, with the amount of pooling dependent on the degree of heterogeneity in response rates across data sources (less borrowing occurs if response rates are very heterogeneous). This also has the effect of shrinking parameter estimates towards the grand mean, mitigating overfitting and improving inference for individual parameters, particularly when data are limited (Gelman et al., 2013).

To illustrate the utility of BHMs in the HEOR space, we focus on some recent applications to analyses of basket trials. Basket trial designs include patients with multiple cancer types which share a common targetable mutation or biomarker. In these basket trials, sample sizes tend to be extremely limited, treatment responses are expected to vary among tumour types, and control arms are often omitted. Murphy et al. (2021) use a BHM approach in a single-arm basket trial setting for evaluating response for NTRK fusion-positive patients receiving larotrectinib. Their approach allows for partial borrowing of information on response rates across tumour types to produce estimates of response for individual tumour types, the overall basket of represented tumours, and for an unrepresented histology. BHM approaches were also well-received in a NICE technical appraisal for larotrectinib (UK National Institute for Health and Care Excellence, 2020).

In the BHM approach to analysis of basket trials, exchangeability of tumour types may be a clinically tenuous assumption (although perhaps an acceptable approximation in light of data limitations if the BHM is flexible enough to describe the data). Neuenschwander et al. (2016) propose an exchangeable-non-exchangeable (EXNEX) model which allows for relaxation of strong exchangeability assumptions, and we envision future methodological developments in this area. Mackay et al. (Mackay et al., 2022; Mackay et al., 2023) have recently proposed an extension of BHM modelling for histology-independent therapies to allow for indirect treatment comparisons (ITC) between multiple basket trials. The approach allows for adjustment for potential confounding due to differences in tumour type compositions between trials while preserving limited precision/power by means of partial pooling. The reader is directed to Murphy et al. (2022) for a more detailed discussion of modelling approaches for histology-independent therapies in an HTA context.

While hierarchical models can be implemented under both a Bayesian and classical frequentist approach, a key advantage of Bayesian approaches is the ability to incorporate prior information and perform probabilistic sensitivity analyses when faced with challenging settings with limited available data. For example, use of weakly informative priors can avoid issues of extreme overfitting to the data. Additionally, it can be difficult to reliably estimate the heterogeneity parameters for a hierarchical model when the number of groups (e.g., tumour types) is very small. In these situations, multiple prior distributions can be used to assess how sensitive conclusions are to assumptions about the degree of heterogeneity in outcomes across groups.

Beyond applications to basket trials, BHMs have been used to incorporate disparate data sources, structural assumptions, and borrowing approaches when no single source is sufficient for inference and decision-making. For example, Heeg et al. (2022) recently used BHMs to partially pool information on specific model parameters across a class of immune-oncology therapies to improve survival extrapolations from immature data.

Can we incorporate non-randomized studies into meta-analyses while mitigating risk of bias to address challenges in assessing comparative efficacy/effectiveness?

Meta-analyses which synthesize the published evidence on relative treatment effects generally rely on RCT evidence only. However, when estimating real world effectiveness or efficacy/effectiveness in key patient populations of interest, or when RCT evidence is lacking due to the rarity of some indications, incorporating information from non-randomized studies (NRS) using real-world data becomes appropriate (Sarri et al., 2022). Relevant NRS would include cohort studies comparing patient outcomes by treatment using appropriate methods to mitigate sources of bias (Faria et al., 2015)—particularly well-designed synthetic control arm analyses (Thorlund et al., 2020) and target trial emulations (Hernán and Robins, 2016). Sarri et al. (2022) provide a structured framework for incorporating NRS into meta-analyses—a process which includes assessing risk of bias in the identified NRS and careful selection of methods to appropriately down-weight the influence of NRS, to incorporate bias adjustments, and to conduct sensitivity analyses to the modelling decisions.

Several promising approaches exist for incorporating NRS into a network meta-analysis (NMA) or pairwise meta-analysis which are

both conducive to down-weighting the NRS either statically or dynamically, and to probabilistic sensitivity analysis. Schmitz et al. (2013) highlight three approaches to incorporating NRS: 1) naïve pooling of the RCT and NRS evidence, 2) incorporation of the NRS using informative priors, and 3) use of a hierarchical model to capture the potential heterogeneity in relative treatment effects between RCT and NRS. They also outline how corrections for systematic and non-systematic bias can be incorporated into approaches 2) and 3).

Schmitz et al. (2013) note that the bias in NRS relative treatment effects (e.g., log-odds ratios, log-hazard ratios, *etc.*) can be modelled using a Gaussian distribution where the mean and variance represent systematic and non-systematic components of the bias, respectively. This allows for incorporation of NRS data into the meta-analysis with potential bias adjustment and down-weighting—either by means of a bias-adjusted priors or through direct incorporation into the likelihood. Efthimiou et al. (2017) highlight additional approaches than can be taken to form priors from NRS data—such as down-weighting of the parameter likelihood from the NRS data by means of a power prior or mixture prior (e.g., robust MAP priors). Verde et al. propose a hierarchical meta-regression (HMR) approach which can be used to estimate a bias-correction term for study design or other study-level covariates, and detect and down-weight outlier studies when there is significant cross-study heterogeneity (Verde et al., 2016; Verde, 2017; Verde, 2019). Additionally, HMRs can be used to extrapolate treatment effects to specific populations when IPD is available for at least one study or real-world data source.

Discussion

As new drug development is focusing more and more on narrower indications, HEOR practitioners are increasingly faced with challenges of limited data. These limitations arise from difficulties recruiting enough patients to conduct adequately powered RCTs (leading to more reliance on single-arm trials for regulatory and HTA submissions), narrowing of indications or subpopulations of interest leading to smaller numbers of relevant studies being identified in systematic literature reviews (and greater risk of disconnected or tenuous networks in NMAs), and more reliance on evidence from ITCs that are unlikely to yield precise estimates of relative treatment effects. Consequently, we present several recommendations for how Bayesian methods (including many of the approaches outlined above) can be used to help mitigate some of these pitfalls.

Firstly, since Bayesian approaches allow for weakly informative priors to be specified before analyzing the data, use of sensible default priors can mitigate some of the risks of model overfitting when data are very limited without imposing overly strong assumptions. With weakly informative defaults, the prior can be easily overwhelmed when informative data are available. An example of this can be found in the Keefe et al. (2021) meta-analysis of diagnostic tests where use of weakly informative priors allows for the meta-analysis to be run even when the model is overparametrized for some classes of diagnostic tests (too few studies relative to the number of parameters). In these cases, the prior is minimally updated (or not updated at all) and continues to reflect agnostic beliefs as to whether the test is predictive. In cases where more studies are available, the prior is updated to reflect the larger evidence base.

Secondly, if strong modelling assumptions are needed to synthesize the limited amount of available data, probabilistic sensitivity analyses should be conducted to assess robustness to deviations from these assumptions. For example, if it is infeasible to conduct a random effects NMA due to too few studies in the network, different heterogeneity assumptions can be assessed by fitting modified random effects NMAs in which different strong priors are used for the heterogeneity parameters, each reflecting a plausible scenario. In this context, fixed effects NMA can be viewed as a special case of random effects NMA, and use of informative priors on heterogeneity parameters allows for sensitivity analysis even when data are too limited to estimate these parameters.

Lastly, if precision/power are anticipated to be extremely limited (e.g., in a rare disease setting), it may be worth considering a context-appropriate decision rule rather than a default p-value threshold. For example, if we are performing an ITC between two treatments that have received regulatory approval based on single-arm trials, and it is infeasible to conduct an adequately powered ITC, it may be sensible to prioritize reimbursement of one drug over the other based on the posterior probability of superiority (a quantity which is directly available in Bayesian inference). This would arguably constitute a "best-available evidence" standard in this example.

In summary, Bayesian methods provide a principled framework for quantifying the amount of evidence in favour of a particular conclusion, are well-suited to combining information from multiple data sources under various structural assumptions, and can facilitate probabilistic sensitivity analyses to probe these structural assumptions. For these reasons we believe that Bayesian methods should play an increasing role in health economics and outcomes research.

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

EM and AS both contributed to the conception, research, expressed opinions, drafting and review. All authors contributed to the article and approved the submitted version.

Funding

EM and AS are employees of Cytel and funding for open-access publication fees are to be provided by Cytel.

Conflict of interest

EM and AS are employees of Cytel and this study received funding from Cytel. The funder was not involved in the article's conception or research, the writing of this article or the decision to submit it for publication.

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

Berger, M. L., and Doban, V. (2014). Big data, advanced analytics and the future of comparative effectiveness research. *J. Comp. Eff. Res.* 3 (2), 167–176. doi:10.2217/cer.14.2

Best, N., Price, R. G., Pouliquen, I. J., and Keene, O. N. (2021). Assessing efficacy in important subgroups in confirmatory trials: An example using Bayesian dynamic borrowing. *Pharm. Stat.* 20 (3), 551–562. doi:10.1002/pst.2093

Dias, S., Welton, N. J., Sutton, A. J., and Ades, A. E. (2011). NICE DSU technical support document 2: A generalised linear modelling framework for pairwise and network meta-analysis of randomised controlled trials. Available from http://www.nicedsu.org.uk/ (Accessed April, 2014).

Efthimiou, O., Mavridis, D., Debray, T. P., Samara, M., Belger, M., Siontis, G. C., et al., and GetReal Work Package 4 (2017). Combining randomized and non-randomized evidence in network meta-analysis. *Statistics Med.* 36 (8), 1210–1226. doi:10.1002/sim.7223

Faria, R., Alava, M. H., Manca, A., and Wailoo, A. J. (2015). NICE DSU technical support document 17: The use of observational data to inform estimates of treatment effectiveness in technology appraisal: Methods for comparative individual patient data. Available from http://www.nicedsu.org.uk/.

Gelman, A., Carlin, J. B., Stern, H. S., Dunson, D. B., Vehtari, A., and Rubin, D. B. (2013). "Hierarchical models," in *Bayesian data analysis*. Editors A. Gelman, J. B. Carlin, H. S. Stern, D. B. Dunson, A. Vehtari, and D. B. Rubin Third edition (Boca Raton, FL: CRC Press), 101–138.

Heeg, B., Verhoek, A., Tremblay, G., Harari, O., Soltanifar, M., Chu, H., et al. (2022). Bayesian hierarchical model-based network meta-analysis to overcome survival extrapolation challenges caused by data immaturity. *J. Comp. Eff. Res.* 12, e220159. doi:10.2217/cer-2022-0159

Hernán, M. A., and Robins, J. M. (2016). Using big data to emulate a target trial when a randomized trial is not available. *Am. J. Epidemiol.* 183 (8), 758–764. doi:10.1093/aje/kwv254

Ibrahim, J. G., and Chen, M. H. (2000). Power prior distributions for regression models. *Stat. Sci.* 15, 46–60. doi:10.1214/ss/1009212673

Ibrahim, J. G., Chen, M. H., Gwon, Y., and Chen, F. (2015). The power prior: Theory and applications. *Statistics Med.* 34 (28), 3724–3749. doi:10.1002/sim.6728

Keefe, D. T., Kim, J. K., Mackay, E., Chua, M., Van Mieghem, T., Yadav, P., et al. (2021). Predictive accuracy of prenatal ultrasound findings for lower urinary tract obstruction: A systematic review and bayesian meta-analysis. *Prenat. Diagn.* 41 (9), 1039–1048. doi:10.1002/pd.6025

Mackay, E., Springford, A., Nagamuthu, C., and Dron, L. (2022). MSR46 A bayesian hierarchical modelling approach for indirect comparison of response outcomes in histology-independent therapies. *Value health* 25 (12), S358–S359. doi:10.1016/j.jval.2022.09.1777

Mackay, E., Springford, A., Nagamuthu, C., Dron, L., and Dias, S. (2023). MSR73 Bayesian hierarchical models for indirect treatment comparisons of histology-independent therapies for survival outcomes. *Value health* 26, S290. [Abstract] Forthcoming. doi:10.1016/j.jval.2023.03.1608

Murphy, P., Claxton, L., Hodgson, R., Glynn, D., Beresford, L., Walton, M., et al. (2021). Exploring heterogeneity in histology-independent technologies and the implications for cost-effectiveness. *Med. Decis. Mak.* 41 (2), 165–178. doi:10.1177/0272989X20980327

Murphy, P., Glynn, D., Dias, S., Hodgson, R., Claxton, L., Beresford, L., et al. (2022). Modelling approaches for histology-independent cancer drugs to inform NICE appraisals: A systematic review and decision-framework. *Health Technol. Assess.* 25, 1–228. doi:10.3310/hta25760

Neuenschwander, B., Branson, M., and Spiegelhalter, D. J. (2009). A note on the power prior. *Statistics Med.* 28 (28), 3562–3566. doi:10.1002/sim.3722

Neuenschwander, B., Capkun-Niggli, G., Branson, M., and Spiegelhalter, D. J. (2010). Summarizing historical information on controls in clinical trials. *Clin. trials* 7 (1), 5–18. doi:10.1177/1740774509356002

Neuenschwander, B., Wandel, S., Roychoudhury, S., and Bailey, S. (2016). Robust exchangeability designs for early phase clinical trials with multiple strata. *Pharm. Stat.* 15 (2), 123–134. doi:10.1002/pst.1730

Richeldi, L., Azuma, A., Cottin, V., Hesslinger, C., Stowasser, S., Valenzuela, C., et al. (2022). Trial of a preferential phosphodiesterase 4B inhibitor for idiopathic pulmonary fibrosis. *N. Engl. J. Med.* 386 (23), 2178–2187. doi:10.1056/NEJMoa2201737

Sarri, G., Patorno, E., Yuan, H., Guo, J. J., Bennett, D., Wen, X., et al. (2022). Framework for the synthesis of non-randomised studies and randomised controlled trials: A guidance on conducting a systematic review and meta-analysis for healthcare decision making. *BMJ evidence-based Med.* 27 (2), 109–119. doi:10.1136/bmjebm-2020-111493

Schmidli, H., Gsteiger, S., Roychoudhury, S., O'Hagan, A., Spiegelhalter, D., and Neuenschwander, B. (2014). Robust meta-analytic-predictive priors in clinical trials with historical control information. *Biometrics* 70 (4), 1023–1032. doi:10.1111/biom.12242

Schmitz, S., Adams, R., and Walsh, C. (2013). Incorporating data from various trial designs into a mixed treatment comparison model. *Statistics Med.* 32 (17), 2935–2949. doi:10.1002/sim.5764

Spiegelhalter, D. J., Myles, J. P., Jones, D. R., and Abrams, K. R. (1999). Methods in health service research. An introduction to bayesian methods in health technology assessment. *BMJ* 319 (7208), 508–512. doi:10.1136/bmj.319.7208.508

Thorlund, K., Dron, L., Park, J. J., and Mills, E. J. (2020). Synthetic and external controls in clinical trials–a primer for researchers. *Clin. Epidemiol.* 12, 457–467. doi:10.2147/CLEP.S242097

UK National Institute for Health and Care Excellence (2020). Larotrectinib for treating NTRK fusion-positive solid tumours: Technology appraisal guidance. Available from https://www.nice.org.uk/guidance/ta630/resources/larotrectinib-for-treating-ntrk-fusionpositive-solid-tumours-pdf-82609071004357.

US Food and Drug Administration (2018). BLA 125370/s-064 and BLA 761043/s-007 multi-disciplinary review and evaluation Benlysta® (belimumab) for intravenous infusion in children 5 to 17 years of age with SLE. Available from https://www.fda.gov/media/127912/download.

Verde, P. E., Ohmann, C., Morbach, S., and Icks, A. (2016). Bayesian evidence synthesis for exploring generalizability of treatment effects: A case study of combining randomized and non-randomized results in diabetes. *Statistics Med.* 35 (10), 1654–1675. doi:10.1002/sim.6809

Verde, P. E. (2017). "Two examples of Bayesian evidence synthesis with the hierarchical meta-regression approach," in *Bayesian inference* (InTech). doi:10.5772/intechopen.70231

Verde, P. E. (2019). The hierarchical metaregression approach and learning from clinical evidence. *Biometrical J.* 61 (3), 535–557. doi:10.1002/bimj.201700266

Viele, K., Berry, S., Neuenschwander, B., Amzal, B., Chen, F., Enas, N., et al. (2014). Use of historical control data for assessing treatment effects in clinical trials. *Pharm. Stat.* 13 (1), 41–54. doi:10.1002/pst.1589

frontiers | Frontiers in Pharmacology

TYPE Original Research
PUBLISHED 03 October 2023
DOI 10.3389/fphar.2023.1211491

Check for updates

OPEN ACCESS

EDITED BY
Amr Makady,
Janssen Pharmaceutica NV, Belgium

REVIEWED BY
Patrick P. T. Jeurissen,
Radboud University Medical Centre, Netherlands
Peter Stopfer,
Boehringer Ingelheim, Germany
Xiujuan Lei,
Shaanxi Normal University, China

*CORRESPONDENCE
You Chen,
you.chen@vanderbilt.edu

RECEIVED 24 April 2023
ACCEPTED 18 September 2023
PUBLISHED 03 October 2023

CITATION
Jeong E, Malin B, Nelson SD, Su Y, Li L and Chen Y (2023), Revealing the dynamic landscape of drug-drug interactions through network analysis.
Front. Pharmacol. 14:1211491.
doi: 10.3389/fphar.2023.1211491

COPYRIGHT
© 2023 Jeong, Malin, Nelson, Su, Li and Chen. This is an open-access article distributed under the terms of the Creative Commons Attribution License (CC BY). The use, distribution or reproduction in other forums is permitted, provided the original author(s) and the copyright owner(s) are credited and that the original publication in this journal is cited, in accordance with accepted academic practice. No use, distribution or reproduction is permitted which does not comply with these terms.

Revealing the dynamic landscape of drug-drug interactions through network analysis

Eugene Jeong[1], Bradley Malin[1,2,3], Scott D. Nelson[1], Yu Su[4], Lang Li[5] and You Chen[1,3]*

[1]Department of Biomedical Informatics, School of Medicine, Vanderbilt University Medical Center, Nashville, TN, United States, [2]Department of Biostatistics, School of Medicine, Vanderbilt University Medical Center, Nashville, TN, United States, [3]Department of Computer Science, School of Engineering, Vanderbilt University, Nashville, TN, United States, [4]Department of Computer Science and Engineering, College of Engineering, The Ohio State University, Columbus, OH, United States, [5]Department of Biomedical Informatics, College of Medicine, The Ohio State University, Columbus, OH, United States

Introduction: The landscape of drug-drug interactions (DDIs) has evolved significantly over the past 60 years, necessitating a retrospective analysis to identify research trends and under-explored areas. While methodologies like bibliometric analysis provide valuable quantitative perspectives on DDI research, they have not successfully delineated the complex interrelations between drugs. Understanding these intricate relationships is essential for deciphering the evolving architecture and progressive transformation of DDI research structures over time. We utilize network analysis to unearth the multifaceted relationships between drugs, offering a richer, more nuanced comprehension of shifts in research focus within the DDI landscape.

Methods: This groundbreaking investigation employs natural language processing, techniques, specifically Named Entity Recognition (NER) via ScispaCy, and the information extraction model, SciFive, to extract pharmacokinetic (PK) and pharmacodynamic (PD) DDI evidence from PubMed articles spanning January 1962 to July 2023. It reveals key trends and patterns through an innovative network analysis approach. Static network analysis is deployed to discern structural patterns in DDI research, while evolving network analysis is employed to monitor changes in the DDI research trend structures over time.

Results: Our compelling results shed light on the scale-free characteristics of pharmacokinetic, pharmacodynamic, and their combined networks, exhibiting power law exponent values of 2.5, 2.82, and 2.46, respectively. In these networks, a select few drugs serve as central hubs, engaging in extensive interactions with a multitude of other drugs. Interestingly, the networks conform to a densification power law, illustrating that the number of DDIs grows exponentially as new drugs are added to the DDI network. Notably, we discovered that drugs connected in PK

Abbreviations: A, Alimentary tract and metabolism; ATC, Anatomical Therapeutic Chemical; B, Blood and blood forming organs; C, Cardiovascular system; CYP, Cytochrome P450; D, Dermatologicals; DDI, Drug-Drug Interaction; DPL, Densification Power Law; G, Genito urinary system and sex hormones; H, Systemic hormonal preparations, excluding sex hormones and insulins; J, Antiinfective for systemic use; L, Antineoplastic and immunomodulating agents; M, Musculo-skeletal system; N, Nervous system; NA, Network Analysis; NER, Name Entity Recognition; NLP, Natural Language Processing; P, Antiparasitic products, insecticides, and repellents; PD, Pharmacodynamic; PK, Pharmacokinetic; R, Respiratory system; RE, Relation Extraction; S, Sensory organs; SOTA, State-Of-The-Art; T5, Text-To-Text Transfer Transformer; UMLS, Unified Medical Language System; V, Various; WHO, World Health Organization.

and PD networks predominantly belong to the same categories defined by the Anatomical Therapeutic Chemical (ATC) classification system, with fewer interactions observed between drugs from different categories.

Discussion: The finding suggests that PK and PD DDIs between drugs from different ATC categories have not been studied as extensively as those between drugs within the same categories. By unearthing these hidden patterns, our study paves the way for a deeper understanding of the DDI landscape, providing valuable information for future DDI research, clinical practice, and drug development focus areas.

KEYWORDS

pharmacokinetic drug-drug interaction, pharmacodynamic drug-drug interaction, network analysis, natural language Processing, research trend

1 Introduction

Drug-drug interactions (DDIs) occur when the effect of one drug is altered by the presence of another drug (van Mil, 2016). DDIs can be broadly classified into two types: 1) pharmacokinetic (PK), which occurs when one drug modifies the disposition (i.e., absorption, distribution, metabolism, and/or excretion) of another drug (Nebert and Russell, 2002; Nigam, 2015), and 2) pharmacodynamic (PD), which occur when the pharmacological effects (on cells, organs, and systems) of one drug are altered or additive by the presence of another (Niu et al., 2019). These interactions can generate a wide range of outcomes, often causing adverse effects and deteriorating patients' health. Consequently, DDIs have been the subject of numerous studies over the past several decades, with progress in high-throughput screening methods, the rapid growth of biomedical databases, and an increase in clinical studies contributing to the discovery of novel DDIs and insights into their underlying PK and PD mechanisms (Becker et al., 2007).

The vast amount of data generated by the numerous studies on DDIs has made it challenging for researchers to analyze research trends and evolutions, which makes it difficult to gain a comprehensive understanding of the overall landscape of DDIs, identify under-explored areas, discern research trends, and pinpoint areas of focused interest. To address this issue, some studies have used bibliometric analysis (Wang et al., 2022; Sun et al., 2022; Pirri et al., 2020; KURUTKAN, 2023), a quantitative method that evaluates and analyzes various aspects of scientific publications. Bibliometric indicators such as the number of publications, citations, and authors can provide a valuable quantitative overview of DDI research. However, this approach has limitations in its ability to capture the complex relationships between drugs and the evolving nature of DDI research, despite its numerical precision and ease of use.

To thoroughly examine the DDI research landscape, we constructed DDI networks based on evidence extracted from PubMed article abstracts by natural language processing (NLP) models and analyzed them using network analysis (NA). NLP models facilitate the automation of information extraction from extensive unstructured text data, enabling researchers to analyze large datasets more quickly and efficiently (Boyce et al., 2012). Network analysis, on the other hand, serves as a powerful model for analyzing complex interactions between drugs, providing a more comprehensive picture of the structure and allowing researchers to represent and explore complex data in a more intuitive and accessible way (Jeong et al., 2017; Chen et al., 2020; Yan et al., 2021).

By utilizing DDI networks, we can gain a complete understanding of the DDI research landscape and the chronological development of the field. This approach provides a comprehensive view of the dynamic landscape of drug-drug interactions and allows for the identification of shifts in the DDI landscape. Our integration of NLP and NA allows researchers to identify areas of focused interest and under-explored areas, recognize emerging areas of concern or novel research trajectories, and spot gaps in the field that may harbor potential yet under-studied drug interactions. Ultimately, this approach may inform decision-making in drug development, clinical practice, and DDI research prioritization.

2 Materials and methods

2.1 Retrieving DDI evidence from PubMed abstracts

We applied a three-step procedure to collect evidence on DDIs from PubMed abstracts published between January 1962 and July 2023: 1) identification of candidate articles about DDIs using a PubMed query (Figure 1A), 2) screening of the articles containing sentences with drug entities using a named entity recognition (NER) model (Figure 1B), and 3) determination of eligible sentences about DDIs using a relation extraction (RE) model (Figure 1C).

2.1.1 PubMed query

We designed a query in accordance with Duda et al. (2005) to retrieve a set of DDI articles with high sensitivity, the broadest search to include all DDI-relevant articles: "drug interactions" [TIAB] OR "drug interactions" [MeSH Terms] OR "drug interaction" [TW] NOT food-drug interactions [MeSH Terms] NOT herb-drug interactions [MeSH Terms] NOT Review [PT] NOT Systematic Review [PT]. This query was chosen to ensure that no pertinent documents were missed.

2.1.2 NER model

In order to search for evidence of DDIs in retrieved articles, it is necessary to first identify drug entities within sentences. To accomplish this, a NER model, a type of NLP model, that is, used to identify and extract entities, is required for the efficient and accurate detection of drug entities.

The SpaCy Python library is an open-source library designed to support a variety of tasks such as POS Tagging, NER, and Dependency Parsing (Honnibal and Montani, 2017). ScispaCy (Neumann et al., 2019) is an extension of spaCy developed for biomedical, scientific, or

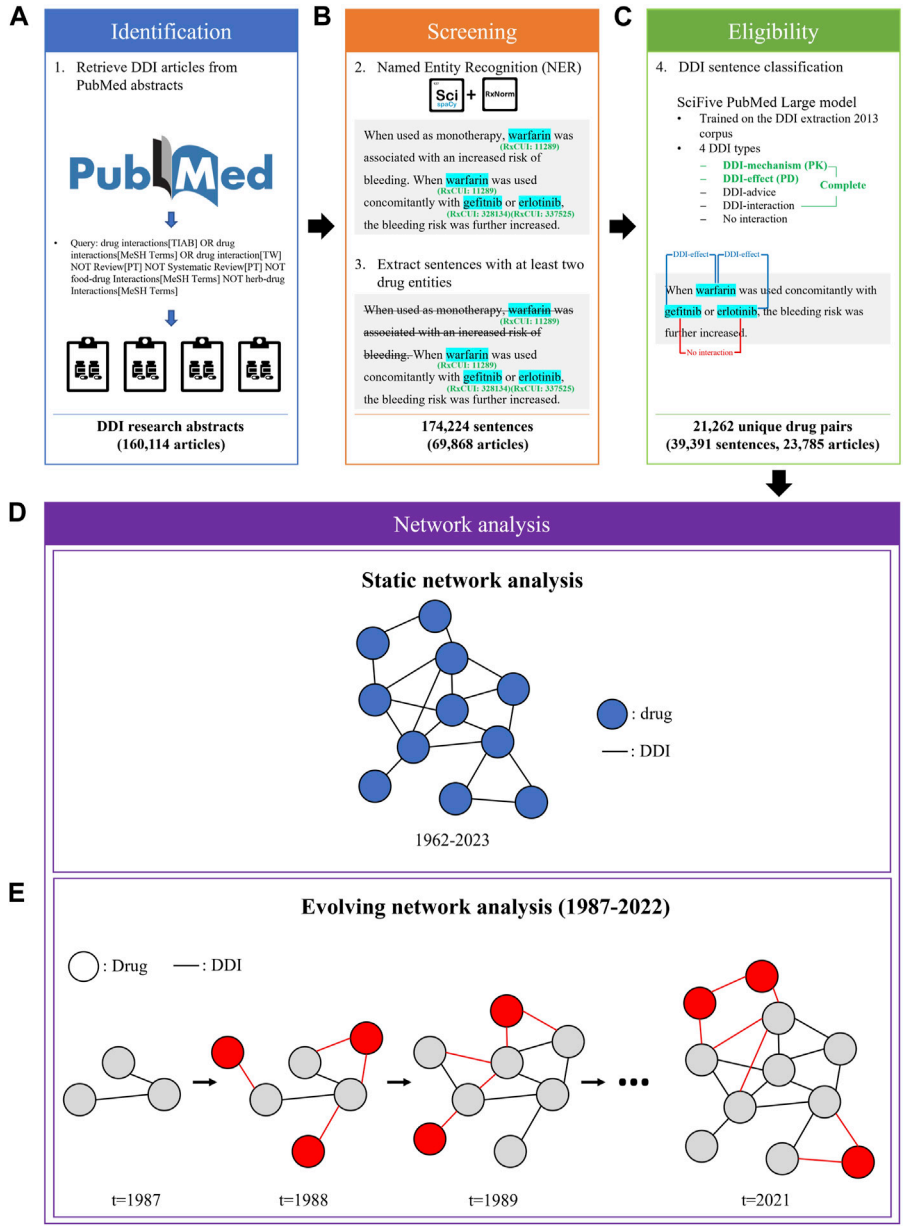

FIGURE 1
The process of DDI evidence extraction and dynamic network analysis. **(A)** A custom query was employed to retrieve articles related to DDIs from PubMed API. **(B)** Sentences with at least two drug entities from abstracts were extracted using the ScispaCy NER model. **(C)** The SciFive DDI RE model was applied to extract DDIs from DDI sentences. **(D)** A static network was constructed based on the entire extracted DDI sentences from 1962 to 2023. **(E)** An evolving network analysis was conducted to examine the 36-year trend in DDI research.

clinical text. It has become the *de facto* standard for practical NLP due to its speed, reliability, and near-state-of-the-art performance (SOTA). Entity linker is a SciSpacy feature that maps entities mentioned in the text to standard, canonical identifiers in a knowledge base or database. These databases for biomedical texts could include UMLS (Unified Medical Language System), RxNorm, and others. The linker conducts a string overlap-based search (char-3grams) on named entities, comparing them with the concepts in a knowledge base via an approximate nearest neighbors search. We implemented RxNorm entity linker in the ScispaCy, which contains ~100 k concepts focused on normalized names for clinical drugs. The ScispaCy-large model was used to perform NER, sentence tokenization, and entity-linking features for every sentence from abstracts. Given that the ScispaCy only provides CUI for RxNorm entities, we used the MRCONSO.RRF file from UMLS Metathesaurus (Bodenreider, 2004) to map CUI to RxCUI. To minimize the chance of duplicating clinically similar RxNorm concepts, we further linked RxNorm concepts to RxNorm ingredients using the RXNREL.RRF file from UMLS Metathesaurus. Finally, we extracted sentences containing at least two drug entities for further analysis of potential DDIs.

2.1.3 RE model

To identify eligible evidence of DDIs from sentences containing at least two drug entities, a RE model, a type of NLP model, that is, used to

identify and extract the relationships between entities in text, was utilized to find the DDI relationship within the sentences. The SciFive PubMed Large model is a domain-specific text-to-text transfer transformer (T5) model (Raffel et al., 2020) that is, pre-trained on PubMed abstracts using 1.2 million steps to optimize the pre-trained weights from the T5 model in the context of biomedical literature. The DDI extraction 2013 corpus is comprised of 792 texts taken from the DrugBank database and 233 Medline papers; it has been created for the SemEval 2013 DDIExtraction challenge, whose primary objective is to provide a common framework for the evaluation of information extraction techniques applied to the recognition of pharmacological substances and the detection of DDIs from biomedical texts, and has been used as the gold standard for evaluating DDI extraction task performance (Segura-Bedmar et al., 2013). Two expert pharmacists with extensive experience in pharmacovigilance annotated drug-drug interactions, covering both pharmacokinetic and pharmacodynamic interactions. The five classifications consist of four distinct types of interactions and one type of non-interaction in the corpus, as follows: 1) No interaction: a sentence does not represent an interaction between two drugs, 2) DDI-mechanism: a sentence describes a pharmacokinetic mechanism, 3) DDI-effect: a sentence describes the effect of the DDI or pharmacodynamic mechanism, 4) DDI-advice: a sentence describes a recommendation or advice regarding a drug interaction, and 5) DDI-int: a sentence describes a drug interaction without providing any other information. The SciFive PMC Large model achieved a level of performance that was similar to SOTA on DDI relation extraction using the DDI extraction 2013 corpus (precision: 83.88, recall: 83.45, and F1 score: 83.63). We applied the pre-trained weights of the SciFive PMC Large model distributed by the authors (Phan et al., 2021), and further fine-tuned the model parameters using the DDI extraction 2013 corpus to determine the reliability of each candidate DDI evidence. If a candidate DDI sentence contained more than two RxNorm ingredients, all possible drug-drug combinations were investigated, implying that a single sentence could contain both a drug pair that does not interact and a drug pair that does interact.

To validate the performance of the SciFive model, we randomly selected extracted DDI evidence and manually annotated them with the help of two reviewers (one with an M.S. in biomedical informatics and one with a Ph.D. in computer science). Both reviewers had 3 years of experience in drug-interaction research. The level of agreement between the two reviewers was measured using Cohen's Kappa.

2.1.4 Mapping RxNorm ingredients to ATC first levels

RxNorm ingredients were mapped to first-level Anatomical Therapeutic Chemical (ATC) classes using the RxNorm API (https://mor.nlm.nih.gov/RxNav/) for drug classification purposes. The ATC first level contains 14 major anatomical or pharmacological groups. If a RxNorm ingredient had multiple ATC first levels, then all ATCs were counted separately. If a RxNorm ingredient was unmapped to any ATC first levels, then it was assigned to a "No ATC" group.

2.2 Network construction

2.2.1 Static networks

Based on all extracted DDI sentences from 1962 to 2023, we constructed three static networks: 1) one for PKs, 2) one for PDs, and 3) one for the complete set of DDIs, including PK and PD, as well as those classified as DDI-advice or DDI-int (Figure 1D). In such networks, each node represents a drug and each edge exists between two nodes if there was at least one sentence from the literature with evidence of a DDI between the two drugs.

2.2.2 Evolving networks

To model the dynamic changes in the DDI networks, we created evolving networks of drugs based on DDIs extracted from each year (Figure 1E). The network T_{i+1} is an augmentation of the prior network T_i, where i represents the year. For example, the network of 1988 represented the addition of new drugs and DDIs published in 1988 to the network of 1987. Similarly, the network of 1989 expanded upon the 1988 network, and this pattern continued in subsequent years. Due to a lack of sufficient data to create yearly networks for years prior to 1987, we chose 1987 as the earliest investigated year for the evolving network analysis. In addition, we have excluded 2023 data from the evolving network analysis due to lacking data for the entire year.

2.3 Network-level properties

In order to provide a more comprehensive understanding of the structure of DDI research, we measured various network structural properties in this study. These properties included the number of nodes and edges, assortativity based on degree and ATC first level categories, average local clustering coefficient, power law exponent γ, network diameter, and the densification power law (DPL). The number of nodes and edges was specifically measured to gain insights into the size of the networks. The degree assortativity is the tendency for nodes of high degree (resp. low degree) in a graph to be connected to high degree nodes (resp. to low degree ones), while ATC-group assortativity is the tendency for nodes to be connected to drugs in the same ATC categories. The average local clustering coefficient measures how close its neighbors are to form a clique. If the neighborhood is fully connected, the clustering coefficient is 1, whereas a value close to 0 indicates that the neighborhood has few connections. The diameter of a network is defined as the smallest distance between the two furthest nodes in the network. This distance is determined by computing the shortest path length between every node and all other nodes and selecting the longest path length as the network's diameter. A smaller network diameter suggests that the drugs in the network are more closely related and may have a higher potential for interactions, while a larger diameter may indicate that the drugs are more diverse and less likely to interact. To determine whether the number of edges grows faster than the number of nodes in the networks, we measured the DPL. The DPL is a concept from the temporal graph evolution (Leskovec et al., 2007) domain. This law indicates that the number of edges should grow according to a power law over the number of nodes over time:

$$\boldsymbol{e(t) \propto n(t)^a} \tag{1}$$

where $e(t)$ and $n(t)$ denote the number of edges and nodes, respectively, of the graph at time t, and a is an exponent (a = 1 represents a constant average degree throughout time, whereas a = 2 represents to an extremely dense graph in which each node has edges to a constant fraction of all nodes on average.) Numerous studies have shown that many real-world evolving networks exhibit a densification power law property (Leskovec et al., 2005; Leskovec

TABLE 1 Two-by-two contingency table for evaluating ATC 1-ATC 2 pairs.

	ATC category #2	No ATC category #2
ATC category #1	a	b
No ATC category #1	c	d

TABLE 2 Structural network properties of the static PK, PD, and complete DDI networks.

Structural network property	PK	PD	Complete
Nodes	1,620	2,011	2,212
Edges	7,579	15,676	21,262
Assortativity (Degree)	−0.151	−0.0754	−0.124
Assortativity (ATC 1st level)	0.087	0.123	0.111
Power law exponent	2.5	2.82	2.36
Avg clustering coefficient	0.23	0.26	0.290
Diameter	9	8	8

et al., 2007; Qu et al., 2014; Qu et al., 2015). Network analysis was conducted using the *igraph* package in R (Csardi and Nepusz, 2006).

2.4 ATC categories-level properties

Our analysis focuses on DDI networks at the level of ATC classification groups. We aim to determine whether the observed DDI interactions occur within the same therapeutic class or across multiple classes. This approach allows us to investigate the potential for interactions between drugs with similar or different mechanisms of action and may provide insights into the overall safety and efficacy of drug combinations within specific therapeutic categories.

2.4.1 The Krackhardt E/I ratio

The Krackhardt E/I Ratio (Krackhardt and Stern, 1988), also known as the E-I index, is a measure of homophily that quantifies the extent to which one node is linked to similar or dissimilar nodes. The E-I index is computed as:

$$\mathbf{E-I}\ \boldsymbol{index} = \frac{\mathbf{EL-IL}}{\mathbf{EL+IL}} \quad (2)$$

where EL and IL denote the number of external links and internal links, respectively. The E-I index ranges from −1 to 1, and if it is positive, it indicates that there are more external links than internal links (heterophily). If the value is close to 0, it indicates that links are distributed equally; and if it is negative, it indicates that there are more internal links than external links (homophily).

2.4.2 Fisher's exact test for ATC-ATC pairs

To determine the most interconnected pairs of ATC categories (those with a higher chance of having DDIs between drugs from the two categories compared to other categories), all possible ATC category-ATC category combinations were extracted from the network and generated in the 2-by-2 contingency table (Table 1). Numbers are assigned to one of the contingency table cells based on the number of interactions between ATC categories. For example, *a* denotes the number of interactions between the #1 category and #2 category, and b denotes the number of interactions that the #1 category has with ATC categories other than the #2 category. A Fisher's exact test with Bonferroni correction was relied upon to determine significance. The ATC-ATC pairs with *p*-values less than 0.05 after Bonferroni correction and odds ratios greater than 1 were considered statistically significant.

2.5 Drug-level properties

While ATC-group-level analyses examine classes of drugs, drug-level analyses focus on individual drugs. This approach provides a more detailed understanding of specific drug interactions and is essential for identifying key drugs in the DDI network. By examining the interactions of individual drugs, we can gain insights into the mechanisms of action that underlie drug interactions and identify drugs that are more likely to be involved in multiple interactions.

2.5.1 Centrality measures

In network analysis, several types of centrality measures can be used to understand the relative importance of drugs within the DDI network. In this study, we concentrated on three centrality measures: degree centrality, betweenness centrality, and eigenvector centrality. The degree centrality is a simple centrality measure that counts how many neighbors a drug has, finding drugs that are likely to be the center and can quickly connect with the wider network. The betweenness centrality measures the number of times a drug lies on the shortest path between other drugs. This measure shows which drugs are bridges between drugs in a network, showing drugs that influence the flow in the DDI network. Eigenvector centrality measures a drug's influence based on the number of links it has to other drugs in the network. A high eigenvector score means that a drug is connected to many drugs that themselves have high scores.

2.5.2 Emerging and declining drugs in the DDI research field

To identify drugs that have recently emerged in DDI research, we calculated the degree, betweenness, and eigenvector centrality growth rates for each drug in the yearly networks over the past 5 years (2018–2022). Drugs with a rapid growth rate are likely to be part of a new trend, attracting increased attention in recent years. The growth rate (slope) was estimated using linear regression. To identify the drugs that are receiving less attention in DDI research, we analyzed the lowest increase or highest decrease rate in centrality measures.

3 Results

3.1 The DDI sentences extracted from PubMed

We retrieved 160,114 candidate articles from the PubMed API through a search query designed for high sensitivity. Next, we

applied ScispaCy for NER and extracted 174,224 sentences (69,868 articles) that contained at least two drug entities from the abstracts of the DDI articles. Finally, we used the SciFive model to extract 2,212 unique drugs and 21,262 unique DDIs (PK: 7,579, PD: 15,676) from 174,224 sentences. Among the 21,262 unique DDIs, 2,445 exhibited both PK and PD DDIs (Supplementary Figures S1, S2). To validate the performance of the SciFive model, we randomly selected 1,296 DDIs (36 for each year from 1987 to 2022) from the 21,233 DDIs. The level of agreement between the two reviewers was found to be extremely high with $\kappa = 0.95$; $p < 0.001$. The SciFive model achieved F1 scores of 0.892. All DDI sentences are provided in Supplementary Table S1.

3.2 Static network analysis

3.2.1 An analysis of static DDI networks reveals scale-free structure, ATC category-based assortativity, and degree-based dissortativity

Table 2 shows the structural properties of the static DDI networks. All PK, PD, and complete DDI networks were scale-free ($2 < \gamma < 3$), with power law exponents (γ) of 2.56, 2.77, and 2.36. This indicates that a small number of drugs had many connections to other drugs, while most drugs had relatively few DDIs. Additionally, in the three networks, the ATC category-based assortativities were positive, suggesting that drugs from the same ATC category were more commonly investigated for DDIs than those from different ATC categories. Moreover, all three networks showed negative degree assortativity, meaning that few drugs were frequently confirmed to have DDIs with a large number of other drugs, each of which was rarely investigated to have a large number of DDIs.

3.2.2 The average clustering coefficients reveal a prevalence of real DDIs among neighbors in static networks

The average clustering coefficients for the PK, PD, and complete networks are 0.23, 0.26, and 0.29, respectively. These are significantly higher than the clustering coefficients [0.005 (0.003–0.008), 0.007 (0.004–0.009), and 0.008 (0.006–0.011)] of random networks generated by Erdős-Rényi algorithms with the same number of nodes and edges. We performed 100 random network simulations.

The larger clustering coefficients of the PK, PD, and complete DDI networks suggest that about 30% of the potential connections among a drug's neighbors in the network are actual DDIs. This means that, when examining the neighbors of a drug in the network, there is at least a 30% chance of finding a real DDI between them.

3.2.3 Network diameter: comparing DDI static networks to the six degrees of separation

The diameter of all three networks, ranging from 8 to 9 (Table 2), slightly exceeds the well-known six degrees of separation observed in our world (Kleinfeld, 2002). The six degrees of separation theory is a concept that suggests any two people on Earth are, on average, separated by no more than six social connections, indicating that networks are both extensive and interconnected. As more drugs and their DDIs are investigated and added to the network, there may be a possibility of reducing the diameter from its current range to 6.

3.2.4 ATC drug category E-I homophily index indicates a higher likelihood of DDIs within the same ATC drug category in static networks

The E-I homophily index values, which measure the degree to which a drug forms DDIs with others in the same category, were smaller than 1 (except for the V in the PD subnetwork) (Table 3). This suggests a tendency for drugs to establish connections with those belonging to the same group (homogeneous interactions).

3.2.5 Identifying ATC category pairs with the highest likelihood of DDIs in static networks

The J-D (Antiinfectives for systemic use—Dermatologicals), A-N (Alimentary tract and metabolism—Nervous system), and M-N (Musculo-skeletal system—Nervous system) pairs were significant and had the highest odds ratios in all three networks (Table 3). Supplementary Table S2 presents all significant ATC-ATC pairs in the static PK, PD, and complete DDI networks.

3.3 Evolving network analysis

3.3.1 Evolving power law exponent and assortativity indicate stable scale-free structure, ATC category-based assortativity, and degree-based dissortativity over time

Figure 2 depicts the properties of network evolution. The power law exponent (γ) of the PK DDI network remained stable between 2 and 3, while in the PD and complete DDI networks, γ fluctuated until 2001 but has stabilized between 2 and 3 since then (Figure 2A). The ATC category-based assortativities increased over time, suggesting a growing likelihood of DDIs among drugs within the same category (Figure 2B). The degree assortativities declined over time, indicating an increase in the dissortativity of the networks (Figure 2C).

3.3.2 Evolving network clustering coefficients indicate an increasing prevalence of real DDIs among neighbors

The average clustering coefficient gradually increased (within a range of [0,1]) in the PK and complete DDI networks, while it slightly decreased over time in the PD network. However, the clustering coefficients of the PD network were consistently higher than those in the PK and complete networks (Figure 2D). When simulating 100 times with random networks containing the same number of nodes and edges for each year, the average clustering coefficients were not only low but also declined as networks expanded. This finding suggests that DDI networks differ from random networks and evolve towards an increased likelihood of DDIs between neighboring drugs (Supplementary Figure S3).

3.3.3 Evolving network diameter narrows the gap to six degrees of separation

Despite the growth in network size over time (Figures 2E, F), the diameters, which represent the longest length of the shortest paths between any two drugs, have experienced a slight decrease, moving from a range of 8–9 to 7–8 (Figure 2G).

TABLE 3 Network properties at the level of ATC category in the static PK, PD, and complete DDI networks.

ATC 1st level code	Anatomical or pharmacological groups	PK DDI network				PD DDI network				Complete DDI network			
		Node	Edge	E-I index	Sig. pair[a] (OR)	Node	Edge	E-I index	Sig. Pair (OR)	Node	Edge	E-I index	Sig. Pair (OR)
A	Alimentary tract and metabolism	1,515	2,456	0.799	N (5.54)	2,495	4,596	0.797	N (2.57)	3,118	6,354	0.801	N (1.63)
B	Blood and blood forming organs	655	922	0.902	V (6.17)	1,369	2,096	0.846	D (2.23)	1,661	2,788	0.857	D (1.41)
C	Cardiovascular system	1,537	2,362	0.699	B (5.1)	2,579	4,482	0.702	H (2.11)	3,147	6,185	0.708	N (1.22)
D	Dermatologicals	688	1,007	0.904	J (5.23)	1,498	2,328	0.832	J (3.51)	1,813	3,061	0.851	J (2.02)
G	Genito urinary system and sex hormones	453	571	0.897	H (6.57)	895	1,543	0.898	A (1.69)	1,033	1,923	0.9	
H	Systemic hormonal preparations, excluding sex hormones and insulins	63	70	0.941	V (12.3)	118	152	0.96	B (2.1)	159	212	0.942	
J	Antiinfective for systemic use	652	1,054	0.676	P (8.4)	763	1,071	0.808	D (3.51)	1,112	2,025	0.746	S (1.98)
L	Antineoplastic and immunomodulating agents	806	1,217	0.719	S (5.43)	1,673	3,175	0.393	P (2.66)	1,940	4,066	0.496	P (1.66)
M	Musculo-skeletal system	332	430	0.928	N (5.39)	721	1,117	0.89	N (2.33)	873	1,425	0.906	N (1.7)
N	Nervous system	1,665	2,209	0.66	A (5.54)	2,879	4,864	0.657	A (2.57)	3,450	6,401	0.664	M (1.7)
P	Antiparasitic products, insecticides, and repellents	58	68	0.875	J (8.4)	176	219	0.904	V (3.83)	205	263	0.896	V (2.58)
R	Respiratory system	421	583	0.929	N (4.38)	807	1,239	0.887	C (1.76)	987	1,698	0.896	A (1.18)
S	Sensory organs	875	1,463	0.895	L (5.43)	1,677	2,738	0.833	J (3.34)	1,966	3,812	0.86	P (1.99)
V	Various	133	146	0.972	H (12.3)	358	482	1	P (3.83)	434	586	0.993	P (2.57)
No ATC	No ATC	349	408	0.95		800	1,140	0.942		990	1,463	0.944	

[a]The ATC-ATC, pair with the highest odds ratio and adjusted p-value <0.05.

3.3.4 Evolving node and edge counts indicate densification power law in DDI networks

We observed growth in the size of the three networks (in terms of node and edge counts) over the years (Figures 2E, F). The PK, PD, and complete DDI networks exhibited a densification power law with high densification exponents (1.84, 1.75, and 1.9, respectively), signifying that these networks become increasingly dense as they expand in size, thereby raising the likelihood of actual DDIs between two drugs (Figure 3).

We observed that the A, C (Cardiovascular system), and N categories consistently had the largest number of nodes (drugs) and edges (DDIs) from 1987 to 2022 (Figures 4A, B). In contrast, the P (Antiparasitic products, insecticides, and repellents) and H (Systemic hormonal preparations, excluding sex hormones and insulins) categories exhibited the smallest number of nodes and edges during the same period. Notably, the number of drugs and DDIs within the L (Antineoplastic and immunomodulating agents) category experienced exponential growth since 2007, while the size of all other ATC categories remained stable, exhibiting a steady growth rate over the years.

3.3.5 Trends in ATC drug category E-I index scores and their implications

The E-I index scores remained below 1 throughout the years (Figure 4C). A significant decrease in the E-I index of the L category was observed in the PD and complete DDI networks, suggesting an increased focus on investigating DDIs within the L category, rather than those involving drugs from L and other categories. The E-I index score for the J category showed a marked downward trend in the PK network. In all three networks, the E-I index scores for the C category experienced the highest growth rate over the years, indicating that the number of DDIs between drugs from the C category and other categories has been increasing more rapidly than the number of DDIs between drugs within the C category itself.

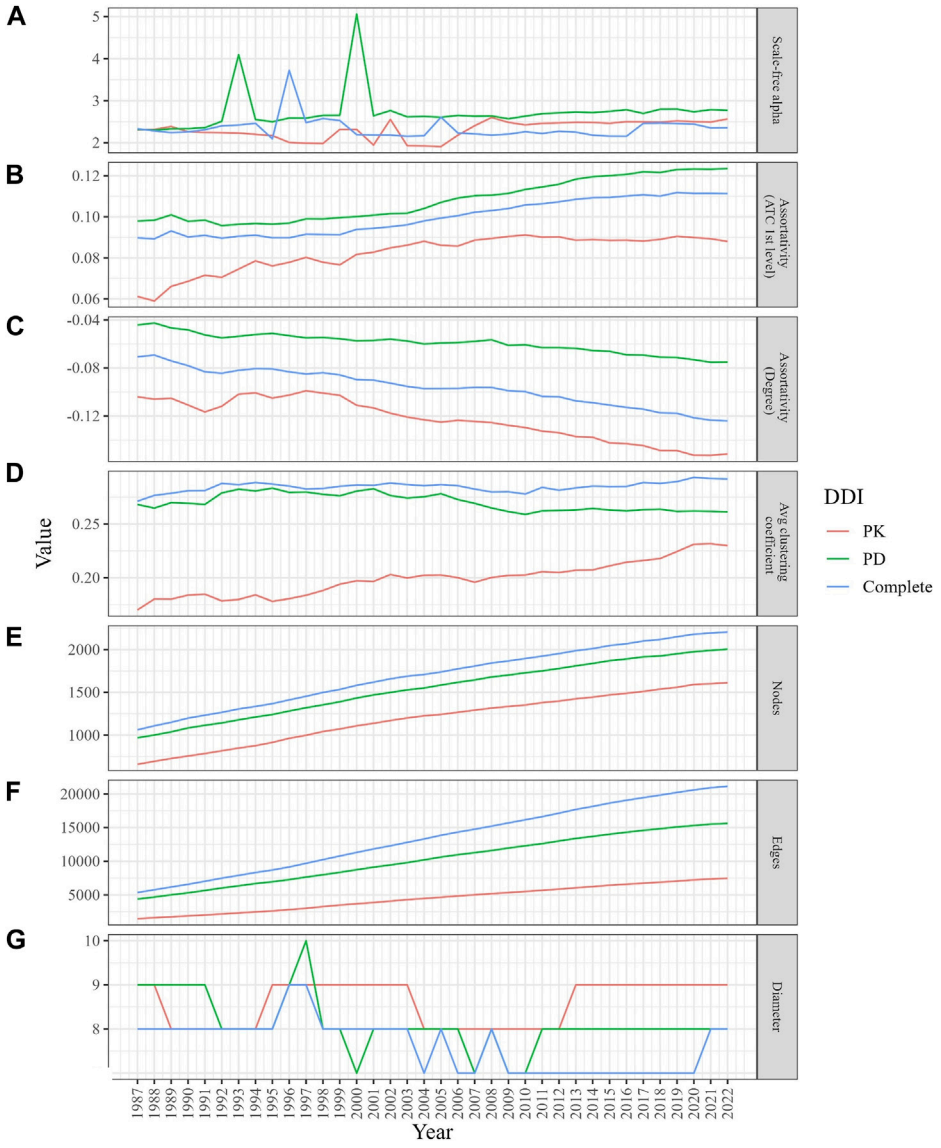

FIGURE 2
Changes in the structural properties of the evolving DDI networks (PK, PD, and complete) from 1987 to 2022. **(A)** The graph demonstrates a stable scale-free alpha (between 2 and 3) for the PK DDI network, while the PD and complete DDI networks show fluctuations until 2001. **(B)** The graph displays a higher likelihood of DDIs among drugs in the same ATC category. **(C)** The graph indicates a decrease in degree assortativity. **(D)** The graph indicates an increase in local clustering coefficients, especially in the PD network. **(E,F)** The graphs illustrate how the size of the network grows over time. **(G)** Despite network growth, the diameters remained stable.

3.3.6 Evolving trends in the ATC category pairs with the highest odds ratios

From the 1980s to the early 2000s, the R-P (Respiratory system - Antiparasitic products, insecticides, and repellents) and P-S (Antiparasitic products, insecticides, and repellents—Sensory organs) pairs displayed the highest odds ratio in the PK DDI network, indicating a higher likelihood of DDIs between drugs from these categories compared to other categories. During the same period, P-V (Antiparasitic products, insecticides, and repellents—Various) and D-J pairs exhibited the highest odds ratio in the PD and complete networks from 1987 to 2011 (Figure 5). From 2013 to 2018, the D-J pair had the highest odds ratio in the PD and complete networks, while the H-V pair showed the highest odds ratio in the PK network from 2014 to 2022.

3.4 Key influential drugs and trends in DDI networks

3.4.1 Rifampin and Morphine: highly influential drugs in static DDI networks

We found that rifampin ranked first in all three centralities in the PK DDI network, while morphine exhibited the highest values in all three centralities in the PD DDI network. In the complete DDI

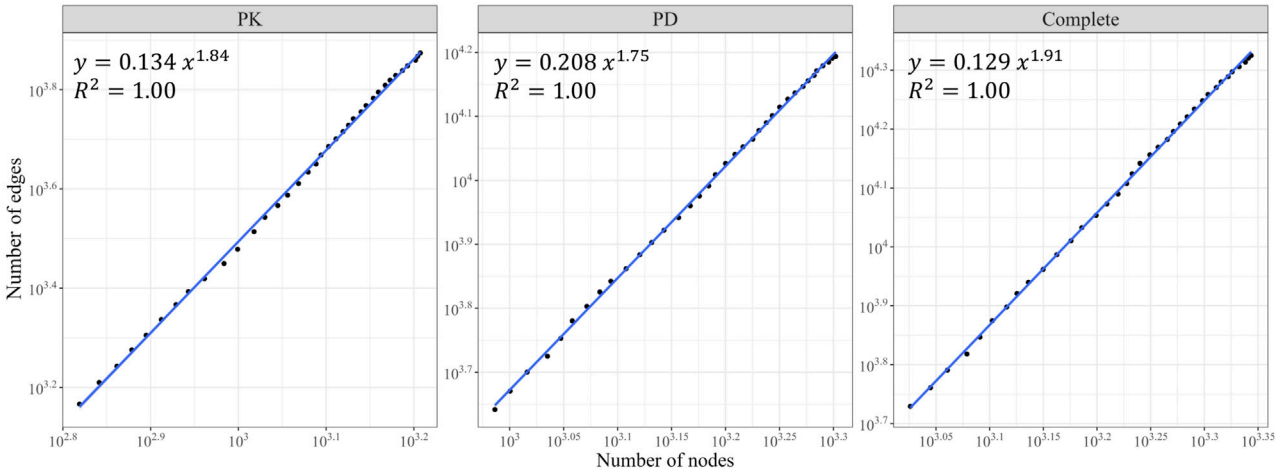

FIGURE 3
The Densification Power Law. The number of edges is plotted against the number of nodes for the PK, PD, and complete DDI networks on a log-log scale. All three networks exhibited a densification power law with a rapid rate of interaction growth (high densification exponents).

network, rifampin had the highest degree and eigenvector centrality values, while morphine showed the highest betweenness.

3.4.2 Cimetidine, morphine, ethanol, and rifampin: highly influential drugs over time in evolving DDI networks

Figure 6 displays the drugs with the highest degree, betweenness, and eigenvector centralities for each year. In the PK DDI network from 1987 to 2012, cimetidine demonstrated the highest degree and eigenvector centrality, suggesting that it was extensively investigated for DDIs with numerous other high-degree drugs. Additionally, cimetidine exhibited high betweenness, serving as a connecting point or bridge for various DDIs. Since 2015, rifampin has held the highest degree, betweenness, and eigenvector centrality, indicating its involvement in many DDIs and its interactions with drugs that also have multiple DDIs. By contrast, morphine maintained the highest degree and eigenvector centrality in the PD and complete DDI networks from 1987 to 2022. Ethanol exhibited the highest betweenness in the complete DDI network from 1987 to 2019, while in the most recent 3 years, rifampin emerged with the highest betweenness in the complete DDI network.

3.4.3 Rifampin and fluoxetine: emerging drugs in evolving DDI networks

In the PK and complete DDI network, rifampin exhibited the highest growth rate across degree and betweenness centrality measures. Meanwhile, fluoxetine showed the highest growth rate in eigenvector centrality within the PD and complete DDI networks.

3.4.4 Declining attention on drugs in evolving networks

The eigenvector centrality of cimetidine experienced the greatest decrease in the PK network, while reserpine exhibited the least increase in eigenvector centrality in the PD and complete networks (Table 4). Supplementary Table S3 presents the rate of increase (or decrease) for each drug's centrality scores in the PK, PD, and complete DDI networks.

4 Discussion

In this study, we employ NLP techniques to extract PK and pharmacodynamic PD DDI evidence from PubMed articles, subsequently characterizing key trends and patterns through static and evolving network analyses. Our findings highlight the scale-free nature of PK and PD networks, with a small number of drugs serving as central hubs, engaging in numerous interactions with other drugs. This observation suggests that the research has focused on specific drugs and their interactions, which could guide future studies to either further explore these central hubs or investigate less-studied drugs. We demonstrate that these networks conform to a densification power law, indicating an exponential growth of DDIs as new drugs are introduced, and emphasizing the increasing complexity of the DDI landscape. Notably, our analysis reveals that drugs within PK and PD networks predominantly belong to the same ATC categories, with fewer interactions observed between drugs from different categories. This insight suggests that DDIs between drugs from distinct ATC categories might be under-explored in the existing literature, warranting further investigation. Moreover, we identify highly influential drugs within static and evolving DDI networks, providing valuable information for future DDI research, clinical practice, and potential areas of focus in drug development.

Our network analysis identified that drugs like rifampin and morphine had high centrality measures, indicating their prominence in DDI research. Rifampin is an antibiotic agent used for treating tuberculosis and other bacterial infections. It is frequently administered in conjunction with other antituberculosis drugs or other families of drugs and has a significant potential for drug interactions due to its well-known induction of drug metabolism through cytochrome P450 (CYP)1A2, CYP2C8, CYP2C9, CYP2C19, CYP3A4, and some glucuronidation pathways (Venkatesan, 1992). It is difficult to predict which medications will be affected by the selective enzyme-induction effect of rifampin (Venkatesan, 1992). Morphine, on the other hand, is the first-choice opioid for the management of cancer pain

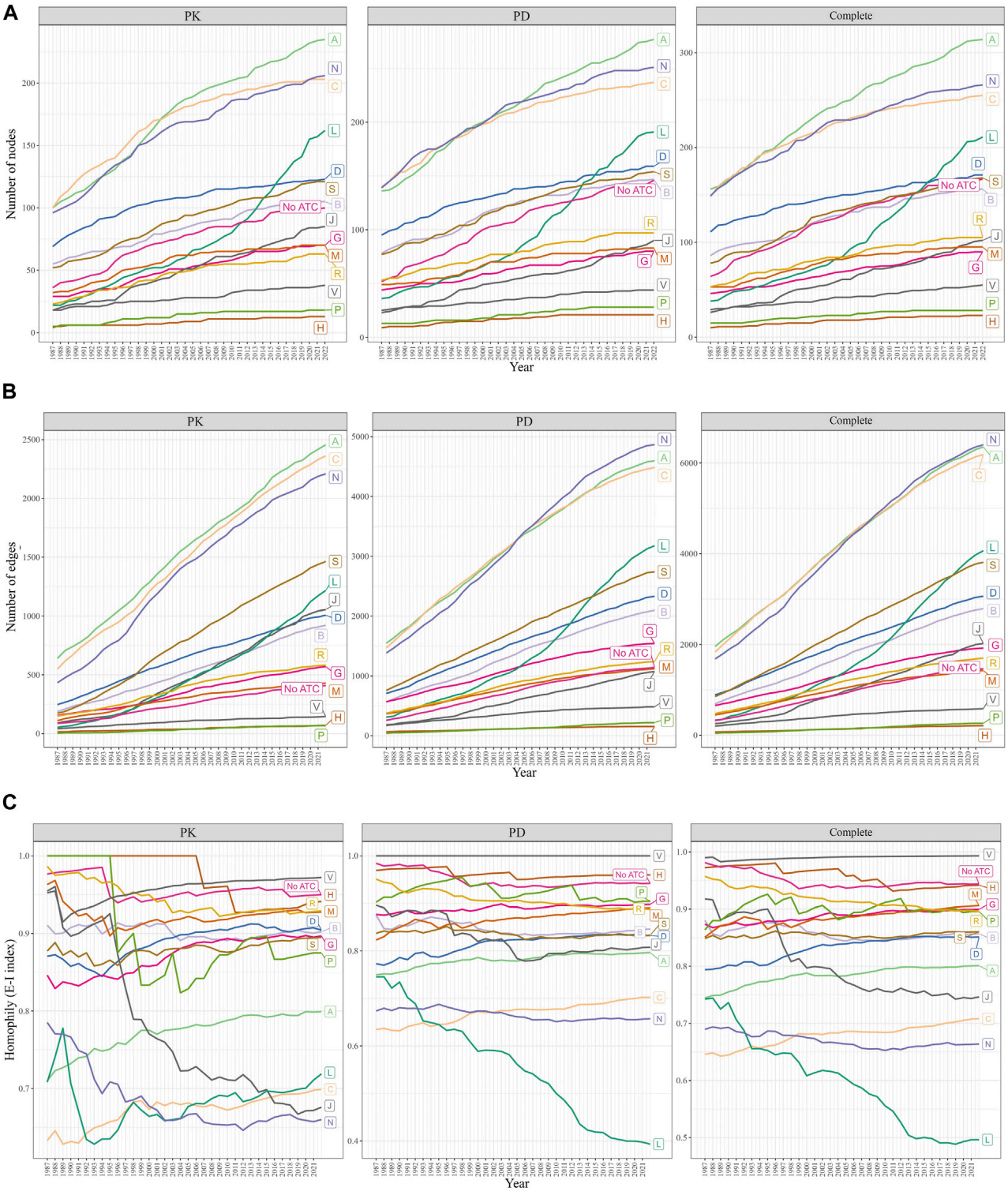

FIGURE 4
Changes in the size and homophily of the ATC first-level categories. **(A)** Change in the number of nodes. **(B)** Changes in the number of edges. **(C)** Change in the EI-index was measured from 1987 to 2022. A drug may have multiple ATC first-level categories, or none ("No ATC"). The A, C, and N categories were the largest, while the L category grew exponentially since 2007. E-I index scores less than 1 indicated intra-category DDIs, with the L category's E-I decreasing and the C category's E-I increasing rapidly.

according to the World Health Organization (WHO) guidelines (Staff and Organization, 1996). The risk of DDIs is high in cancer patients due to a large number of concomitant drugs (Kotlinska-Lemieszek et al., 2014). In the static PD and complete networks, the most researched DDI was morphine–naloxone. This DDI was intensively studied between the 1980s and the early 2000s.

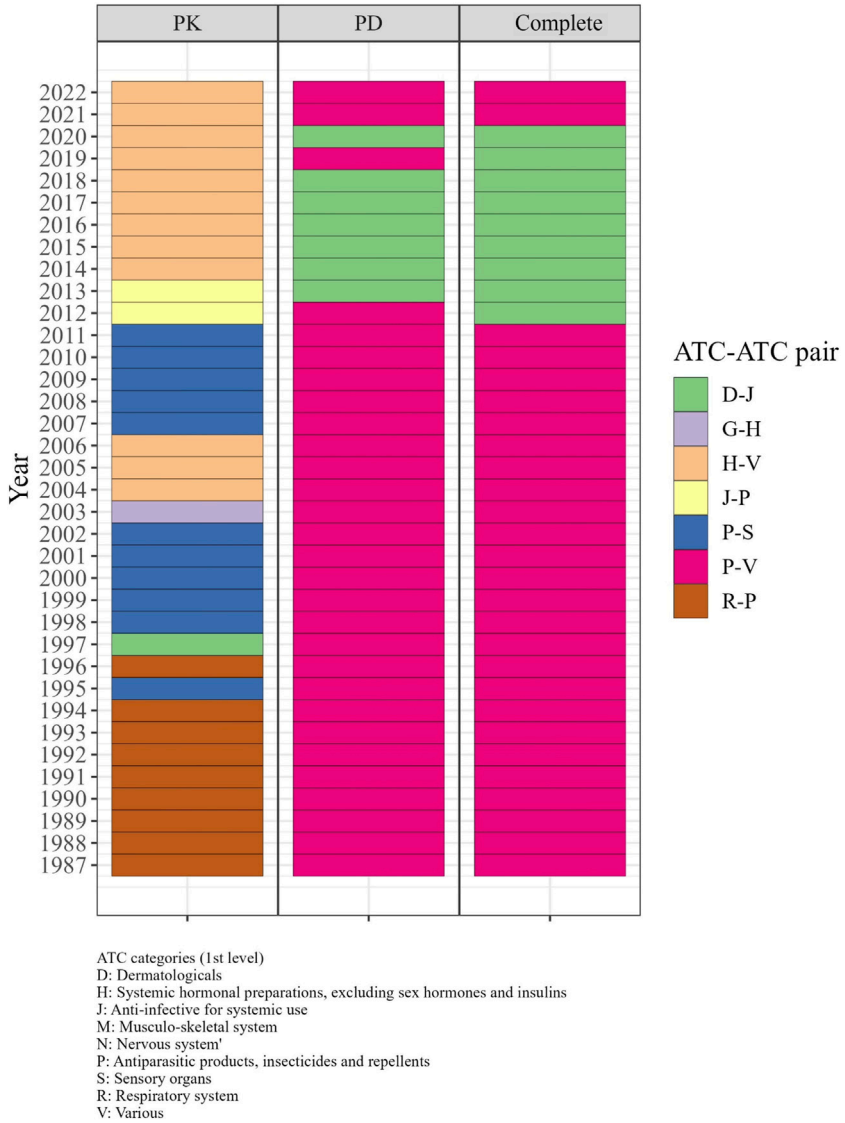

FIGURE 5
The significant ATC-ATC pairs with the highest odds ratios in each year from 1987 to 2022. In the 1980s–2000s, R-P and P-S pairs in PK, and P-V and D-J pairs in PD and complete networks had the highest odds ratios for DDIs. From 2013–2018, D-J pairs in PD and complete networks and H-V pairs in PK showed the highest odds ratios.

Morphine is the classic opioid agonist that provides considerable analgesia and respiratory depression (Sartain et al., 2003), while naloxone is an opioid antagonist capable of reversing the powerful opioid effects of morphine and inducing the opposite effect of hyperalgesia and reversing respiratory depression (Westbrook and Greeley, 1990). Morphine and naloxone were the most notable opioid drugs studied in the past for pain modulation, opioid tolerance, and opioid dependency, especially since naloxone is considered an antidote for morphine and other opioids (Westbrook and Greeley, 1990).

Our investigation also indicated dynamic changes in the DDI research over time using an evolving network analysis, which the traditional static network analysis is unable to provide. While the PK, PD, and complete DDI networks were scale-free, they also followed the densification power law, wherein the number of DDIs grows faster than the number of drugs—networks become denser over time. Despite the network size growth over time, the average local clustering coefficient of all three networks remains high, indicating that the DDI networks are developing small-world network characteristics. This suggests that the drugs studied in DDI research trends are becoming increasingly interconnected and that the scientific community is becoming more adept at examining and comprehending the complex relationships between drugs. Degree assortativity has decreased over the years, while ATC-group assortativity has increased, suggesting that the number of connections between drugs with a high and low degree has been increasing (decreased degree assortativity), and the number of connections between drugs in the same ATC categories has been increasing (increased ATC assortativity).

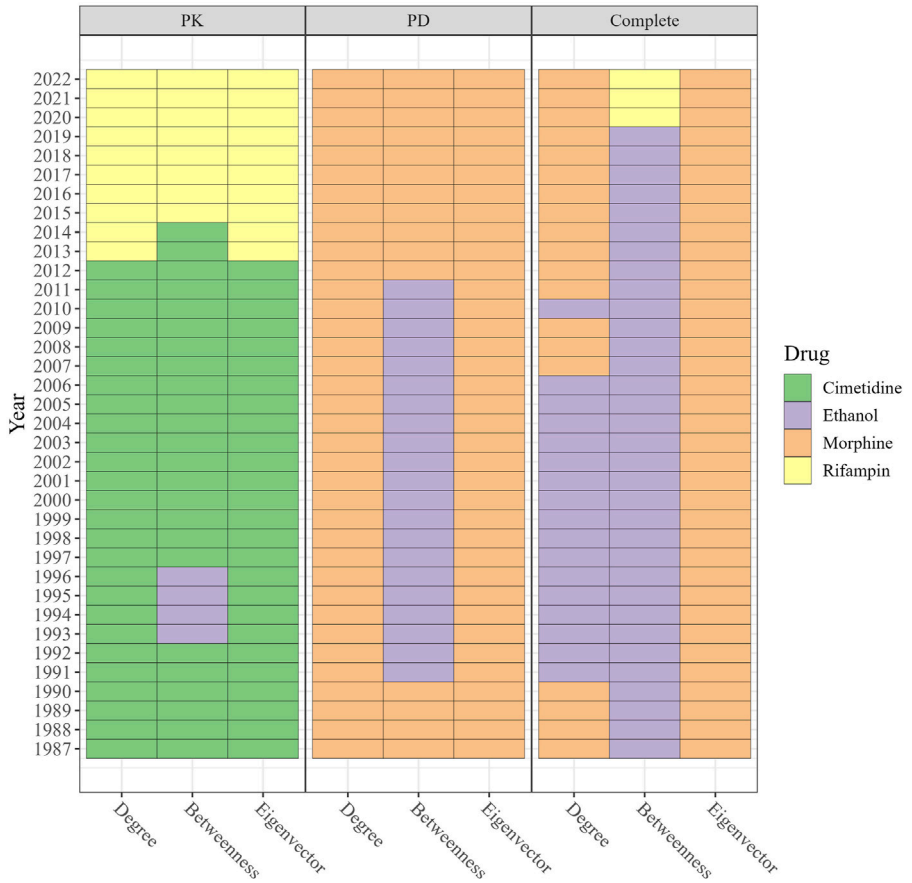

FIGURE 6
The drugs with the highest degree, betweenness, and eigenvector in PK DDI, PD DDI, and complete DDI networks in each year ranged from 1987 to 2022. The color in the cell represents the drug index. From 1987–2012, cimetidine exhibited the highest degree and eigenvector in the PK network, while rifampin has dominated since 2015. Morphine maintained the highest degree and eigenvectors in the PD and complete networks, with ethanol having the highest betweenness until 2019 when rifampin surpassed it.

TABLE 4 The drugs with the highest average, as well as growth rate, in the three types of centrality (degree, betweenness, and eigenvector) over the last 5 years.

		PK	PD	Complete
Largest increase	Degree	Rifampin	Cisplatin	Rifampin
	Betweenness	Rifampin	Morphine	Rifampin
	Eigenvector	Ketoconazole	Fluoxetine	Fluoxetine
Lowest increase (or largest decrease)	Degree	Chlordiazepoxide	Trimethaphan	Meprobamate
	Betweenness	Hexobarbital	Oxygen	Rimantadine
	Eigenvector	Cimetidine	Reserpine	Reserpine

Even though the A, C, and N categories still comprised the majority of DDI research, the ATC-level analysis revealed that the number of drugs and DDIs in the L category has increased dramatically since 2002. This may be because combination therapy, a treatment modality that combines two or more therapeutic agents, is a cornerstone of cancer therapy (Bayat Mokhtari et al., 2017). This also explains the decreasing trend in the E-I index of the L category, which indicates that the DDIs between drugs in the same L category have recently been investigated. It is worth noting that the number of DDI studies may be influenced by prescription frequency. For example, According to Bodenreider and Rodriguez (2014), despite the fact that the dataset was based on emergency room patients for 3 months in 2011, the A, C, and N categories were the most commonly prescribed drug categories, so the sheer number of DDIs found in categories A, C, and N might be inflated due to the fact that these drugs are more commonly prescribed, leading to more observations and subsequent publications. However, we have found that high-

frequency prescribed drugs are not always investigated in DDI research, and low-frequency prescribed drugs can also be highly investigated for DDIs. For instance, the L category drugs were prescribed at a very low rate, but our research showed that the number of L category related DDI studies was very high in 2011. Conversely, the H category drugs were frequently prescribed, but their DDIs were rarely investigated in 2011. These findings suggest that other factors, such as safety concerns or emerging research interests, may play a role in driving DDI research beyond drug prescription frequency alone.

Between 1987 to 2012, cimetidine had the highest degree, betweenness, and eigenvector centralities in the PK DDI network, but it was replaced by rifampin. Cimetidine has numerous drug interactions due to its nonselective inhibition of cytochrome P450 enzymes (Levine and Bellward, 1995). The introduction of longer-acting H2 receptor antagonists with fewer side effects and drug interactions has diminished the usage of cimetidine, and it is no longer one of the most regularly used H2 receptor antagonists. Similarly, the therapeutic applications of trimethaphan (a vasodilator), which showed the lowest eigenvector increase in the PK network, are extremely limited due to competition from newer drugs with more selective actions and effects produced (Wilkins et al., 2007).

Rimantadine demonstrated the greatest decrease in betweenness in the complete network, which may be due to the fact it is not recommended for use in the United States since 2009 because of widespread antiviral resistance to this class of antivirals among circulating flu A viruses (Bloom et al., 2010). Colistin–meropenem was the most actively researched DDI in the PD and complete DDI networks over the past 5 years. Numerous studies demonstrated that the combination of different antibiotics with colistin, such as meropenem produced favorable results (Biancofiore et al., 2007). Recently, researchers have questioned whether the colistin–meropenem combination has a synergistic effect (better than monotherapy) against bacteria (Soudeiha et al., 2017). The controversial opinions expressed by researchers may have prompted the recent active investigation of this DDI.

Despite its contributions, our study has several limitations. First, although we employed the ScispaCy and SciFive models, the results may include false positives and overlook articles due to false negatives, as the model is not perfect. While we confirmed the performance of the NLP models through manual evaluations of a set of randomly selected DDIs, a thorough manual examination of all DDI sentences would be necessary to improve the quality of the results. Second, some relevant publications may have been excluded from this study if they did not fall within the search criteria. For instance, our conclusions are based on the assumption that all DDI articles contained at least one sentence with at least two drug entities in their abstracts. However, there may be DDI articles that lack such a sentence or contain a sentence with at least two drug entities only in the full text, and our study would not include these articles. However, extracting DDIs from full-text articles with acceptable performance is challenging for NLP models. Despite the existence of advanced NLP models such as SciFive, knowledge graphs, and large language models, their performance in extracting DDIs from full-text articles is unknown. Furthermore, many sentences in full-text articles describe or introduce DDIs from cited papers, which can skew the results. Third, we acknowledge that the 5-year investigation window size we chose to inform readers about recent DDI research trends was arbitrary. Even though we believed that a 5-year period could provide recent trends in DDI research because longer timeframes could capture more historical trends and shorter timeframes could not reveal trends adequately, the selection of a 5-year investigation window size may not fully represent the recent DDI research trend. As a result, the recent trends in this paper should be interpreted using the 5-year investigation window. Lastly, the quality of DDI evidence extracted from the literature is dependent on the quality of the original research, which may be limited or inconsistent. This may lead to variability in the quality and relevance of DDI evidence extracted from the literature, potentially resulting in incomplete or biased analyses. Future studies should focus on enhancing data quality, including the manual curation of DDI evidence from published literature, to develop high quality DDI networks. Incorporating the clinical implications of DDIs into network analysis is also crucial. This would highlight the clinical significance of these interactions, providing insights that could be instrumental in optimizing patient care.

Data availability statement

The datasets presented in this study can be found in online repositories. The names of the repository/repositories and accession number(s) can be found in the article/Supplementary Material.

Author contributions

EJ and YC were responsible for the overall design, development, and evaluation of this study. EJ, SN, YS, BM, LL, and YC performed the data analysis, methods design, experiment design, evaluation, and interpretation of the experiments and writing of the manuscript. All authors contributed to the article and approved the submitted version.

Funding

This research was supported, in part, by the National Library of Medicine of the National Institutes of Health under Award Number R01LM014199 and T15LM007450.

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.

Supplementary material

The Supplementary Material for this article can be found online at: https://www.frontiersin.org/articles/10.3389/fphar.2023.1211491/full#supplementary-material

References

Bayat Mokhtari, R., Homayouni, T. S., Baluch, N., Morgatskaya, E., Kumar, S., Das, B., et al. (2017). Combination therapy in combating cancer. *Oncotarget* 8, 38022–38043. doi:10.18632/oncotarget.16723

Becker, M. L., Kallewaard, M., Caspers, P. W., Visser, L. E., Leufkens, H. G., and Stricker, B. H. (2007). Hospitalisations and emergency department visits due to drug-drug interactions: a literature review. *Pharmacoepidemiol Drug Saf.* 16, 641–651. doi:10.1002/pds.1351

Biancofiore, G., Tascini, C., Bisa, M., Gemignani, G., Bindi, M. L., Leonildi, A., et al. (2007). Colistin, meropenem and rifampin in a combination therapy for multi-drug-resistant Acinetobacter baumannii multifocal infection. A case report. *Minerva Anestesiol.* 73, 181–185.

Bloom, J. D., Gong, L. I., and Baltimore, D. (2010). Permissive secondary mutations enable the evolution of influenza oseltamivir resistance. *Science* 328, 1272–1275. doi:10.1126/science.1187816

Bodenreider, O., and Rodriguez, L. M. (2014). Analyzing U.S. Prescription lists with RxNorm and the ATC/DDD index. *AMIA Annu. Symp. Proc.* 2014, 297–306.

Bodenreider, O. (2004). The unified Medical Language System (UMLS): integrating biomedical terminology. *Nucleic Acids Res.* 32, D267–D270. doi:10.1093/nar/gkh061

Boyce, R., Gardner, G., and Harkema, H. (2012). "Using natural language processing to identify pharmacokinetic drug-drug interactions described in drug package inserts," in Proceedings of the 2012 Workshop on Biomedical Natural Language Processing, Montreal, Quebec, Canada, 2012 May 07 (Association for Computational Linguistics), 206–213.

Chen, Y., Yan, C., and Patel, M. B. (2020). Network analysis subtleties in ICU structures and outcomes. *Am. J. Respir. Crit. Care Med.* 202, 1606–1607. doi:10.1164/rccm.202008-3114LE

Csardi, G., and Nepusz, T. (2006). The igraph software package for complex network research. *InterJournal, complex Syst.* 1695, 1–9.

Duda, S., Aliferis, C., Miller, R., Statnikov, A., and Johnson, K. (2005). Extracting drug-drug interaction articles from MEDLINE to improve the content of drug databases. *AMIA Annu. Symp. Proc.* 2005, 216–220.

Honnibal, M., and Montani, I. (2017). spaCy 2: Natural language understanding with Bloom embeddings, convolutional neural networks and incremental parsing. *appear* 7, 411–420.

Jeong, E., Ko, K., Oh, S., and Han, H. W. (2017). Network-based analysis of diagnosis progression patterns using claims data. *Sci. Rep.* 7, 15561. doi:10.1038/s41598-017-15647-4

Kleinfeld, J. (2002). Could it be a big world after all? The six degrees of separation myth. *Society* 12, 5–2.

Kotlinska-Lemieszek, A., Paulsen, O., Kaasa, S., and Klepstad, P. (2014). Polypharmacy in patients with advanced cancer and pain: a European cross-sectional study of 2282 patients. *J. Pain Symptom Manage* 48, 1145–1159. doi:10.1016/j.jpainsymman.2014.03.008

Krackhardt, D., and Stern, R. N. (1988). Informal networks and organizational crises: An experimental simulation. *Soc. Psychol. Q.* 51, 123–140. doi:10.2307/2786835

Leskovec, J., Kleinberg, J., and Faloutsos, C. (2007). Graph evolution: Densification and shrinking diameters. *ACM Trans. Knowl. Discov. Data (TKDD)* 1, 2–es. doi:10.1145/1217299.1217301

Leskovec, J., Kleinberg, J., and Faloutsos, C. (2005). "Graphs over time: densification laws, shrinking diameters and possible explanations," in Proceedings of the eleventh ACM SIGKDD international conference on Knowledge discovery in data mining, 177–187.

Levine, M., and Bellward, G. D. (1995). Effect of cimetidine on hepatic cytochrome P450: evidence for formation of a metabolite-intermediate complex. *Drug Metab. Dispos.* 23, 1407–1411.

Nebert, D. W., and Russell, D. W. (2002). Clinical importance of the cytochromes P450. *Lancet* 360, 1155–1162. doi:10.1016/S0140-6736(02)11203-7

Neumann, M., King, D., Beltagy, I., and Ammar, W. (2019). *ScispaCy: Fast and robust models for biomedical natural language processing. arXiv preprint arXiv:1902.07669.*

Nigam, S. K. (2015). What do drug transporters really do? *Nat. Rev. Drug Discov.* 14, 29–44. doi:10.1038/nrd4461

Niu, J., Straubinger, R. M., and Mager, D. E. (2019). Pharmacodynamic drug-drug interactions. *Clin. Pharmacol. Ther.* 105, 1395–1406. doi:10.1002/cpt.1434

Phan, L. N., Anibal, J. T., Tran, H., Chanana, S., Bahadroglu, E., Peltekian, A., et al. (2021). *Scifive: A text-to-text transformer model for biomedical literature. arXiv preprint arXiv:2106.03598.*

Qu, Y., Guan, X., Zheng, Q., Liu, T., Zhou, J., and Li, J. (2015). Calling network: A new method for modeling software runtime behaviors. *ACM SIGSOFT Softw. Eng. Notes* 40, 1–7. doi:10.1016/j.pupt.2015.07.004

Qu, Y., Zheng, Q., Liu, T., Li, J., and Guan, X. (2014). In-depth measurement and analysis on densification power law of software execution. *Proc. 5th Int. Workshop Emerg. Trends Softw. Metrics*, 55–58. doi:10.1145/2593868.2593878

Raffel, C., Shazeer, N., Roberts, A., Lee, K., Narang, S., Matena, M., et al. (2020). Exploring the limits of transfer learning with a unified text-to-text transformer. *J. Mach. Learn. Res.* 21, 1–67.

Sartain, J. B., Barry, J. J., Richardson, C. A., and Branagan, H. C. (2003). Effect of combining naloxone and morphine for intravenous patient-controlled analgesia. *Anesthesiology* 99, 148–151. doi:10.1097/00000542-200307000-00024

Segura-Bedmar, I., MartïNEZ FernāNDEZ, P., and Herrero Zazo, M. (2013). *Semeval-2013 task 9: Extraction of drug-drug interactions from biomedical texts (ddiextraction 2013).* Association for Computational Linguistics.

Soudeiha, M. A. H., Dahdouh, E. A., Azar, E., Sarkis, D. K., and Daoud, Z. (2017). *In vitro* evaluation of the colistin-carbapenem combination in clinical isolates of A. Baumannii using the checkerboard, etest, and time-kill curve techniques. *Front. Cell Infect. Microbiol.* 7, 209. doi:10.3389/fcimb.2017.00209

Staff, W. H. O., and Organization, W. H. (1996). *Cancer pain relief: With a guide to opioid availability*. World Health Organization.

Van Mil, J. W. (2016). Stockley's drug interactions 11th edition. *Int. J. Clin. Pharm.* 38, 1003–1004. doi:10.1007/s11096-016-0325-2

Venkatesan, K. (1992). Pharmacokinetic drug interactions with rifampicin. *Clin. Pharmacokinet.* 22, 47–65. doi:10.2165/00003088-199222010-00005

Westbrook, R., and Greeley, J. (1990). Some effects of the opioid antagonist, naloxone, upon the rat's reactions to a heat stressor. *Q. J. Exp. Psychol. Sect. B* 42, 1–40. doi:10.1080/14640749008401869

Wilkins, B. W., Hesse, C., Sviggum, H. P., Nicholson, W. T., Moyer, T. P., Joyner, M. J., et al. (2007). Alternative to ganglionic blockade with anticholinergic and alpha-2 receptor agents. *Clin. Auton. Res.* 17, 77–84. doi:10.1007/s10286-006-0387-7

Yan, C., Zhang, X., Gao, C., Wilfong, E., Casey, J., France, D., et al. (2021). Collaboration structures in COVID-19 critical care: Retrospective network analysis study. *JMIR Hum. Factors* 8, e25724. doi:10.2196/25724

frontiers | Frontiers in Pharmacology

TYPE Original Research
PUBLISHED 30 October 2023
DOI 10.3389/fphar.2023.1255021

Check for updates

OPEN ACCESS

EDITED BY
Blythe Adamson,
Flatiron Health, United States

REVIEWED BY
Faiz Ahmed Mohamed Elfaki,
Qatar University, Qatar
Jim Nutaro,
Oak Ridge National Laboratory (DOE), United States

*CORRESPONDENCE
Fanny Franchini,
fanny.franchini@unimelb.edu.au

†These authors have contributed equally to this work and share first authorship

RECEIVED 08 July 2023
ACCEPTED 10 October 2023
PUBLISHED 30 October 2023

CITATION
Franchini F, Fedyashov V, IJzerman MJ and Degeling K (2023), Implementing competing risks in discrete event simulation: the event-specific probabilities and distributions approach. *Front. Pharmacol.* 14:1255021. doi: 10.3389/fphar.2023.1255021

COPYRIGHT
© 2023 Franchini, Fedyashov, IJzerman and Degeling. This is an open-access article distributed under the terms of the Creative Commons Attribution License (CC BY). The use, distribution or reproduction in other forums is permitted, provided the original author(s) and the copyright owner(s) are credited and that the original publication in this journal is cited, in accordance with accepted academic practice. No use, distribution or reproduction is permitted which does not comply with these terms.

Implementing competing risks in discrete event simulation: the event-specific probabilities and distributions approach

Fanny Franchini[1,2*†], Victor Fedyashov[3†], Maarten J. IJzerman[1,2,4,5] and Koen Degeling[1,2]

[1]Cancer Health Services Research, Centre for Health Policy, Melbourne School of Population and Global Health, Faculty of Medicine, Dentistry and Health Sciences, The University of Melbourne, Melbourne, VIC, Australia, [2]Cancer Health Services Research, Centre for Cancer Research, Faculty of Medicine, Dentistry and Health Sciences, The University of Melbourne, Melbourne, VIC, Australia, [3]ARC Training Centre in Cognitive Computing for Medical Technologies, The University of Melbourne, Parkville, VIC, Australia, [4]Department of Cancer Research, Peter MacCallum Cancer Centre, Melbourne, Australia, [5]Erasmus School of Health Policy & Management, Erasmus University, Rotterdam, Netherlands

Background: Although several strategies for modelling competing events in discrete event simulation (DES) exist, a methodological gap for the event-specific probabilities and distributions (ESPD) approach when dealing with censored data remains. This study defines and illustrates the ESPD strategy for censored data.

Methods: The ESPD approach assumes that events are generated through a two-step process. First, the type of event is selected according to some (unknown) mixture proportions. Next, the times of occurrence of the events are sampled from a corresponding survival distribution. Both of these steps can be modelled based on covariates. Performance was evaluated through a simulation study, considering sample size and levels of censoring. Additionally, an oncology-related case study was conducted to assess the ability to produce realistic results, and to demonstrate its implementation using both frequentist and Bayesian frameworks in R.

Results: The simulation study showed good performance of the ESPD approach, with accuracy decreasing as sample sizes decreased and censoring levels increased. The average relative absolute error of the event probability (95%-confidence interval) ranged from 0.04 (0.00; 0.10) to 0.23 (0.01; 0.66) for 60% censoring and sample size 50, showing that increased censoring and decreased sample size resulted in lower accuracy. The approach yielded realistic results in the case study.

Discussion: The ESPD approach can be used to model competing events in DES based on censored data. Further research is warranted to compare the approach to other modelling approaches for DES, and to evaluate its usefulness in estimating cumulative event incidences in a broader context.

KEYWORDS

discrete event simulation, competing risks modelling, censored data, frequentist implementation, bayesian framework

1 Introduction

Discrete event simulation (DES) is increasingly used to model disease, treatment, and care delivery pathways in healthcare (Günal and Pidd, 2017; Vazquez-Serrano et al., 2021). Given its event-based handling of time and the ability to account for resource capacity constraints, it is an effective and efficient individual-level (or microsimulation) modelling technique for a range of decision problems (Marshall et al., 2020). The increased flexibility of DES compared to more traditional approaches, such as state-transition modelling, also implies that certain decisions regarding the model structure and methodologies used may not necessarily be applicable to such traditional approaches and adjustments must be made when implementing such a dynamic model (Karnon et al., 2012).

Competing risks or events are common in healthcare and clinical studies (Pintilie, 2006; Koller et al., 2012; Coemans et al., 2022) and refer to a situation where there are multiple possible outcomes that can occur, and the occurrence of one outcome precludes the occurrence of the others or changes their likelihood. One of the advantages of DES is the ability to implement competing risks using different approaches (Barton et al., 2004; Karnon et al., 2012). In implementing decision-analytic models, every transition in the model pathway typically involves competing events. More specifically, if it is possible to move to more than one model state from a certain state, the transitions to these subsequent states are competing risks. For example, for a model structure commonly used in oncology defined by three health states (i.e., disease free, recurrence, and death), the possible transitions to the recurrence or death state from the disease-free state are competing risks. Similarly, in a model of patient flows in an emergency department, discharging a patient after triaging by a nurse may be a competing event relative to the patient being referred to an emergency doctor for further investigations.

The ability to model competing risks using different strategies allows the modeler to select the approach that best suits the available evidence and context (Caro and Möller, 2014). Each strategy necessitates defining a data analysis framework and the required simulation steps, collectively referred to as a modelling approach. In terms of competing risks, two broad approaches to time-to-event estimation are commonly used. When considering competing risks, there are two broad approaches to time-to-event estimation (Barton et al., 2004). The first calculates individual time estimates for each potential subsequent event and proceeds based on which event is predicted to occur earliest. The second approach also generates an overall time estimate for the next event but employs an additional sampling process to identify the specific type of event likely to happen. Importantly, the likelihood of each event type occurring can be influenced by this initially sampled time-to-event.

These approaches can be broadly categorised into specific modelling strategies (Barton et al., 2004; Degeling et al., 2019; Degeling et al., 2022).

1. Strategy 1—Event-Specific Distributions (ESD): it involves sampling times to each event and simulating the first event to occur. It uses event-specific distributions to sample time-to-event for each competing event and then selects the earliest to simulate.
2. Strategy 2—Event-Specific Probabilities and Distributions (ESPD): the event type is sampled first based on specific probabilities, followed by sampling the time-to-event from the corresponding distribution. The resulting model is a mixture of event-specific time-to-event distributions, weighted by their probabilities.
3. Strategy 3—Unimodal and Multimodal Distribution and Regression (UDR & MDR): the time-to-event is sampled first, using either a unimodal or multimodal distribution. It then employs a regression model to determine the specific event that corresponds to the sampled time.
4. Strategy 4–using discrete time cycles with transition probabilities: it operates in discrete time cycles and uses transition probabilities for state changes. This strategy resembles a discrete-time state-transition model (Siebert et al., 2012) more than a traditional DES. While it can be useful in certain contexts, it sacrifices the continuous-time advantages and complex event dependencies that DES is designed to capture.

Previous research has focused on the ESD, ESPD, UDR, and MDR modelling strategies in the context of uncensored individual patient data (Degeling et al., 2019), demonstrating that accuracy depended on the number of competing events, overlap of time-to-event distributions for the competing events, and sample size. While these studies have shown that the ESPD approach performs well and is straightforward to implement for uncensored data (Degeling et al., 2019), there is a methodological gap when it comes to censored data, which is a common challenge in long-term studies and real-world settings. The impact of data censoring on model accuracy has been examined for the ESD and UDR approaches (Degeling et al., 2022), but no framework currently exists for implementing the ESPD approach in the presence of censoring.

The primary objective of this study is to adapt the ESPD approach for handling censored data, which will offer several advantages. Firstly, the ESPD approach has proven to be effective and straightforward for uncensored data (Degeling et al., 2019), yet its application is limited by the lack of a framework to handle censored data. Given the commonality of censored data in long-term and real-world studies, our study could significantly expand the method's applicability. Secondly, while existing methods like ESD and UDR have frameworks to deal with censored data, they do not offer the same advantages as the ESPD in terms of ease of implementation and effective uncertainty estimation around time-to-event parameters. Addressing this limitation involves tackling technical challenges, one of which is the current absence of a well-defined likelihood function tailored for the ESPD approach in censored data scenarios, a gap that our study aims to fill.

By addressing this methodological gap, we provide a more versatile toolset for analysts in this field. On the practical side, we offer implementations of this adapted ESPD approach in both Bayesian and frequentist methods using R, thereby catering to a wide range of statistical preferences and needs. The paper is structured to provide comprehensive evidence for the tailored ESPD approach. We start by defining the ESPD approach in the methods section, followed by a simulation study for accuracy assessment. To demonstrate its utility in real-world scenarios, a case study is included for illustration. The paper concludes with a general discussion that synthesises our findings and outlines directions for future research.

2 Methods

We follow standard notation for survival analyses, where T is the event time or censoring time, a continuous random variable that is distributed according to a particular probability density function $f(t)$, with cumulative distribution function 1) $F(t)$ and survival function 2) $S(t)$ defined as:

$$\boldsymbol{F(t) = P(T \le t) = \int_0^t f(x)dx} \tag{1}$$

$$\boldsymbol{S(t) = P(T > t) = 1 - F(t) = \int_t^\infty f(x)dx} \tag{2}$$

Furthermore, let $K = 1, \ldots, k$ be the index set for k mutually exclusive independent competing events and let C_j be the event indicator which shows whether person $i, i = 1, \ldots, n$ experienced event j, $j = 1, \ldots, k$, or not. For notational simplicity we encode our events with the vector c, i.e., if competing event j is experienced, then $c_j = 1, c_{i,i\neq j} = 0$.

Building on 1) and 2), we introduce the ESPD strategy. It offers a two-step procedure for generating events in a competing risk scenario. First, the type of event is selected based on some (unknown) mixture proportions. Second, the time of occurrence for the chosen event is sampled from a corresponding survival distribution (e.g., Weibull, Gompertz, *etc.*). Effectively, this results in times that are a *mixture of m distributions*, where $m = k$ aligns naturally with the k mutually exclusive independent competing events described in the survival analysis framework. Additionally, we allow both the mixture proportions and the survival distributions to depend on covariates X, such as age, disease stage, *etc.*, thereby making the final model a multivariable mixture model.

To successfully implement the ESPD strategy in the context of censored data, the next step is to define a robust likelihood function. Unlike other methods like ESD or UDR, where frameworks for handling censored data are already established, the ESPD strategy lacks such a framework. As a result, our study introduces a tailored likelihood function, to allow for more accurate and reliable parameter estimation. This involves parameterising two critical components: the mixture proportions or event risks i) and the time-to-event distributions ii).

The first component involves modeling the type of event i). Specifically, we employ a multinomial distribution with event probabilities $\pi = (\pi_1, \ldots, \pi_k)$, such that $\sum_{j=1}^k \pi_j = 1$. To allow mixture proportions π to depend on some vector of covariates X^π, a linear relationship is assumed. The model for mixture proportions is constructed using the ***softmax*** function, which takes as input the product of the vector of covariates X^π, and a vector of coefficients $\boldsymbol{\beta^\pi}$. We model it as follows:

$$\boldsymbol{\pi = softmax(\beta^\pi X^\pi)} \tag{3}$$

The relationship between event probabilities and covariates is expressed by the ***softmax*** Eq 3. It serves to map the linear combination of predictor variables to a probability set that always sums up to 1. By ensuring this, the function guarantees a positive probability distribution. These probabilities are subsequently used to estimate the mixture proportions or event risks.

The linear combination of the covariates (known from the data) and their coefficients (to be estimated) creates a score (or logit) for each event, which can be represented as $z = \beta^\pi X^\pi$. If we consider two competing risks and two covariates, our score vector can be detailed as $z = [z_1, z_2]$, where z_1 represents the score for the first event (e.g., recurrence) and z_2 is for the second event (e.g., death).

Further, the transformation using the ***softmax*** function for a given score x_i in a vector x can be given by $\sigma(x_i) = \frac{e^{x_i}}{\sum_{j=1}^L e^{x_j}}$, where, L corresponds to the total number of events, and e is the base of the natural logarithm.

When the transformation is applied to the score vector z, the resulting probabilities for the two events are $\pi_1 = \frac{e^{z_1}}{e^{z_1}+e^{z_2}}$ and $\pi_2 = \frac{e^{z_2}}{e^{z_1}+e^{z_2}}$. Within this context, π_1 gives the probability of the occurrence of the first event, while π_2 provides the probability for the second event.

For the second component ii), conditioned on the occurrence of a specific event j, a particular survival function $\boldsymbol{S_j(t, \theta_j)}$ is used to model the time-to-event distribution 4):

$$\boldsymbol{P\left(T > t \middle| c_j = 1\right) = S_j(t, \theta_j) = \int_t^\infty f_j(s, \theta_j)ds} \tag{4}$$

Different risks may have different distributions and the parameter vectors for these distributions $\theta_j, j = 1, \ldots, k$ can incorporate dependence on (potentially risk-specific) covariates as well.

For non-censored data 5), the probability of observing an event of type j ($c_j = 1$) at time t is:

$$\boldsymbol{P\left(t \middle| c_j = 1\right) = \pi_j f_j(t, \theta_j)} \tag{5}$$

While for censored data 6), since no event is observed $\left(\sum_{j=0}^k c_j = 0\right)$, we know that whichever event got selected, the corresponding *"failure"* occurred after the end of the experiment. In other words, no event has occurred yet:

$$\boldsymbol{P\left(t \middle| \sum_{j=0}^k c_j = 0\right) = \sum_{j=1}^k \pi_j S_j(t, \theta_j)} \tag{6}$$

Ignoring covariates for simplicity, the combined likelihood (for k competing risks) can be written as follows:

$$\boldsymbol{P(t, c \mid \pi, \theta) = \underbrace{\prod_{j=1}^k \left[\pi_j f_j(t, \theta_j)\right]^{c_j}}_{A} \times \underbrace{\left[\left\{\sum_{j=1}^k \pi_j S_j(t, \theta_j)\right\}^{1-\sum_{i=1}^k c_i}\right]}_{B}} \tag{7}$$

Here, term in $\boldsymbol{A}$ applies if an event is observed, i.e., $c_j = 1$ for some j, while $\boldsymbol{B}$ will become 1 since $\mathbf{1}_{\sum_{j=1}^k c_j = 0} = 0$). If no event is observed, i.e., $\sum_{j=0}^k c_j = 0$, only B will contribute.

In Eq 7, we presented the combined likelihood for handling an arbitrary number k of competing risks. When covariates are involved in the analysis, they would generally be incorporated into both π and θ, so that each parameter can be parameterised with covariates.

Considering that many practical applications consider two competing risks, $k = 2$, we also provide the simplified likelihood function 8) for such case. The incorporation of covariates to the likelihood function 8) is implemented in the accompanying R code for both frequentist and Bayesian analyses.

$$P(t,c|\pi,\theta) = (\pi_1 f_1(t,\theta_1))^{c_1} \times (\pi_2 f_2(t,\theta_2))^{c_2} \times (\pi_1 S_1(t,\theta_1) + \pi_2 S_2(t,\theta_2))^{1-c_1-c_2} \quad (8)$$

When considering the two competing risks setting based on Eq 8, there would be seven parameters to be estimated. These parameters, encompassing coefficients in our expression for π, as well as associated shapes and scales, are intrinsically linked to the observables t (time) and c (event type). The latter are observations found in the study dataset, while the parameters are to be estimated from the dataset. Our goal is to determine the best-fit parameters that align with the observed data. This entails solving an optimization problem to obtain the maximum likelihood estimates from the dataset.

3 Results

3.1 Simulation study

A simulation study was performed to verify the accuracy and performance of the ESPD modelling approach. All files related to the simulation study can be accessed on GitHub: https://github.com/koendegeling/CompetingEvents_ESPD.

The considered hypothetical scenario included $k = 2$ competing risks: recurrence (recur) and death before recurrence (death), and we opted for parameter values that are consistent with practical, real-world data, particularly in the oncology setting. The chosen coefficients were selected to reflect realistic relationships between disease stage and time-to-event risks, providing a credible foundation for our model. Simulated patients had equal probabilities of being diagnosed at disease stage IA, IB, or II, and their age was normally distributed, with mean 60 years and a standard deviation of 5 (normalised to mean 0). The true parameter values used to simulate the population were:

$$\log\left(\frac{\pi_{recur}}{1-\pi_{recur}}\right) = -0.4 + 0.4\,stageIB + 0.8\,stageII$$

$$F_{recur}(t|\theta_{recur}) = Weibull\left(t, \theta_{recur}^{shape} = exp(0.7), \theta_{recur}^{scale} = exp(2 - 0.2\,stageIB - 0.6\,stageII)\right)$$

$$F_{death}(t|\theta_{death}) = Gompertz\left(t, \theta_{death}^{shape} = 0.1, \theta_{death}^{rate} = exp(-3.5 + 0.1\,age)\right)$$

Based on these true parameters, a population (s_{pop}) of $n_{sim} = 1,000,000$ individuals was simulated. Subsequently, the performance of the ESPD approach was assessed for a range of scenarios defined by the proportion of censored observations ($p_{censored} = 0.0, 0.1, 0.3, 0.6$) and various sample sizes ($n_{sample} = 50, 100, 200, 500$) using the following procedure:

- For all combinations of $p_{censored}$ and n_{sample}:
- For $n_{run} = 10,000$ iterations:
- Draw a sample $s_{uncensored}$ from population s_{pop} according to n_{sample}
- Censor sample $s_{uncensored}$ according to $p_{censored}$ to obtain sample $s_{censored}$
- Analyse $s_{censored}$ according to the ESPD approach
- Based on the estimated parameters, simulate a new sample s_{sim} of size n_{sim}
- Assess the performance by comparing the outcomes of s_{sim} to the population s_{pop}

Censoring was performed through an independent process where censoring times were sampled from an exponential distribution defined by a censoring rate. If the sampled censoring time was lower than that of the event, the observation was censored at the censoring time. The censoring rate was increased incrementally until the required proportion of censored observations was achieved.

Further, the performance of the approach was assessed in terms of the probability of recurrence, as well as the mean and distribution of the time-to-recurrence and time-to-death. The performance of the event probability and mean time-to-events was quantified using a range of error measures, for which lower values corresponds to a better performance.

- Error (E) or bias: $E = sim - pop$
- Absolute error (AE): $AE = |sim - pop|$
- Relative error (RE): $RE = \frac{sim-pop}{pop}$
- Relative absolute error (RAE): $RAE = \frac{|sim-pop|}{pop}$

In which pop refers to the simulated population of individuals with aforementioned characteristics, which is simulated based on the true parameter values, while sim refers to the simulations based on the parameter values as estimated by the ESPD approach.

Lastly, considering these measures do not consider the variance and spread of the distributions of the time-to-events, we also quantified the performance of these distributions by the Kullback-Leibler divergence (KLD), or relative entropy, which is widely used to assess the likeliness of distributions (Kullback and Leibler, 1951):

$$KLD\left(f_{pop}(t)\middle|f_{sim}(t)\right) = \int_0^\infty f_{pop}(x)\log\left(\frac{f_{pop}(x)}{f_{sim}(x)}\right)dx = \int_0^\infty f_{pop}(x) \times \left(\log\left(f_{pop}(x)\right) - \log\left(f_{sim}(x)\right)\right)dx$$

The KLD is a measure of the distance between probability distributions. In general, the smaller the value of KLD, the closer the simulated distribution is to the true population distribution, and the better the model is at representing the data (Cover, 1999). Thus, a smaller KLD indicates a better fit between the simulated and the true distributions, whereas a larger KLD indicates a worse fit.

Given that none of the performance measures considers second-order uncertainty and given that the frequentist implementation is more computationally efficient in obtaining point-estimates compared to the Bayesian implementation, the former was used in the simulation study. Although the Bayesian implementation is illustrated for the case study, a formal comparison of the two implementations in the simulation study was beyond the scope of this study.

Overall, the ESPD approach performed well. We observed that higher proportions of censoring and lower sample sizes both negatively impacted the accuracy of the approach across all performance measures. All results of the simulation study are available in Table 1.

Regarding the event probability, on average, there was no error in the E or RE up to 30% censoring. For 60% censoring, the RE ranged from −0.02 (95% confidence interval: 0.16; 0.13) for sample size 500, to 0.05 (−0.44; 0.66) for size 50. The average RAE ranged from 0.04 (0.00; 0.10) for multiple scenarios, to 0.23 (0.01; 0.66) for 60% censoring and size 50.

For the mean time-to-event, similar as for the event probability, on average there was basically no error in the E or RE up to 30% censoring. For 60% censoring, unrealistic results

TABLE 1 Simulation study results, based on varying degree of censoring (Prop censor), and different sample size (Size). E: error (bias); AE: absolute error; Prob: probability; Prop: proportion; RE: relative error; RAE: relative absolute error; TTR: time-to-recurrence; TTD: time-to-death.

Prop censor	Size	Prob TTR.E	Prob TTR.AE	Prob TTR.RE	Prob TTR.RAE	Mean TTR.E	Mean TTR.AE	Mean TTR.RE	Mean TTR.RAE	Mean TTD.E	Mean TTD.AE	Mean TTD.RE	Mean TTD.RAE	TTR.KLD	TTD.KLD
0	50	0.00 (−0.14; 0.14)	0.06 (0.00; 0.16)	0.00 (−0.28; 0.28)	0.11 (0.00; 0.32)	−0.01 (−1.11; 1.16)	0.46 (0.02; 1.29)	0.00 (−0.22; 0.23)	0.09 (0.00; 0.26)	0.03 (−2.71; 2.86)	1.13 (0.04; 3.19)	0.00 (−0.22; 0.23)	0.09 (0.00; 0.26)	0.07 (0.01; 0.24)	0.07 (0.02; 0.23)
0	100	0.00 (−0.10; 0.10)	0.04 (0.00; 0.11)	0.00 (−0.20; 0.19)	0.08 (0.00; 0.23)	−0.01 (−0.79; 0.81)	0.32 (0.01; 0.91)	0.00 (−0.16; 0.16)	0.06 (0.00; 0.18)	0.01 (−1.95; 1.99)	0.79 (0.03; 2.23)	0.00 (−0.16; 0.16)	0.06 (0.00; 0.18)	0.04 (0.01; 0.13)	0.05 (0.02; 0.11)
0	200	0.00 (−0.07; 0.07)	0.03 (0.00; 0.08)	0.00 (−0.14; 0.14)	0.06 (0.00; 0.16)	0.00 (−0.55; 0.56)	0.22 (0.01; 0.63)	0.00 (−0.11; 0.11)	0.04 (0.00; 0.13)	0.00 (−1.39; 1.39)	0.57 (0.02; 1.59)	0.00 (−0.11; 0.11)	0.05 (0.00; 0.13)	0.03 (0.01; 0.08)	0.04 (0.02; 0.07)
0	500	0.00 (−0.04; 0.04)	0.02 (0.00; 0.05)	0.00 (−0.09; 0.09)	0.04 (0.00; 0.10)	0.00 (−0.34; 0.35)	0.14 (0.01; 0.39)	0.00 (−0.07; 0.07)	0.03 (0.00; 0.08)	0.01 (−0.86; 0.86)	0.35 (0.01; 0.98)	0.00 (−0.07; 0.07)	0.03 (0.00; 0.08)	0.03 (0.01; 0.05)	0.03 (0.02; 0.05)
0.1	50	0.00 (−0.15; 0.15)	0.06 (0.00; 0.17)	0.00 (−0.30; 0.29)	0.12 (0.00; 0.34)	0.01 (−1.17; 1.35)	0.51 (0.02; 1.46)	0.00 (−0.24; 0.27)	0.10 (0.00; 0.29)	0.06 (−2.91; 3.10)	1.22 (0.05; 3.48)	0.00 (−0.23; 0.25)	0.10 (0.00; 0.28)	0.08 (0.01; 0.29)	0.08 (0.02; 0.29)
0.1	100	0.00 (−0.10; 0.10)	0.04 (0.00; 0.12)	0.00 (−0.21; 0.20)	0.08 (0.00; 0.24)	0.00 (−0.83; 0.86)	0.34 (0.01; 0.97)	0.00 (−0.17; 0.17)	0.07 (0.00; 0.20)	0.02 (−2.06; 2.08)	0.85 (0.03; 2.40)	0.00 (−0.17; 0.17)	0.07 (0.00; 0.19)	0.05 (0.01; 0.14)	0.05 (0.02; 0.14)
0.1	200	0.00 (−0.07; 0.07)	0.03 (0.00; 0.08)	0.00 (−0.14; 0.14)	0.06 (0.00; 0.16)	0.00 (−0.58; 0.59)	0.24 (0.01; 0.68)	0.00 (−0.12; 0.12)	0.05 (0.00; 0.14)	0.01 (−1.48; 1.47)	0.60 (0.02; 1.70)	0.00 (−0.12; 0.12)	0.05 (0.00; 0.14)	0.03 (0.01; 0.08)	0.04 (0.02; 0.08)
0.1	500	0.00 (−0.05; 0.04)	0.02 (0.00; 0.05)	0.00 (−0.09; 0.09)	0.04 (0.00; 0.10)	0.00 (−0.36; 0.37)	0.15 (0.01; 0.41)	0.00 (−0.07; 0.07)	0.03 (0.00; 0.08)	0.02 (−0.90; 0.94)	0.37 (0.01; 1.04)	0.00 (−0.07; 0.08)	0.03 (0.00; 0.08)	0.03 (0.01; 0.05)	0.03 (0.02; 0.05)
0.3	50	0.00 (−0.16; 0.18)	0.07 (0.00; 0.20)	0.00 (−0.33; 0.36)	0.14 (0.01; 0.40)	0.38 (−1.35; 2.99)	0.94 (0.03; 2.99)	0.08 (−0.27; 0.60)	0.19 (0.01; 0.60)	0.08 (−3.98; 4.11)	1.59 (0.05; 4.78)	0.01 (−0.32; 0.33)	0.13 (0.00; 0.38)	0.12 (0.01; 0.45)	0.15 (0.03; 0.76)
0.3	100	0.00 (−0.12; 0.12)	0.05 (0.00; 0.13)	0.00 (−0.23; 0.23)	0.09 (0.00; 0.27)	0.02 (−0.95; 1.17)	0.42 (0.02; 1.23)	0.00 (−0.19; 0.23)	0.08 (0.00; 0.25)	0.06 (−2.43; 2.67)	1.03 (0.04; 2.94)	0.00 (−0.20; 0.22)	0.08 (0.00; 0.24)	0.06 (0.01; 0.19)	0.07 (0.02; 0.24)
0.3	200	0.00 (−0.08; 0.08)	0.03 (0.00; 0.09)	−0.01 (−0.16; 0.15)	0.06 (0.00; 0.18)	−0.01 (−0.68; 0.71)	0.28 (0.01; 0.78)	0.00 (−0.14; 0.14)	0.06 (0.00; 0.16)	0.11 (−1.58; 1.79)	0.69 (0.03; 1.95)	0.01 (−0.13; 0.14)	0.06 (0.00; 0.16)	0.04 (0.01; 0.10)	0.05 (0.02; 0.10)
0.3	500	0.00 (−0.05; 0.05)	0.02 (0.00; 0.06)	0.00 (−0.10; 0.10)	0.04 (0.00; 0.12)	0.00 (−0.42; 0.43)	0.17 (0.01; 0.49)	0.00 (−0.08; 0.09)	0.03 (0.00; 0.10)	0.07 (−0.99; 1.13)	0.43 (0.02; 1.21)	0.01 (−0.08; 0.09)	0.03 (0.00; 0.10)	0.03 (0.01; 0.06)	0.04 (0.02; 0.06)
0.6	50	0.02 (−0.22; 0.33)	0.11 (0.01; 0.33)	0.05 (−0.44; 0.66)	0.23 (0.01; 0.66)	3050.09 (−1.76; 24.99)	3050.68 (0.05; 24.99)	612.99 (−0.35; 5.02)	613.10 (0.01; 5.02)	−0.15 (−7.55; 9.81)	3.11 (0.11; 10.21)	−0.01 (−0.61; 0.79)	0.25 (0.01; 0.82)	0.26 (0.02; 1.00)	0.49 (0.03; 2.97)
0.6	100	0.00 (−0.17; 0.24)	0.08 (0.00; 0.24)	−0.01 (−0.34; 0.48)	0.16 (0.01; 0.49)	1.15 (−1.36; 5.45)	1.70 (0.03; 5.45)	0.23 (−0.27; 1.09)	0.34 (0.01; 1.09)	0.44 (−5.56; 6.83)	2.21 (0.08; 7.47)	0.04 (−0.45; 0.55)	0.18 (0.01; 0.60)	0.14 (0.01; 0.45)	0.23 (0.03; 1.52)
0.6	200	−0.02 (−0.13; 0.12)	0.05 (0.00; 0.14)	−0.03 (−0.25; 0.24)	0.10 (0.00; 0.29)	0.18 (−1.00; 2.42)	0.61 (0.02; 2.42)	0.04 (−0.20; 0.49)	0.12 (0.00; 0.49)	0.73 (−2.75; 5.34)	1.57 (0.06; 5.51)	0.06 (−0.22; 0.43)	0.13 (0.00; 0.44)	0.07 (0.01; 0.25)	0.10 (0.02; 0.41)
0.6	500	−0.01 (−0.08; 0.06)	0.03 (0.00; 0.08)	−0.02 (−0.16; 0.13)	0.06 (0.00; 0.17)	0.02 (−0.64; 0.78)	0.29 (0.01; 0.82)	0.00 (−0.13; 0.16)	0.06 (0.00; 0.17)	0.40 (−1.44; 2.53)	0.85 (0.03; 2.55)	0.03 (−0.12; 0.20)	0.07 (0.00; 0.21)	0.04 (0.01; 0.09)	0.05 (0.02; 0.14)

10.3389/fphar.2023.1255021

TABLE 2 Summary of the data used from the publicly available melanoma dataset.

Variable	N = 205 [n (%); median (IQR)]
Demographics	
Sex (male)	79 (39%)
Diagnosis age (years)	54 (42, 65)
Tumour thickness (mm)	1.94 (0.97, 3.56)
Ulcerated tumour	90 (44%)
Outcomes	
Time to last status assessment (years)	5.5 (4.2, 8.3)
Patient status	
Deceased	57 (28%)
Recurred	134 (65%)
Censored	14 (6.8%)

were obtained for sample size 50. Other than that, for 60% censoring, the average RE ranged from 0.00 (−0.13; 0.16) for time-to-recurrence and size 500, to 0.23 (−0.27; 1.09) for time-to-recurrence and size 100. Excluding the scenario of 60% censoring and size 50, the average RAE ranged from 0.03 (0.00; 0.08) for multiple scenarios, to 0.34 (0.01; 1.09) for time-to-recurrence under 60% censoring and for size 100.

In terms of the *KLD*, similar trends were observed. The *KLD* ranged from 0.03 (0.01; 0.05) for multiple scenarios, to 0.49 (0.03; 2.97) for time-to-death for the scenario of 60% censoring and size 50. As the KLD values were relatively close to 0 for most scenarios, this suggested that the ESPD model performed well in approximating the true population distribution of time-to-events. However, for the scenario with 60% censoring and sample size 50, the KLD value for time-to-death was relatively larger than 0.1 and higher compared to other scenarios, thus suggesting that the model did not fit the data as well, which is a similar pattern observed with alternative modelling strategies.

3.2 Case study

The overall aim of the case study is to provide users with an understanding of the steps involved in implementing the ESPD strategy in a simple example in R. This is provided for both frequentist and Bayesian frameworks and considering Weibull distribution only, for simplicity. In this section, we describe key steps towards the ESPD Weibull implementation as an illustration, and advice the reader to refer to the publicly available scripts for a full description and detailed step-by-step implementation of the strategy. Finally, so that users can fully apply the ESPD approach, we also provide a custom R function that allows for fitting various distributions, including other than Weibull, together with visual fits and Akaike Information Criterion (AIC) scores. These files are also accessible on the listed GitHub repository.

We use a publicly available dataset *melanoma*, available from the *boot* package in R (Hinkley and Anthony, 1997; Ripley and Angelo, 2022). The *melanoma* dataset was originally analysed by Andersen et al. (1993) and consists of measurements made on patients with malignant melanoma, which all had their tumour removed by surgery in Denmark from 1962 to 1977 (Andersen et al., 1993). Several covariates are available, as summarised in Table 2. In terms of outcomes, we consider the following patient status: deceased, disease recurrence, and alive without disease recurrence (i.e., censored).

In estimating the probability of recurrence and distributions of the time-to-recurrence and time-to-death, we assume that.

- The mixture proportions are modelled based on covariates age, sex, ulceration status, and tumour thickness,
- Weibull distributions are appropriate for the time-to-event distributions, assuming a single shape parameter across all groups, where the scale parameter is modelled based on the same covariates as the mixture proportions.

3.2.1 Frequentist implementation

The first step is to define the log-likelihood function in R. The *step_by_step_frequentist* notebook provides a thorough step-by-step implementation of the log-likelihood from its simplest version to incorporating censoring, followed by adding covariates, and finally to considering two competing risks. We highly recommend users who are unfamiliar with these concepts to go through the R code and different steps in the notebook.

Box 1 defines the function in its complete form, which returns the log-likelihood for a set of parameters given the data.

BOX 1 Definition of the log-likelihood function, incorporating the competing events and covariates.

```
ll_weibull_mix_cov <- function(coefs, times, events, X) {
  # Extract the coefficients for the different parameters from the vector
  coefs_logit_recur <- coefs[1:5]          # logit of recurrence probability
  coefs_shape_recur <- coefs[6]            # shape parameter of Weibull distribution for recurrence
  coefs_scale_recur <- coefs[7:11]         # scale parameter of Weibull distribution for recurrence
  coefs_shape_death <- coefs[12]           # shape parameter of Weibull distribution for death
  coefs_scale_death <- coefs[13:17]        # scale parameter of Weibull distribution for death

  # Obtain the patient-level parameters based on the patient covariates
  # and the extracted parameter coefficients
  p_recur     <- sigmoid(coefs_logit_recur)     # probability of recurrence for a given patient
  shape_recur <- exp(coefs_shape_recur)
  scale_recur <- exp(X %*% coefs_scale_recur)
  shape_death <- exp(coefs_shape_death)
  scale_death <- exp(X %*% coefs_scale_death)

  # Log likelihood for those who recurred
  ll_recur <- p_recur * dweibull(times, shape_recur, scale_recur) ^ (events == 'recur')
  # probability density function of the Weibull distribution for recurrence times
  # for patient who experienced recurrence

  # Log likelihood for those who died
  ll_death <- (1 - p_recur) * dweibull(times, shape_death, scale_death) ^ (events == 'death')
  # probability density function of the Weibull distribution for death times
  # for patients who experienced death

  # Log likelihood for censored event
  ll_cens <- (p_recur * pweibull(times, shape_recur, scale_recur, FALSE) +
              (1 - p_recur) * pweibull(times, shape_death, scale_death, FALSE)) ^ (events == 'cens')
  # survival function of the Weibull distribution for recurrence and death times
  # for patients who are censored
  # Combine the log likelihoods
  ll <- sum(log(ll_recur) + log(ll_death) + log(ll_cens))

  return(ll)
}
```

Here, vector t contains the event or censoring times, vector e contains the event data (possible values: *recur* for disease recurrence, *death* for deceased patients, or *cens* for censored patients), and X is the covariance matrix. Furthermore, $coefs$

represents the vector of coefficients that are to be estimated, which need to be defined through a single vector for most optimisation functions. Therefore, the first step in the function is to extract the coefficients for the different parameters from the *coefs* vector. For each parameter that is modelled based on the 4 covariates, there are 5 coefficients: 1 for the intercept and one for each covariate. For the shape parameters that are not modelled based on covariates, there simply is one coefficient. Subsequently, the coefficients are transformed into the parameters for the mixture distribution. Transformations are required because the mixture proportions π are modelled as a logistic regression model and the resulting log-odds need to be transformed to probabilities. Similarly, the shape and scale parameters of the Weibull distributions need to be non-negative and are, therefore, typically modelled using coefficients that are log-transformed. That way, the coefficients can have any negative or positive value in the optimization process, whilst the corresponding parameters will be non-negative. In R, the %*% operator is used for matrix multiplications. This is used in the code to apply covariate matrix X to the coefficients, resulting in vectors of patient-specific parameter values, such as *p_recur*, *shape_recur*, *etc.* For readability of the code, the log-likelihood is obtained in 3 separate steps, one for each of the possible events.

The next step is to apply the ll_weibull_mix_cov function to the data to find the optimal coefficients by maximizing the likelihood function. For this we use the maxLik function of the maxLik package (Henningsen et al., 2010), which was developed with this exact objective in mind, and which conveniently returns the variance-covariance matrix together with the coefficient estimates. In this function, we need to specify the log-likelihood function, start values for the coefficients, and any arguments that need to be passed on to the function, which are *t, e,* and *X* in this case. Because we optimize the function defined by coefficients and not the parameters on real scale, we can simply specify a zero as the starting value for each parameter. Once the optimization is performed, point estimates for the coefficient values can be extracted from the optimization object. This process is illustrated in Box 2.

BOX 2 Performing the maximisation of the likelihood function and extracting the results.

```
# Define the start values for the coefficients and name them
start_values <- setNames(
  object = rep(x = 0, times = 17),
  nm = c(
    'logit_recur_intercept', 'logit_recur_age', 'logit_recur_sex', 'logit_recur_ulceration',
    'logit_recur_thickness', 'logshape_recur', 'logscale_recur_intercept', 'logscale_recur_age',
    'logscale_recur_sex', 'logscale_recur_ulceration', 'logscale_recur_thickness', 'logshape_death',
    'logscale_death_intercept', 'logscale_death_age', 'logscale_death_sex',
    'logscale_death_ulceration', 'logscale_death_thickness'))

# Perform the likelihood optimisation
pars_maxLik <- maxLik(
  logLik = ll_weibull_mix_cov,        # pass the log-likelihood function to be maximised
  start = start_values,               # pass the starting values for the coefficients
  times = times,                      # pass the observed event times
  events = events,                    # pass the observed event types
  X = X)                              # pass the patient-level covariates

# Extract the coefficients and variance-covariance matrix
coef(pars_maxLik)                     # extract the estimated coefficients
vcov(pars_maxLik)                     # extract the estimated variance-covariance matrix
```

Although this step-by-step implementation in R is relatively straightforward, a general function that can be used to apply the ESPD approach for modelling two competing events has been made available with the tutorial on GitHub. The function is available in the script ESPD_frequentist.R, providing all the functionality one may require, such as allowing for different parametric families of distributions for individual risks. Further information is available in the corresponding script.

3.2.2 Bayesian implementation

The Bayesian implementation is fully detailed in the notebook *case_study_bayesian.Rmd*, together with the Stan model *weibull_mix_cov.stan* on GitHub. Stan is a probabilistic programming language for specifying statistical models, providing full Bayesian inference, approximate Bayesian inference and penalised maximum likelihood estimation with optimisation (Team, 2023). In R, Stan can be called through various libraries and in this implementation, we use *CmdStanR* (Češnovar, 2022), which does not interface directly with C++ and is thus user friendly for beginners. In Stan, a typical simulation is a two-step process, by which we first fit the model on existing data to obtain posterior estimates of all parameters, and then sample from the resulting distribution to obtain a synthetic dataset.

The Bayesian implementation inherently captures parameter uncertainty in a principled manner through posterior distributions. These distributions can be used directly to inform parameter values in a probabilistic analysis of the simulation model (Briggs et al., 2012).

3.2.3 Case study results

In this section, we highlight the application and interpretation of Bayesian implementation outcomes. Though the conclusions are applicable to the frequentist case, we believe this example presents an educational opportunity for readers to compare optimisation (Box 2) results in R. This demonstration's objective is less about unearthing groundbreaking findings and more about illuminating the nuances of a practical implementation.

Our discussion focuses on Figure 1; Table 3. In Figure 1, the posterior distribution of various model parameters is displayed. On the x-axis, we see the parameter values, and the y-axis portrays their density. These parameters are part of the Bayesian ESPD model for competing risks, which uses Weibull distributions. Each event, be it recurrence or death, has its parameters estimated individually. In these plots, the parameters alpha (α) and mu (μ) are prominent, serving as the fundamental shape and scale components of the Weibull distribution. In contrast, beta (β) encompasses the regression coefficients tied to the model's covariates, and the pi (π) parameter defines coefficients for the mixture proportions derived from the covariates.

By analysing the coefficients for both risks side-by-side, we aimed to derive an intuitive understanding of their implications. The relationship between these variables and event types can be observed through their respective coefficient values. Specifically, the magnitude and direction of their coefficient values in the model offer insights into their relationships with the event types. A higher coefficient value for a variable suggests a stronger association with the outcome.

A particularly relevant observation from Figure 1 is the mixing proportion intercept (π_{int}) with its mean value hovering around −1.5, indicating a higher likelihood of recurrence as opposed to death. Figure 1 also underscores that ulceration and

thickness are pivotal factors influencing the outcomes, which can be seen based on the β coefficients. These coefficients depict how changes in ulceration and thickness are associated with changes in the hazard of the events. Furthermore, by examining the π coefficients for these covariates, we can glean insights into their influence on the likelihood of one event type over another, such as recurrence *versus* death.

Examining the Weibull parameters (α and μ), we conclude that the mean death time is shorter than the recurrence time. For a complete view of the distribution, we suggest readers to simulate times for death and recurrence based on mean Weibull parameters with zero covariate effects and plotting a histogram. This can be accomplished using the rweibull_cov function from GitHub. Finally, the likelihood of recurrence increases with the observation period length.

In Table 3, we present the probabilities of recurrence for censored individuals from the melanoma dataset, as sampled from the Bayesian posterior distribution. Recurrence is frequently the more probable outcome. The primary covariates influencing both the event type and its timing are ulceration and thickness, reflecting findings from Figure 1. Patients with a tumour thickness significantly above the mean exhibit a heightened risk of death before recurrence, indicating that thickness may be an indication of melanoma severity.

4 Discussion

Competing risks data are common in medical research that aims to investigate an outcome of interest and, hence, decision-analytic models of healthcare pathways commonly include multiple competing events. For example, in oncology, recurrence is a competing risk to death prior to recurrence, which is typically modelled based on background mortality. Here, we addressed a methodological gap by defining and illustrating a modelling approach for implementing the 'event first, time second' strategy for modelling competing events in DES when the parameters are to be estimated based on censored data. The resulting ESPD modelling approach was mathematically defined for any number of competing risks in Eqs 7, 8, and implementations in both the frequentist and Bayesian framework were provided for two competing risks, including when considering covariates. Finally, the approach was evaluated in a simulation study and illustrated in a case study for which the corresponding R code has been made available with this manuscript.

The results of the simulation study indicate that the frequentist implementation of the ESPD approach performs well under various degrees of censoring and sample sizes. However, its accuracy diminishes with decreasing sample sizes and increasing levels of censoring (Table 1). These results are consistent with past studies on implementing competing events in DES with uncensored and censored data using alternative strategies (Degeling et al., 2019; Degeling et al., 2022). Our findings reiterate the importance for modellers to recognise that datasets characterised by high censoring levels and small sample sizes might render the ESPD approach less reliable for simulations, which also holds for other methods previously investigated (Degeling et al., 2019). Although a formal comparison of the frequentist and Bayesian implementations was beyond the scope of the simulation study, the case study demonstrated that both implementations yielded comparable and realistic results. Further research may compare the frequentist and Bayesian implementations more systematically to identify whether either may be preferable in certain scenarios. Significantly, our study introduces an additional method to the existing techniques for addressing censoring, filling the gap where no method was previously delineated for such censoring.

Further research is also warranted to quantitatively compare the performance of the ESPD to previously defined modelling approaches for implementing competing events in DES based on censored data (Degeling et al., 2022), in line with previous work focused on uncensored data (Degeling et al., 2019). This would also inform selection between the different modelling approaches. Such guidance is already available for the ESD, ESPD, UDR and MDR approaches for scenarios in which they are informed by uncensored data, as well as for the ESD and UDR approaches when informed by censored data. Based on the previous work for uncensored data and the results of our simulation study, we expect that the ESPD approach will have good accuracy and be relatively straightforward to implement and interpret when used for censored data compared to the other approaches, but this is to be confirmed in a comparative simulation study. In this context, it is important to note that the interpretation of likelihood-based measures, such as the AIC may be different between the approaches. For the ESD approach, the likelihood only considers the time-to-event for each event independently and not the type-of-event, whereas the likelihood in the UDR approach considers the likelihood of the time-to-event and event-type separately, and the ESPD considers the time-to-event and event-type of all events jointly. Regardless, despite high-level guidance on the selection of different approaches being useful, validation of the results will remain pertinent in the modelling process.

By demonstrating the implementation of the ESPD approach using both the frequentist and Bayesian frameworks in R and Stan, we enable novice and more advanced R users to leverage this modelling strategy. The frequentist implementation allows for a fast and relatively easy retrieval of the point estimates and the variance-covariance matrix of the coefficient, especially with the provided general functions that facilitates incorporation of covariates and different distribution types. Whilst more challenging to implement, some may argue that the Bayesian version provides a more natural and principled way of combining useful prior information into the estimate, which may be more accurate than a frequentist estimate, if such information is available. Furthermore, some consider the interpretation of a Bayesian result more straightforward, as it provides a framework about the unknown parameter conditional on the observed data, rather than about the observed data conditional on the unknown parameter. By providing both implementations, we provide modellers with the freedom to use the framework they prefer.

The ESPD approach was developed for modelling competing events in DES. However, the event-specific probabilities may also be considered as cumulative event incidences in an epidemiological context. The cumulative incidence of competing risks has generally been modelled using cause-specific hazard models and sub-distribution hazard models (Fine and Gray, 1999; Pintilie, 2006; Lau et al., 2009; Austin and Fine, 2017). The ESPD approach may

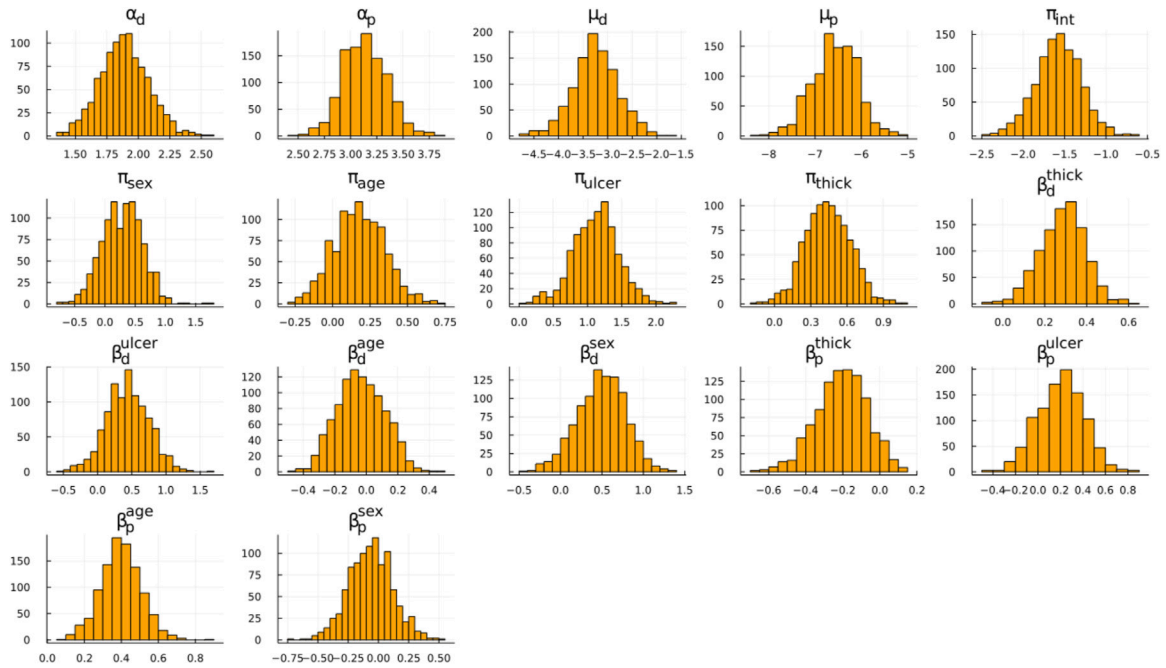

FIGURE 1
Posterior distribution samples for the ESPD model. Weibull parameters: α, μ; Mixing proportion: π; Adjustment coefficients: β; Subscripts: *p* (recurrence), *d* (death).

TABLE 3 Mean posterior probabilities for recurrence event for each censored individual.

Age	Sex	Thickness	Ulcer	Time to last status assessment (years)	Probability of recurrence
76	Male	6.76	Present	0.03	0.35
56	Male	0.65	Absent	0.08	0.83
71	Female	2.90	Absent	0.27	0.80
60	Female	3.22	Present	0.64	0.59
64	Female	0.16	Present	0.97	0.68
72	Male	12.56	Present	1.35	0.29
86	Female	8.54	Present	2.26	0.44
64	Male	1.29	Absent	3.91	0.87
76	Female	1.29	Present	4.18	0.70
71	Male	4.84	Present	5.10	0.84
66	Female	0.65	Absent	5.71	0.88
49	Male	1.62	Absent	8.64	0.97
49	Male	6.12	Absent	8.72	0.99
54	Female	1.45	Absent	9.47	0.89

provide an interesting alternative to these models, where the cumulative event incidences can be modelled directly as the mixture proportions. This could be utilized for estimating probabilities of treatment sequences from real-world data where typically a substantial proportion of patients is still on treatment, i.e., censored for the competing events of switching to a subsequent treatment line and death without further treatment. This is relevant for disease areas where patients typically receive multiple lines of therapy, such as oncology.

In summary, our study has filled a methodological gap by providing a tutorial and framework for modelling competing events in discrete event simulations with censored data. The ESPD

approach, which samples the event first and time-to-event second, was found to be accurate and produced realistic results in both simulation and case studies. The ESPD approach has been implemented in both a frequentist and Bayesian framework using R, making it easily accessible for others to use and expand upon in future research. Not only is the ESPD strategy applicable for modelling competing events in DES, but it also has potential to be used in other contexts to estimate cumulative event incidences. Future studies should perform and report on cross-validation of the ESPD approach compared to the other strategies, which will ultimately ensure individual patient data are appropriately modelled.

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

FF: Conceptualization, Formal Analysis, Funding acquisition, Investigation, Methodology, Project administration, Resources, Software, Validation, Visualization, Writing–original draft, Writing–review and editing. VF: Conceptualization, Formal Analysis, Investigation, Methodology, Project administration, Resources, Software, Validation, Visualization, Writing–review and editing. MI: Conceptualization, Funding acquisition, Investigation, Project administration, Supervision, Writing–review and editing. KD: Conceptualization, Formal Analysis, Funding acquisition, Investigation, Methodology, Resources, Software, Supervision, Validation, Visualization, Writing–original draft, Writing–review and editing.

Funding

The time of FF, KD, and MI on this project was supported by the Medical Research Future Fund, Preventative and Public Health THSCOR 2019 (Targeted Health Systems and Community Organisation Research), grant number 2020/MRF1199701.

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.

References

Andersen, P. K., Borgan, O., Gill, R. D., and Keiding, N. (1993). *Statistical models based on counting processes*. Berlin, Germany: Springer-Verlag.

Austin, P. C., and Fine, J. P. (2017). Practical recommendations for reporting Fine-Gray model analyses for competing risk data. *Stat. Med.* 36 (27), 4391–4400. doi:10.1002/sim.7501

Barton, P., Jobanputra, P., Wilson, J., Bryan, S., and Burls, A. (2004). The use of modelling to evaluate new drugs for patients with a chronic condition: the case of antibodies against tumour necrosis factor in rheumatoid arthritis. *Health Technol. Assess. Winch. Engl.* 8 (11), 1–91. doi:10.3310/hta8110

Briggs, A. H., Weinstein, M. C., Fenwick, E. A. L., Karnon, J., Sculpher, M. J., Paltiel, A. D., et al. (2012). Model parameter estimation and uncertainty: a report of the ISPOR-SMDM modeling good research practices task force-6. *Value Health* 15 (6), 835–842. doi:10.1016/j.jval.2012.04.014

Caro, J. J., and Möller, J. (2014). Decision-analytic models: current methodological challenges. *Pharmacoeconomics* 32 (10), 943–950. doi:10.1007/s40273-014-0183-5

Češnovar, J. G. (2022). *Cmdstanr: R interface to 'CmdStan*.

Coemans, M., Verbeke, G., Döhler, B., Süsal, C., and Naesens, M. (2022). Bias by censoring for competing events in survival analysis. *BMJ* 378, e071349. doi:10.1136/bmj-2022-071349

Cover, T. M. (1999). *Elements of information theory*. Hoboken: John Wiley & Sons.

Degeling, K., Ijzerman, M. J., Groothuis-Oudshoorn, C. G. M., Franken, M. D., Koopman, M., Clements, M. S., et al. (2022). Comparing modeling approaches for discrete event simulations with competing risks based on censored individual patient data: a simulation study and illustration in colorectal cancer. *Value Health* 25 (1), 104–115. doi:10.1016/j.jval.2021.07.016

Degeling, K., Koffijberg, H., Franken, M. D., Koopman, M., and Ijzerman, M. J. (2019). Comparing strategies for modeling competing risks in discrete-event simulations: a simulation study and illustration in colorectal cancer. *Med. Decis. Mak.* 39 (1), 57–73. doi:10.1177/0272989X18814770

Fine, J. P., and Gray, R. J. (1999). A proportional hazards model for the subdistribution of a competing risk. *J. Am. Stat. Assoc.* 94 (446), 496–509. doi:10.1080/01621459.1999.10474144

Günal, M. M., and Pidd, M. (2017). Discrete event simulation for performance modelling in health care: a review of the literature. *J. Simul.* 4 (1), 42–51. doi:10.1057/jos.2009.25

Henningsen, A., and Toomet, O. (2010). maxLik: a package for maximum likelihood estimation in R. *Comput. Stat.* 26 (3), 443–458. doi:10.1007/s00180-010-0217-1

Hinkley, D. V., and Anthony, C. D. (1997). *Bootstrap methods and their applications*. Cambridge, UK: Cambridge University Press.

Karnon, J., Stahl, J., Brennan, A., Caro, J. J., Mar, J., and Möller, J. (2012). Modeling using discrete event simulation: a report of the ISPOR-SMDM modeling good research practices task force–4. *Med. Decis. Mak.* 32 (5), 701–711. doi:10.1177/0272989X12455462

Koller, M. T., Raatz, H., Steyerberg, E. W., and Wolbers, M. (2012). Competing risks and the clinical community: irrelevance or ignorance? *Stat. Med.* 31 (11-12), 1089–1097. doi:10.1002/sim.4384

Kullback, S., and Leibler, R. A. (1951). On information and sufficiency. *Ann. Math. Statistics* 22 (1), 79–86. doi:10.1214/aoms/1177729694

Lau, B., Cole, S. R., and Gange, S. J. (2009). Competing risk regression models for epidemiologic data. *Am. J. Epidemiol.* 170 (2), 244–256. doi:10.1093/aje/kwp107

Marshall, D. A., Grazziotin, L. R., Regier, D. A., Wordsworth, S., Buchanan, J., Phillips, K., et al. (2020). Addressing challenges of economic evaluation in precision medicine using dynamic simulation modeling. *Value Health* 23 (5), 566–573. doi:10.1016/j.jval.2020.01.016

Pintilie, M. (2006). *Competing risks: a practical perspective*. Hoboken: John Wiley & Sons.

Ripley, B., and Angelo, J. C. (2022). *Boot: bootstrap R (S-plus) functions*.

Siebert, U., Alagoz, O., Bayoumi, A. M., Jahn, B., Owens, D. K., Cohen, D. J., et al. (2012). State-transition modeling: a report of the ISPOR-SMDM modeling good research practices task force-3. *Value Health* 15 (6), 690–700. doi:10.1177/0272989X12455463

Team, S. D. (2023). *RStan: the R interface to Stan*.

Vazquez-Serrano, J. I., Peimbert-Garcia, R. E., and Cardenas-Barron, L. E. (2021). Discrete-event simulation modeling in healthcare: a comprehensive review. *Int. J. Environ. Res. Public Health* 18 (22), 12262. doi:10.3390/ijerph182212262

frontiers | Frontiers in Pharmacology

TYPE Original Research
PUBLISHED 01 November 2023
DOI 10.3389/fphar.2023.1225919

Check for updates

OPEN ACCESS

EDITED BY
Blythe Adamson,
Flatiron Health, United States

REVIEWED BY
Yoshihiro Noguchi,
Gifu Pharmaceutical University, Japan
Huang Shang-yi,
National Taiwan University Hospital, Taiwan

*CORRESPONDENCE
Xiao Wang,
wangxiao0719@163.com
Kuan Li,
li_kuan1989@126.com

†These authors have contributed equally to this work and share first authorship

RECEIVED 20 May 2023
ACCEPTED 18 October 2023
PUBLISHED 01 November 2023

CITATION
Su S, Wu L, Zhou G, Peng L, Zhao H, Wang X and Li K (2023), Indication and adverse event profiles of denosumab and zoledronic acid: based on U.S. FDA adverse event reporting system (FAERS). *Front. Pharmacol.* 14:1225919. doi: 10.3389/fphar.2023.1225919

COPYRIGHT
© 2023 Su, Wu, Zhou, Peng, Zhao, Wang and Li. This is an open-access article distributed under the terms of the Creative Commons Attribution License (CC BY). The use, distribution or reproduction in other forums is permitted, provided the original author(s) and the copyright owner(s) are credited and that the original publication in this journal is cited, in accordance with accepted academic practice. No use, distribution or reproduction is permitted which does not comply with these terms.

Indication and adverse event profiles of denosumab and zoledronic acid: based on U.S. FDA adverse event reporting system (FAERS)

Si Su[1,2†], Liuqing Wu[3†], Guibao Zhou[2], Lingling Peng[2], Huanzhe Zhao[2], Xiao Wang[1,2*] and Kuan Li[2*]

[1]School of Pharmacy, Guangdong Medical University, Zhanjiang, China, [2]Department of Pharmacy, Shenzhen People's Hospital (The Second Clinical Medical College, Jinan University; The First Affiliated Hospital, Southern University of Science and Technology), Shenzhen, Guangdong, China, [3]Longgang Central Hospital of Shenzhen, Shenzhen, Guangdong, China

Objective: To investigate adverse events (AEs) associated with denosumab (Dmab) and zoledronic acid (ZA), compare their association strengths, and explore potential applications to provide clinical reference.

Methods: We collected data from FAERS from January 2004 to November 2022 and mined AE signals for Dmab and ZA using ROR values. We compared signal intensity for same AEs and investigated off-label use. We also examined their AEs in adjuvant therapy for breast and prostate cancer.

Results: 154,735 reports of primary suspect drugs were analyzed in the FAERS database (Dmab: 117,857; ZA: 36,878). Dmab and ZA had 333 and 1,379 AE signals, with 189 overlaps. The AEs of Dmab included death (ROR:3.478), osteonecrosis of jaw (ROR:53.025), back pain (ROR:2.432), tooth disorder (ROR:16.18), bone pain (ROR:6.523). For ZA, the AEs included osteonecrosis (ROR:104.866), death (ROR: 3.645), pain (ROR:3.963), osteonecrosis of jaw (ROR: 91.744), tooth extraction (ROR: 142.143). Among overlap signals, Dmab showed higher strength in exostosis of the jaw (ROR: 182.66 vs. 5.769), atypical fractures (ROR: 55.589 vs. 9.123), and atypical femur fractures (ROR:49.824 vs. 4.968). And ZA exhibited stronger associations in abscess jaw (ROR: 84.119 vs. 11.12), gingival ulceration (ROR: 74.125 vs. 4.827), increased bone formation (ROR: 69.344 vs. 3.218). Additionally, we identified 528 off-label uses for Dmab and 206 for ZA, with Dmab mainly used in prostate cancer (1.04%), breast cancer (1.03%), and arthritis (0.42%), while ZA in breast cancer (3.21%), prostate cancer (2.48%), and neoplasm malignant (0.52%). For Dmab in breast cancer treatment, AEs included death (11.6%), disease progression (3.3%), and neutropenia (2.7%), while for ZA included death (19.8%), emotional disorder (12.9%), osteomyelitis (11.7%). For prostate cancer treatment, Dmab`s AEs were death (8.9%), prostate cancer metastatic (1.6%), renal impairment (1.7%), while ZA`s included death (34.4%), general physical health deterioration (19.9%), and hemoglobin decreased (18.9%).

Conclusion: Our analysis of FAERS database provided postmarketing surveillance data and revealed different strengths of reported AE signals between Dmab and ZA

in some of their common AEs. It's also worth noting that both drugs have potential off-label applications, which could introduce new AEs. This highlights the necessity for safety monitoring when using Dmab and ZA off-label.

KEYWORDS

denosumab, zoledronic acid, adverse events, off-label use, pharmacovigilance

1 Introduction

Denosumab (Dmab), the first and only one receptor activator of NF-κB ligand (RANKL) inhibitor so far, was approved for marketing by the U.S. Food and Drug Administration (FDA) in 2010 and Zoledronic acid (ZA) approved in 2001. They have similar efficacy such as applying for prevention and treatment of osteoporosis in postmenopausal females, osteoporosis in males, glucocorticoid-induced osteoporosis, hypercalcemia of malignancy, and preventing skeletal-related events (SREs) secondary to solid tumors metastases (Greear and Bankole, 2022; Hildebrand et al., 2022). However the mechanism differs between the two (Baron et al., 2011), with Dmab exerting its anti-bone resorption effect by attaching to RANKL which activate osteoclasts through the binding with RANK, thereby suppressing bone resorption (Jamal et al., 2011; Pang et al., 2020). Zoledronic acid, on the other hand, binding of inorganic pyrophosphate to hydroxyapatite crystals in bone, especially in the sites where bone is remodeling actively, and thus play an anti-bone resorption role (Drake et al., 2008). Dmab and ZA have two different drug specifications each. Dmab is available as Xgeva (120 mg) for preventing bone-related events in cancer patients and Prolia (60 mg) for treating osteoporosis. Similarly, ZA has two different specifications; Reclast (5 mg) for treating osteoporosis and Zometa (4 mg) for cancer-related bone damage.

In the past decade, significant efficacy of both drugs has been extensively documented, whereas, novel AEs not well studied were gradually raised during the clinical application. Furthermore, novel mechanisms as well as application also emerged. We hope this analysis based on FAERS database will provide safety profile in support of future studies in the application of Dmab and ZA. And to provide reference directions for exploring their potential clinical applications.

2 Materials and methods

2.1 Data sources and procedures

The data for this retrospective pharmacovigilance study were obtained from FAERS, a global spontaneous reporting system that collects safety information on approved drugs and therapeutic biologic products from various sources including manufacturers, healthcare professionals, and consumers. FAERS is the primary source of post-marketing safety monitoring and evaluation for the FDA, and it provides signal detection and quantification of the association between drugs and AEs (Tang et al., 2022). The database contains seven categories of data including demographic and management information, drug information, adverse events, patient outcomes, report sources, treatment start and end dates, and indication.

2.2 Data extraction and processing

To extract adverse event (AE) reports from the FDA Open-FDA program, we utilized the online tool OpenVigil 2.1 (http://openvigil.sourceforge.net/). Individual safety reports (ISRs) for Dmab and ZA were extracted from the FAERS database. ISRs are the count of raw data extracted by OpenVigil 2.1 and an ISR code represents an AE report.

The study retrieved data from FAERS covering the period between January 2004 and November 2022. The search for Dmab included its generic name "DENOSUMAB" and commodity names "Xgeva," "Ranmark," and "Prolia," while for ZA, the search included its generic drug name "ZOLEDRONIC ACID," "ZOLEDRONATE," and trade names "ACLASTA," "RECLAST", and "ZOMETA." Drugs irrelevant to the study and those with uncertain names were excluded. Only drugs listed as the "primary suspect" were included in the analysis as they were most likely associated with the AEs (Verden et al., 2018; Omar et al., 2021).

2.3 AE signals detection

Disproportionality analysis was conducted to identify potential safety signals for the drugs, with RORs as measures of association (van Puijenbroek et al., 2002; Hauben, 2003; Tang et al., 2022). The analysis of the association between drug exposure and adverse events (referred to as "signals") in OpenVigil relies on the use of a 2 × 2 contingency table (Böhm, 2018; Noguchi et al., 2021) (Refer to Table 1), which can be effortlessly generated within the platform. The higher the ROR values, the stronger the correlation between the drug and target AE. Significant signals were identified based on criteria including AE reports >3, ROR and PRR >2.0, ROR lower bound of 95% confidence interval (CI) value exceeds 1.0, and $\chi^2 > 4$ (Böhm, 2018; Shao et al., 2021; Tang et al., 2022). The equations and criteria for the three algorithms are shown in Table 2. Data processing was carried out using Microsoft Excel 2016 and GraphPad Prism 9.

3 Results

3.1 AE reports and clinical information

The FAERS database contained 385,327 reports of primary suspect drugs from its inception until October 2022, with 297,896 AEs associated with Dmab and 87,431 AEs related to ZA. After removing duplicates, a total of 154,735 reports were included, consisting of 117,857 AEs for Dmab and 36,878 AEs for ZA. Process flowchart is shown in Figure 1.

The characteristics and clinical information are summarized in Table 3. The majority of the reports for both drugs were from

TABLE 1 Two-by-two contingency table.

	Drug exposure	No drug exposure	Sums
Adverse event occurred	DE	dE	E
No adverse event occurred	De	de	e
Sums	D	d	N

Note: D represents occurrence of drug exposure and E represents adverse event of interest, d represents no drug exposure and e represents no occurrence of the adverse event.

TABLE 2 Equation and criteria of three algorithms for signal detection.

Algorithms	Equation	Criteria
ROR	ROR = (DE/De)/(dE/de)	ROR ≥ 2
	$95\%CI = e^{In(ROR)\pm1.96\sqrt{\frac{1}{DE}+\frac{1}{De}+\frac{1}{dE}+\frac{1}{de}}}$	95%CI > 1
		DE ≥ 3
PRR	PRR = (DE/D)/(dE/d)	PRR ≥ 2
		DE ≥ 3
χ^2	$\chi^2_{Yates} = N * (\lvert DE*de - dE*De \rvert - N/2)^2/(D * d * E * e)$	$\chi^2 \geq 4$

Note: ROR, reporting odds ratio; PRR, proportional reporting ratio; CI: confidence interval; χ^2, chi-squared; DE, number of co-occurrences.

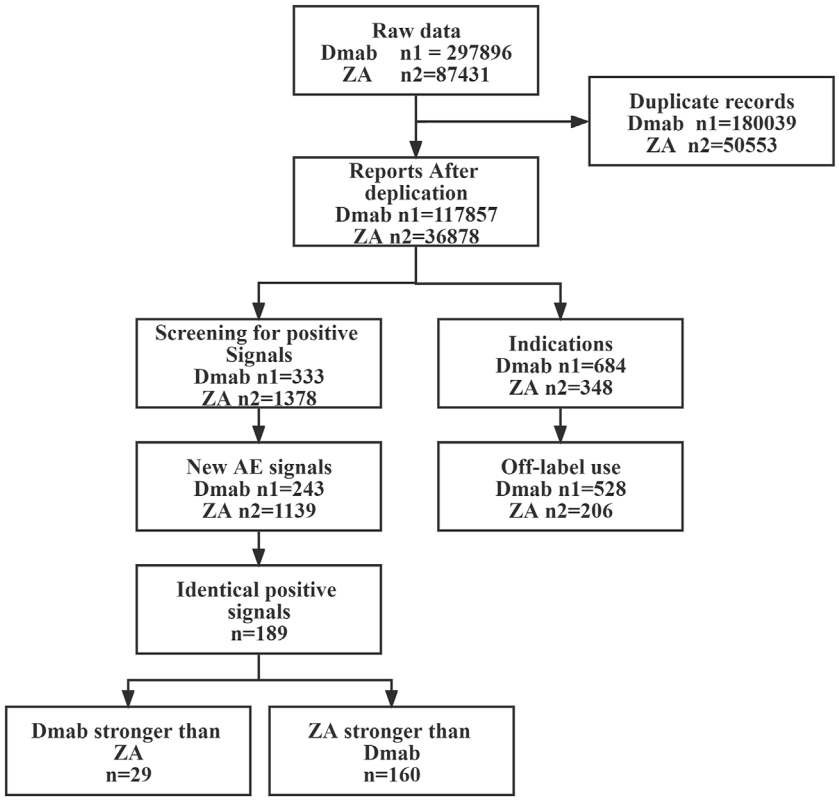

FIGURE 1
Flowchart of data mining process.

TABLE 3 Characteristics and clinical information.

Characteristics	Reports (*N*, %)	
	Denosumab (n = 117,857)	Zoledronic acid (n = 36,878)
Gender		
Female	89,799 (76.19)	24,743 (67.09)
Male	14,136 (11.99)	10,083 (27.34)
Unknow	13,921 (11.81)	2052 (5.56)
Age		
Median (IQR)	68 (59–77)	73 (65–81)
<18	176 (0.15)	47 (0.13)
18–40	789 (0.67)	427 (1.16)
41–65	16,929 (14.36)	7189 (19.49)
>65	48,510 (41.16)	9847 (26.70)
Unknow	51,452 (43.66)	19,368 (52.52)
Report countries		
United States	88,883 (75.42)	12,431 (33.71)
Canada	6223 (5.28)	3894 (10.56)
others	22,667 (19.23)	16,109 (43.62)
Unknow	83 (0.07)	4444 (12.05)

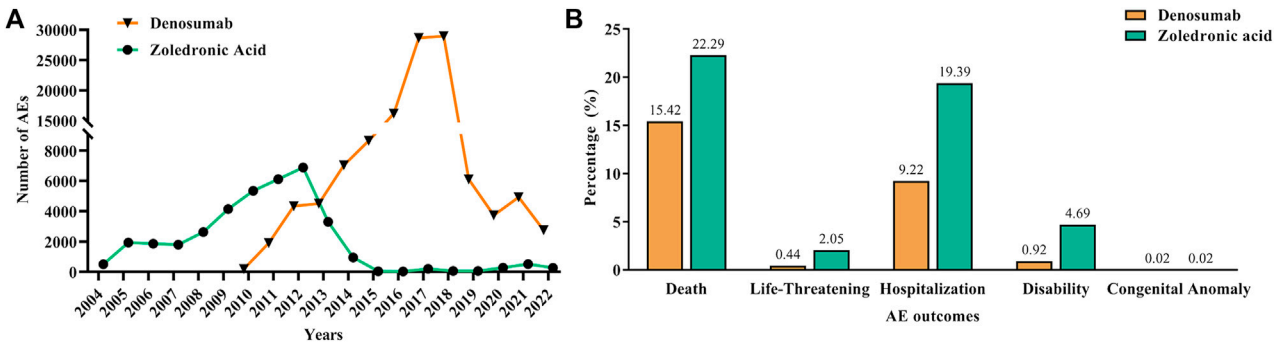

FIGURE 2
Report years and serious AE outcome information of denosumab and zoledronic acid. **(A)**: Number of reported AEs of denosumab and zoledronic acid from 2004 to 2022. **(B)**: The serious AE outcome indicators of denosumab and zoledronic acid.

females (76.19% for Dmab and 67.09% for ZA), and the median age of the reports was 68 and 73 years for Dmab and ZA, respectively, with a focus on the elderly population.

In order to make the changes more intuitive, we visualized the AEs metric data of each year with a line chart, as shown in Figure 2A. The chart shows an increasing trend in AEs for both drugs year by year, but a decline in 2012 and 2018 for ZA and Dmab, respectively. In addition, we also visualized the serious AE outcome metric data for the two drugs, as shown in Figure 2B. Serious AEs were mainly attributed to death (15.42% for Dmab and 22.29% for ZA) and hospitalization (9.22% for Dmab and 19.39% for ZA). Furthermore, ZA had slightly higher proportions of life-threatening (2.05% vs. 0.44%) and disability (4.69% vs. 0.92%) according to the reports from the database compared to Dmab.

3.2 Differences of overall AE signals between dmab and ZA

We then conducted a disproportionality analysis using ROR to detect AE signals, which led to the identification of 333 significant AE signals related to Dmab and 1379 associated with ZA.

TABLE 4 Top 10 significant AE signals of Dmab and ZA.

	AEs	N	ROR (95% CI)	PRR (χ2)
Denosumab	Death*	16,013	3.478 (3.42)	3.142 (23,699.174)
	Osteonecrosis of jaw	6043	53.025 (51.377)	50.358 (193,469.11)
	Back pain	2793	2.432 (2.342)	2.398 (2244.037)
	Tooth disorder	1929	16.18 (15.414)	15.931 (23,239.1)
	Bone pain	1878	6.523 (6.223)	6.435 (8106.468)
	Hypocalcaemia	1805	23.173 (22.009)	22.834 (30,578.129)
	Spinal fracture	1369	15.963 (15.073)	15.79 (16,337.329)
	Pain in jaw	1314	9.335 (8.819)	9.242 (8831.068)
	Fracture	894	6.768 (6.323)	6.725 (4076.903)
	Tooth extraction	815	11.487 (10.681)	11.415 (6932.845)
Zoledronic acid	Osteonecrosis	6980	104.866 (101.892)	85.207 (458,676.62)
	Death*	5283	3.645 (3.539)	3.266 (8597.053)
	Pain	4177	3.963 (3.836)	3.627 (8110.249)
	Osteonecrosis of jaw	3696	91.744 (88.342)	82.649 (236,587.42)
	Tooth extraction	2321	142.143 (135.258)	133.26 (214,317.12)
	Bone disorder	2129	68.221 (65.038)	64.34 (110,354.49)
	Pyrexia	1957	3.932 (3.756)	3.777 (4002.168)
	Arthralgia	1895	3.364 (3.211)	3.242 (2953.963)
	Pain in jaw	1792	44.097 (41.93)	42.002 (63,354.395)
	Fall*	1543	3.039 (2.887)	2.953 (2001.728)

Note: *, The instruction does not mention; 95% CI: only show the low bound of ROR.

Interestingly, 243 new AE signals and 528 off-label use for Dmab that were not registered in the FDA-approved specification were found, along with 1139 new signals and 206 off-label uses for ZA.

The most common AEs associated with Dmab were death, osteonecrosis of the jaw, back pain, tooth disorder, bone pain, and hypocalcemia. For ZA, the most frequent adverse events were osteonecrosis, death, pain, osteonecrosis of the jaw, and tooth extraction. Among these, death and fall were not mentioned in the drug labels for either Dmab or ZA. Top l0 significant AE signals sorted by frequency for both drugs are presented in Table 4.

As there were numerous shared AEs between Dmab and ZA, we conducted a further comparison of the overlapping AE signals. Out of the 189 identical positive AE signals between the two drugs, 29 AEs of Dmab exhibited stronger correlation than ZA, while 160 AEs of Dmab had weaker correlation than ZA, as determined by the ROR value. Table 5 presents the AE signals with significant differences in intensity between the two drugs. The AE signals of Dmab with stronger correlation than ZA ($d > 20$) included exostosis of jaw (ROR: 182.66 vs. 5.769), atypical fracture (ROR: 55.589 vs. 9.123), and atypical femur fracture (ROR: 49.824 vs. 4.968), while the AE signals of ZA with stronger correlation than Dmab ($d > 50$) were related to abscess jaw (ROR: 84.119 vs. 11.12), gingival ulceration (ROR: 74.125 vs. 4.827), increased bone formation (ROR: 69.344 vs. 3.218), and bone disorder (ROR: 68.221 vs. 3.189), among others.

3.3 Off-label use

While analyzing the data, we found off-label use was also a significant signal. Therefore, we further analyzed the data on off-label use. As we observed mixed reports of different specifications for each drug, for example, Dmab had a 60 mg specification for giant cell tumor and hypercalcemia of malignancy, while the 120 mg specification was used for postmenopausal osteoporosis. Similarly, ZA had two different specifications with mixed reports. Therefore, we combined the FDA-approved indications for both specifications of each drug and compared them with the indications in the database to identify off-label uses. We found 528 types of off-label use for Dmab and 206 types for ZA. Table 6 shows the top 10 off-label uses not mentioned in the drug instructions for both drugs, which are frequently used for treating various tumors. Breast cancer (1.03% and 3.21%) and prostate cancer (1.04% and 2.48%) were the most commonly off-label use for both drugs in the database. Other off-label uses for Dmab included arthritis (0.42%), vitamin D deficiency (0.26%), spinal compression fracture (0.25%), gastroesophageal reflux disease (0.24%), plasma

TABLE 5 AE signals with significant differences in ROR values between Dmab and ZA.

Item	AEs	Dmab	ZA
		ROR (95%CI)	ROR (95%CI)
AE signals of Dmab stronger than ZA (*d>20*)	exostosis of jaw	182.66 (138.708)	5.769 (2.146)
	atypical fracture	55.589 (43.652)	9.123 (4.517)
	atypical femur fracture	49.824 (44.731)	4.968 (3.29)
	dental care	66.203 (58.259)	22.846 (17.792)
	dental implantation	49.816 (42.49)	22.907 (16.979)
	bone density abnormal	23.178 (20.97)	2.812 (1.848)
AE signals of ZA stronger than Dmab (*d>50*)	abscess jaw	11.12 (8.861)	84.119 (71.047)
	gingival ulceration	4.827 (3.115)	74.125 (58.52)
	bone formation increased	3.218 (1.427)	69.344 (47.835)
	bone disorder	3.189 (2.884)	68.221 (65.038)
	gingival erosion	4.065 (1.909)	67.58 (45.849)
	gingival erythema	2.791 (1.572)	65.051 (50.643)
	dental fistula	5.062 (3.152)	66.826 (51.022)
	periodontitis	4.408 (3.463)	62.003 (54.294)
	oroantral fistula	3.718 (1.522)	60.457 (38.31)
	bone callus excessive	11.324 (4.864)	65.817 (33.298)
	bone scan abnormal	5.185 (3.029)	59.205 (42.901)
	osteopetrosis	8.412 (3.654)	59.239 (31.974)

Note:*d*, difference of ROR, between denosumab zoledronic acid; 95% CI, only show the low bound of ROR.

TABLE 6 Top 10 off-label uses not mentioned in the drug instructions.

Denosumab (n = 117,857)		Zoledronic acid (n = 36,878)	
Indication	N (%)	Indication	N (%)
Prostate cancer	1225 (1.04)	Breast cancer	1184 (3.21)
Breast cancer	1215 (1.03)	Prostate cancer	916 (2.48)
Arthritis	499 (0.42)	Neoplasm malignant	190 (0.52)
Vitamin d deficiency	303 (0.26)	Renal cancer	166 (0.45)
Spinal compression fracture	299 (0.25)	Lung cancer	147 (0.40)
gastroesophageal reflux disease	286 (0.24)	Plasma cell myeloma	111 (0.30)
Plasma cell myeloma	224 (0.19)	Plasma cytoma	69 (0.19)
Rheumatoid arthritis	221 (0.19)	Colon cancer	28 (0.08)
Other solid tumors*	640 (0.54)	Osteoarthritis	25 (0.07)
Chronic kidney disease	116 (0.10)	Other solid tumors*	280 (0.76)

Note: *, refers to a group of tumors that includes non-small cell lung cancer, bronchial carcinoma, gastric cancer, rectal cancer, lymphoma, and more.

cell myeloma (0.19%), rheumatoid arthritis (0.19%), and chronic kidney disease (0.10%). Off-label uses for ZA included neoplasm malignant (0.52%), renal cancer (0.45%), lung cancer (0.40%), plasma cell myeloma (0.30%), plasma cytoma (0.19%), colon cancer (0.08%), and osteoarthritis (0.07%).

We conducted a comparison of the AEs associated with Dmab and ZA in off-label use for breast cancer and prostate cancer. After comparing with the FDA-approved instructions and removing similar AEs, we found 451 AEs in Dmab and 848 AEs in ZA for breast cancer treatment. For prostate cancer treatment, we found

TABLE 7 Top 6 AEs not registered in specification related to breast cancer treatment.

Indication	Denosumab (n = 1215)		Zoledronic acid (n = 1184)	
	AEs	N (%)	AEs	N (%)
Breast cancer	Death	141 (11.6)	Death	235 (19.8)
	Disease progression	40 (3.3)	Emotional disorder [b]	153 (12.9)
	Breast cancer metastatic	33 (2.7)	Osteomyelitis	135 (11.4)
	Neutropenia	33 (2.7)	Neoplasm progression	127 (10.7)
	Emotional disorder [a]	28 (2.3)	Cardiac disorders [c]	125 (10.6)
	Pyrexia	21 (1.7)	Impaired healing	112 (9.5)

Note: a. Emotional disorder in patients with breast cancer treated with Dmab included confusional state (0.74%), delirium (0.33%), disturbance in attention (0.25%), palpitations (0.25%), irritability (0.16%), anxiety (0.16%), depressed mood (0.08%), hallucination visual (0.08%), major depression (0.08%), paranoia (0.08%) and psychotic disorder (0.08%). **b**. Emotional disorder in patients with breast cancer treated with ZA, included emotional distress (3.04%), confusional state (2.53%), depression (1.86%), depressed mood (2.20%), suicidal ideation (0.84%), personality disorder (0.68%), amnesia (0.59%), abasia (0.42%), mental disorder (0.17%), emotional disorder (0.17%), disturbance in attention (0.17%), suicide attempt (0.08%), mood altered (0.08%) and depression suicidal (0.08%). **c**. Cardiac disorders in patients with breast cancer treated with ZA, included tachycardia (3.04%), cardiac failure congestive (1.35%), arrhythmia (1.01%), palpitations (1.01%), left ventricular dysfunction (0.76%), cardiomegaly (0.59%), cardiac disorder (0.51%), cardiac failure (0.51%), myocardial infarction (0.42%), atrial fibrillation (0.34%), heart rate decreased (0.34%), cardio-respiratory arrest (0.17%), endometrial hypertrophy (0.17%), cardiovascular somatic symptom disorder (0.08%), cardiovascular disorder (0.08%), and cardiac discomfort (0.08%).

TABLE 8 Top 6 AEs not registered in specification related to prostate cancer treatment.

Indication	Denosumab (n = 1225)		Zoledronic acid (n = 916)	
	AEs	N (%)	AEs	N (%)
Prostate cancer	Death	109 (8.9)	Death	315 (34.4)
	Prostate cancer metastatic	20 (1.6)	General physical health deterioration	182 (19.9)
	Renal impairment [a]	26 (1.7)	Hemoglobin decreased	173 (18.9)
	Emotional disorder [b]	20 (1.6)	Prostatic specific antigen increased	169 (18.4)
	Cardiac disorders	16 (1.3)	Cardiac disorders	125 (13.6)
	Disease progression	13 (1.1)	Malaise	113 (12.3)

Note: a. Renal impairment included renal failure (0.82%), renal impairment (0.33%), blood creatinine increased (0.33%), acute kidney injury (0.24%), renal failure acute (0.16%), renal disorder (0.16%), and blood creatinine abnormal (0.08%). **b**. Emotional disorder in patients with prostate cancer treated with Dmab included nervousness (0.08%), anxiety (0.33%), abasia (0.16%), irritability (0.16%), restlessness (0.16%), aggression (0.16%), agitation (0.08%), anger (0.08%), anxiety (0.08%), depressed mood (0.08%), depression (0.08%), and hallucination (0.08%).

341 AEs in Dmab and 583 AEs in ZA that were not mentioned in the drug instructions.

In breast cancer treatment, the top AEs associated with Dmab were death (11.6%), disease progression (3.3%), breast cancer metastatic (2.7%), neutropenia (2.7%), emotional disorder (2.3%), and pyrexia (1.7%). For ZA, the most frequent AEs were death (19.8%), emotional disorder (12.9%), osteomyelitis (11.4%), neoplasm progression (10.7%), cardiac disorders (10.6%), and impaired healing (9.5%). Cardiac disorders in patients with breast cancer treated with ZA included tachycardia (3.04%), congestive heart failure (1.35%), arrhythmia (1.01%), and palpitations (1.01%). Both drugs were associated with varying degrees of mental illness such as emotional distress, depression, and personality disorder, particularly in treating breast cancer, even leading to suicidal ideation. Tables 7, 8 display the six most frequent AEs in breast cancer and prostate cancer treatments, respectively, which were not registered in the drug specifications.

In prostate cancer treatment, the top AEs associated with Dmab were death (8.9%), prostate cancer metastatic (1.6%), and renal impairment (1.7%), while for ZA, the most common AEs were death (34.4%), general physical health deterioration (19.9%), and hemoglobin decreased (18.9%). Additionally, ZA was also associated with increased prostatic specific antigen (18.4%) and cardiac disorders (13.6%), while Dmab was associated with emotional disorder (1.6%) and cardiac disorders (1.3%).

4 Discussion

4.1 Descriptive analysis

In this study, we performed a pharmacovigilance analysis using FAERS to investigate suspected AEs and off-label uses associated with Dmab and ZA. The data covers a substantial timeframe from 2004 to 2022, during which these two medications were administered in clinical practice at different time periods. Notably, the reporting rate for AEs can differ not only among various drugs but also for the same drug as time progresses (Moore et al., 2007; Alatawi and Hansen, 2017). Additionally, media attention, regulatory measures, Risk Evaluation and Mitigation Strategy, new indications, formulation changes, or shifts in marketing approaches can impact the adverse events profiles

(Chhabra et al., 2013). Furthermore, both drug reporting trends exhibit a Weber-like effect (Hoffman et al., 2014; Noguchi et al., 2021), where AEs increase prior to marketing approval and subsequently decrease. Consequently, these variations in usage timelines may have led to different adverse event profiles, potentially impacting the results of our data analysis.

Nonetheless, the current understanding of these drugs is not yet fully comprehensive, and many AEs still require adequate attention. To better understand the AE profile of these drugs, it is recommended to collect as much clinical data as possible and conduct more in-depth analysis and evaluation.

4.2 AE signals with higher ROR values

The most frequent AEs of Dmab were death, osteonecrosis of jaw, back pain, tooth disorder, bone pain and hypocalcaemia and those for ZA were osteonecrosis, death, pain, osteonecrosis of jaw, and tooth extraction. The AEs identified in this analysis were generally in line with the known AEs of these drugs, indicating the validity of the study and suggesting that the findings may accurately reflect real-world clinical practices.

It is known that Dmab and ZA share many similar AEs. In our study, we conducted a comparative analysis to assess the signal strength of AEs between these two drugs. Among the signals of Dmab stronger than ZA ($d > 20$), the significant signals were exostosis of jaw, atypical fracture, and atypical femur fracture, suggesting that Dmab may be more prone to these AEs than ZA. Exostosis of jaw may be associated with the widely recognized osteonecrosis of the jaw (ONJ), which is a rare but serious side effect of anti-bone resorption inhibitors. Although a study demonstrated that patients with bone metastases treated with Dmab or ZA had similar incidences of ONJ (Nicolatou-Galitis et al., 2019), a meta-analysis of patients with solid tumors found that the use of Dmab was linked to a significantly higher risk of ONJ compared to ZA (Boquete-Castro et al., 2016). It is important to note that the incidence of ONJ may also be related to the dosage and duration of drug exposure (Khan et al., 2015). Thus, long-term and high-dose use of Dmab or ZA requires vigilance against ONJ. In contrast, the signals of ZA stronger than Dmab ($d > 50$) were mostly related to oral problems, which may also have potential implications for ONJ. Regular dental examinations should be conducted when using Dmab and ZA.

4.3 Off-label use with higher frequency in the database

Dmab and ZA, have been approved for preventing bone metastases associated with solid tumors. However, our research has found that these drugs are also frequently used in bone metastasis-free cancer. It should be emphasized that in some reports, cases of non-bone metastatic cancers may have been reported ambiguously without clear indication of the presence or absence of bone metastasis, thereby posing a limitation to the study. The theory of cancer treatment may primarily base on preventing cancer treatment-induced bone loss. Furthermore, some studies have shown that both drugs have potential anti-cancer properties (Dedes et al., 2012; Ubellacker et al., 2017; de Groot et al., 2018), but whether they have a positive effect on fighting cancer remains a matter of debate.

Postmenopausal women with breast cancer have a higher risk of osteoporosis due to the decrease in estrogen, as well as the effects of chemotherapy, radiotherapy, endocrine therapy, and the tumor itself (Guise, 2006; Chen et al., 2009; Gralow et al., 2013; Shapiro, 2020). Endocrine therapies such as tamoxifen and aromatase inhibitors have been shown to increase bone loss or fracture risk in both pre- and postmenopausal women with early-stage breast cancer (Powles et al., 1996; Sverrisdóttir et al., 2004; Aihara et al., 2010; Zaman et al., 2012; Tseng et al., 2018). Dmab 60 mg is approved for aromatase inhibitor-induced bone loss in women with breast cancer regardless of whether there is bone metastasis, while ZA did not receive such approval. Interestingly, ZA is reported to be used for preventing bone loss or decreasing fracture in premenopausal women with breast cancer (Gnant et al., 2015; Wilson et al., 2018). There is no clinical evidence that Dmab is suitable for use this population. Evidence suggests that Dmab or ZA could be applied as adjuvant therapy to improve bone density in postmenopausal women with early-stage breast cancer (Brufsky et al., 2009; Waqas et al., 2021). Note that one phase 3 trial shows that Dmab did not improve disease-related outcomes and did not support a role as an antitumor agent in early-stage breast cancer for women with high-risk early breast cancer, in addition to the benefits of delaying cancer bone-related events (Coleman et al., 2020).

Antihormonal treatments for prostate cancer can also cause bone loss. The FDA has approved Dmab (60 mg) for the treatment of bone loss or preventing fracture in non-metastatic prostate cancer, while ZA currently lacks FDA approval. Several small randomized trials have shown that bisphosphonates can increase BMD in patients with non-metastatic prostate cancer (Smith et al., 2001; Smith et al., 2003; Klotz et al., 2013). Note that no benefit has been shown among bisphosphonates in preventing fractures among patients with nonmetastatic prostate cancer (Strum et al., 2018).

As for the treatment of osteoarthritis (OA), a study in rabbits with experimental knee osteoarthritis showed that ZA had protective effect on articular cartilage and subchondral bone (She et al., 2017). An initial trial showed that ZA may be effective in treating osteoarthritis (Aitken et al., 2018). However, we have not yet found strong evidence that osteoarthritis can benefit from ZA. Markers of bone turnover are increased in patients with progressive OA, similar to those in patients with postmenopausal osteoporosis (Bingham et al., 2006). Based on that mechanism, ZA may have a prospective benefit for osteoarthritis. Regarding Dmab, it has rarely been reported in osteoarthritis, but evidence suggests that Dmab may be a potential new therapeutic option for treating rheumatoid arthritis (Hu et al., 2021; Tanaka et al., 2021).

In conclusion, mining new indications from the database has the potential to expand drug application range, promote drug research and development, and improve clinical practice. However, it is crucial to conduct further real-world research to validate these new indications and ultimately benefit patients.

4.4 AEs with higher report frequency in breast cancer and prostate cancer

According to the reports, disease progression was observed more frequently in the treatment of breast cancer and prostate cancer with Dmab or ZA. However, current evidence does not establish a definitive

link between tumor progression and drug exposure. Our study also found a high frequency of neutropenia among breast cancer patients treated with Dmab, which is consistent with reports of neutropenia in a phase III study of multiple myeloma patients treated with both Dmab and ZA (Raje et al., 2018). Mental problems were also reported in breast cancer patients treated with either drug, although drug-induced mental disorders on Dmab or ZA are currently poorly documented. A case report indicated that extreme anxiety and hypocalcemia after denosumab treatment for cancer-related bone metastasis may have contributed to depressive mood (Lin et al., 2015). Although atrial fibrillation is a known AE to Dmab, our study also found a high frequency of heart problems in ZA-treated patients. Previous studies have reported an increased rate of heart failure in zoledronate-treated patients (Black et al., 2007; Rubin et al., 2020), suggesting that more clinical trials are needed to confirm the safety of ZA. Renal toxicity is a potential AE of ZA treatment, although Dmab is considered relatively safe for the kidneys. However, renal toxicity has been observed in the treatment of multiple myeloma using Dmab (Raje et al., 2018). From the pharmacokinetics perspective, Dmab is not metabolized by the kidneys and theoretically has minimal damage to the kidneys, it is still relatively safe.

5 Limitation

It's important to acknowledge several limitations that raise questions about its direct real-world applicability. Looking at the FAERS database, there are several aspects to consider. First, the cases registered in spontaneous reporting systems are only those of drug-induced AEs, not the total number of patients treated with the drugs (Noguchi et al., 2021; Marwitz and Noureldin, 2022; Crisafulli et al., 2023), making it difficult to compare the incidence of AEs between Dmab and ZA. Second, some reports may lack important information such as outcome, indication, dose, age, and sex (Shao et al., 2021; Tang et al., 2022), leading to potential bias in the analysis. Additionally, the accuracy of the data may be compromised due to the involvement of non-professional reporters (Bian et al., 2021) and the absence of a standardized reporting format. Furthermore, it should be noted that some reported AEs may actually be different manifestations of the same underlying condition, such as jaw exostosis, jaw abscess, and exposed bone in jaw, all of which may be related to osteonecrosis of the jaw. Although the study has attempted to integrate such AEs, there is still a possibility of some omissions. In addition, the presence of "notoriety effects" leading to increased reporting of specific adverse events can limit the study due to potential underestimation (Pariente et al., 2007; Noguchi et al., 2021).

Regarding disproportionality analysis, it solely represents statistical correlation between drugs and AEs and do not permit the establishment of causal associations between reported AEs and specific medications (Abe et al., 2015; Michel et al., 2017). Furthermore, it comes with the limitation of false-positive signals and suffers from the limitation of lower specificity (Noguchi et al., 2021).

Also, the study's failure to compare the impact of different specifications on indications may lead to incomplete evaluation of the drugs' safety and efficacy. Additionally, it solely focuses on potential off-label use, neglecting over-the-label use of different specifications and their safety profiles, potentially overlooking certain safety issues and differences in effectiveness. Furthermore, the article may be affected by selection bias in data or inadequate analysis methods, which may impact the accuracy of drug evaluation and conclusions. The article may also not fully consider other factors affecting drug use, such as individual differences among patients, comorbidities, or the influence of other drugs.

However, the FAERS database gathers AE reports associated with drugs and therapeutic biologic products, which is a valuable resource to identify potential safety issues. Despite the aforementioned limitations, disproportionality analysis is now a validated method in the field of drug safety research and surveillance (Montastruc et al., 2011). It has high sensitivity and could serve as a foundation for generating hypotheses in future research endeavors (Abe et al., 2015; Noguchi et al., 2021; Crisafulli et al., 2023). Moreover, it can offer further insights into the influence of regulatory and policy decisions on AE reporting (Marwitz and Noureldin, 2022). Additionally, it's worth noting that there is a correlation between the risk of adverse reactions studied through meta-analysis and disproportionality analysis in many cases (Khouri et al., 2021).

6 Conclusion

In conclusion, our study found that both Dmab and ZA have similar trends in AE distribution. However, Dmab is statistically associated with a higher risk of jaw exostosis and atypical femur fractures, while ZA has a statistical link to more oral problems. It is important to note that both drugs have potential applications beyond their approved indications, particularly in the treatment of various cancers and osteoarthritis, and some new AEs may come with those off-label use, including mental health disorders, neutropenia and kidney damage, and heart problems. Given the correlation between the analysis results from spontaneous reporting systems databases and clinical safety studies (Khouri et al., 2021), our findings highlight the importance of safety monitoring when using Dmab and ZA off-label. Moreover, considering the limited research focused on this specific aspect, our study may serve as a reference point for future investigations, contributing to drug safety vigilance efforts. Finally, due to the inherent limitations of spontaneous reporting databases, which inevitably contain potential biases, there is an urgent need for well-designed comparative safety studies to validate these findings.

Data availability statement

The datasets presented in this study can be found in online repositories. The names of the repository/repositories and accession number(s) can be found below: http://openvigil.sourceforge.net/.

Author contributions

KL and XW conceived and designed the study. SS and LW conducted the database search, performed data analysis, created

figures, and contributed to writing and reviewing the manuscript. GZ, LP, and HZ participated in data interpretation. All authors contributed to the article and approved the submitted version.

Funding

This work was supported by grants from the National Natural Science Foundation of China (Grant NO. 81903581), Shenzhen Science and Technology Program (Grant NOs JCYJ20190807150005699 and RCBS20200714115000009), Shenzhen Key Medical Discipline Construction Fund (Grant NO. SZXK059), Shenzhen Key Laboratory of Prevention and Treatment of Severe Infections (ZDSYS20200811142804014).

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.

References

Abe, J., Umetsu, R., Kato, Y., Ueda, N., Nakayama, Y., Suzuki, Y., et al. (2015). Evaluation of dabigatran- and warfarin-associated hemorrhagic events using the FDA-adverse event reporting system database stratified by age. *Int. J. Med. Sci.* 12 (4), 312–321. doi:10.7150/ijms.10703

Aihara, T., Suemasu, K., Takei, H., Hozumi, Y., Takehara, M., Saito, T., et al. (2010). Effects of exemestane, anastrozole and tamoxifen on bone mineral density and bone turnover markers in postmenopausal early breast cancer patients: results of N-SAS BC 04, the TEAM Japan substudy. *Oncology* 79 (5-6), 376–381. doi:10.1159/000323489

Aitken, D., Laslett, L. L., Cai, G., Hill, C., March, L., Wluka, A. E., et al. (2018). A protocol for a multicentre, randomised, double-blind, placebo-controlled trial to compare the effect of annual infusions of zoledronic acid to placebo on knee structural change and knee pain over 24 months in knee osteoarthritis patients - ZAP2. *BMC Musculoskelet. Disord.* 19 (1), 217. doi:10.1186/s12891-018-2143-2

Alatawi, Y. M., and Hansen, R. A. (2017). Empirical estimation of under-reporting in the U.S. Food and drug administration adverse event reporting system (FAERS). *Expert Opin. drug Saf.* 16 (7), 761–767. doi:10.1080/14740338.2017.1323867

Baron, R., Ferrari, S., and Russell, R. G. (2011). Denosumab and bisphosphonates: different mechanisms of action and effects. *Bone* 48 (4), 677–692. doi:10.1016/j.bone.2010.11.020

Bian, S., Zhang, P., Li, L., Wang, Z., Cui, L., Xu, Y., et al. (2021). Anaphylaxis associated with allergen specific immunotherapy, omalizumab, and dupilumab: a real world study based on the us food and drug administration adverse event reporting system. *Front. Pharmacol.* 12, 767999. doi:10.3389/fphar.2021.767999

Bingham, C. O., 3rd, Buckland-Wright, J. C., Garnero, P., Cohen, S. B., Dougados, M., Adami, S., et al. (2006). Risedronate decreases biochemical markers of cartilage degradation but does not decrease symptoms or slow radiographic progression in patients with medial compartment osteoarthritis of the knee: results of the two-year multinational knee osteoarthritis structural arthritis study. *Arthritis rheumatism* 54 (11), 3494–3507. doi:10.1002/art.22160

Black, D. M., Delmas, P. D., Eastell, R., Reid, I. R., Boonen, S., Cauley, J. A., et al. (2007). Once-yearly zoledronic acid for treatment of postmenopausal osteoporosis. *N. Engl. J. Med.* 356 (18), 1809–1822. doi:10.1056/NEJMoa067312

Böhm, R. (2018). Primer on disproportionality analysis. Available at: https://openvigil.sourceforge.net/#.

Boquete-Castro, A., Gómez-Moreno, G., Calvo-Guirado, J. L., Aguilar-Salvatierra, A., and Delgado-Ruiz, R. A. (2016). Denosumab and osteonecrosis of the jaw. A systematic analysis of events reported in clinical trials. *Clin. oral implants Res.* 27 (3), 367–375. doi:10.1111/clr.12556

Brufsky, A. M., Bosserman, L. D., Caradonna, R. R., Haley, B. B., Jones, C. M., Moore, H. C. F., et al. (2009). Zoledronic acid effectively prevents aromatase inhibitor-associated bone loss in postmenopausal women with early breast cancer receiving adjuvant letrozole: Z-FAST study 36-month follow-up results. *Clin. breast cancer* 9 (2), 77–85. doi:10.3816/CBC.2009.n.015

Chen, Z., Maricic, M., Aragaki, A. K., Mouton, C., Arendell, L., Lopez, A. M., et al. (2009). Fracture risk increases after diagnosis of breast or other cancers in postmenopausal women: results from the Women's Health Initiative. *Osteoporos. Int. a J. established as result Coop. between Eur. Found. Osteoporos. Natl. Osteoporos. Found. U. S. A.* 20 (4), 527–536. doi:10.1007/s00198-008-0721-0

Chhabra, P., Chen, X., and Weiss, S. R. (2013). Adverse event reporting patterns of newly approved drugs in the USA in 2006: an analysis of FDA Adverse Event Reporting System data. *Drug Saf.* 36 (11), 1117–1123. doi:10.1007/s40264-013-0115-x

Coleman, R., Finkelstein, D. M., Barrios, C., Martin, M., Iwata, H., Hegg, R., et al. (2020). Adjuvant denosumab in early breast cancer (D-CARE): an international, multicentre, randomised, controlled, phase 3 trial. *Lancet Oncol.* 21 (1), 60–72. doi:10.1016/S1470-2045(19)30687-4

Crisafulli, S., Khan, Z., Karatas, Y., Tuccori, M., and Trifirò, G. (2023). An overview of methodological flaws of real-world studies investigating drug safety in the post-marketing setting. *Expert Opin. drug Saf.* 22 (5), 373–380. doi:10.1080/14740338.2023.2219892

Dedes, P. G., Gialeli, C., Tsonis, A. I., Kanakis, I., Theocharis, A. D., Kletsas, D., et al. (2012). Expression of matrix macromolecules and functional properties of breast cancer cells are modulated by the bisphosphonate zoledronic acid. *Biochimica biophysica acta* 1820 (12), 1926–1939. doi:10.1016/j.bbagen.2012.07.013

de Groot, A. F., Appelman-Dijkstra, N. M., van der Burg, S. H., and Kroep, J. R. (2018). The anti-tumor effect of RANKL inhibition in malignant solid tumors - a systematic review. *Cancer Treat. Rev.* 62, 18–28. doi:10.1016/j.ctrv.2017.10.010

Drake, M. T., Clarke, B. L., and Khosla, S. (2008). Bisphosphonates: mechanism of action and role in clinical practice. *Mayo Clin. Proc.* 83 (9), 1032–1045. doi:10.4065/83.9.1032

Gnant, M., Mlineritsch, B., Stoeger, H., Luschin-Ebengreuth, G., Knauer, M., Moik, M., et al. (2015). Zoledronic acid combined with adjuvant endocrine therapy of tamoxifen versus anastrozol plus ovarian function suppression in premenopausal early breast cancer: final analysis of the Austrian Breast and Colorectal Cancer Study Group Trial 12. *Ann. Oncol. official J. Eur. Soc. Med. Oncol.* 26 (2), 313–320. doi:10.1093/annonc/mdu544

Gralow, J. R., Biermann, J. S., Farooki, A., Fornier, M. N., Gagel, R. F., Kumar, R. N., et al. (2013). NCCN task force report: bone health in cancer care. *J. Natl. Compr. Cancer Netw. JNCCN.* 11 (3), S1–S32. doi:10.6004/jnccn.2009.0076

Greear, E. L., and Bankole, A. (2022). *Zoledronate. StatPearls. Treasure island (FL).* StatPearls Publishing Copyright © 2022, StatPearls Publishing LLC.

Guise, T. A. (2006). Bone loss and fracture risk associated with cancer therapy. *Oncol.* 11 (10), 1121–1131. doi:10.1634/theoncologist.11-10-1121

Hauben, M. (2003). A brief primer on automated signal detection. *Ann. Pharmacother.* 37 (7-8), 1117–1123. doi:10.1345/aph.1C515

Hildebrand, G. K., and Kasi, A. (2022). *Denosumab. StatPearls. Treasure island (FL).* StatPearls Publishing Copyright © 2022, StatPearls Publishing LLC.

Hoffman, K. B., Dimbil, M., Erdman, C. B., Tatonetti, N. P., and Overstreet, B. M. (2014). The weber effect and the United States food and drug administration's adverse event reporting system (FAERS): analysis of sixty-two drugs approved from 2006 to 2010. *Drug Saf.* 37 (4), 283–294. doi:10.1007/s40264-014-0150-2

Hu, Q., Zhong, X., Tian, H., and Liao, P. (2021). The efficacy of denosumab in patients with rheumatoid arthritis: a systematic review and pooled analysis of randomized or matched data. *Front. Immunol.* 12, 799575. doi:10.3389/fimmu.2021.799575

Jamal, S. A., Ljunggren, O., Stehman-Breen, C., Cummings, S. R., McClung, M. R., Goemaere, S., et al. (2011). Effects of denosumab on fracture and bone mineral density by level of kidney function. *J. bone mineral Res. official J. Am. Soc. Bone Mineral Res.* 26 (8), 1829–1835. doi:10.1002/jbmr.403

Khan, A. A., Morrison, A., Hanley, D. A., Felsenberg, D., McCauley, L. K., O'Ryan, F., et al. (2015). Diagnosis and management of osteonecrosis of the jaw: a systematic review and international consensus. *J. bone mineral Res. official J. Am. Soc. Bone Mineral Res.* 30 (1), 3–23. doi:10.1002/jbmr.2405

Khouri, C., Petit, C., Tod, M., Lepelley, M., Revol, B., Roustit, M., et al. (2021). Adverse drug reaction risks obtained from meta-analyses and pharmacovigilance disproportionality analyses are correlated in most cases. *J. Clin. Epidemiol.* 134, 14–21. doi:10.1016/j.jclinepi.2021.01.015

Klotz, L. H., McNeill, I. Y., Kebabdjian, M., Zhang, L., Chin, J. L., and Canadian Urology Research Consortium (2013). A phase 3, double-blind, randomised, parallel-group, placebo-controlled study of oral weekly alendronate for the prevention of androgen deprivation bone loss in nonmetastatic prostate cancer: the Cancer and Osteoporosis Research with Alendronate and Leuprolide (CORAL) study. *Eur. Urol.* 63 (5), 927–935. doi:10.1016/j.eururo.2012.09.007

Lin, K. F., Chen, K. H., and Huang, W. L. (2015). Organic anxiety in a woman with breast cancer receiving denosumab. *General Hosp. psychiatry* 37 (2), 192.e7–e8. doi:10.1016/j.genhosppsych.2015.01.007

Marwitz, K. K., and Noureldin, M. (2022). A descriptive analysis of concomitant opioid and benzodiazepine medication use and associated adverse drug events in United States adults between 2009 and 2018. *Explor. Res. Clin. Soc. Pharm.* 5, 100130. doi:10.1016/j.rcsop.2022.100130

Michel, C., Scosyrev, E., Petrin, M., and Schmouder, R. (2017). Can disproportionality analysis of post-marketing case reports be used for comparison of drug safety profiles? *Clin. drug Investig.* 37 (5), 415–422. doi:10.1007/s40261-017-0503-6

Montastruc, J. L., Sommet, A., Bagheri, H., and Lapeyre-Mestre, M. (2011). Benefits and strengths of the disproportionality analysis for identification of adverse drug reactions in a pharmacovigilance database. *Br. J. Clin. Pharmacol.* 72 (6), 905–908. doi:10.1111/j.1365-2125.2011.04037.x

Moore, T. J., Cohen, M. R., and Furberg, C. D. (2007). Serious adverse drug events reported to the food and drug administration, 1998-2005. *Archives Intern. Med.* 167 (16), 1752–1759. doi:10.1001/archinte.167.16.1752

Nicolatou-Galitis, O., Schiødt, M., Mendes, R. A., Ripamonti, C., Hope, S., Drudge-Coates, L., et al. (2019). Medication-related osteonecrosis of the jaw: definition and best practice for prevention, diagnosis, and treatment. *Oral Surg. oral Med. oral pathology oral radiology* 127 (2), 117–135. doi:10.1016/j.oooo.2018.09.008

Noguchi, Y., Tachi, T., and Teramachi, H. (2021). Detection algorithms and attentive points of safety signal using spontaneous reporting systems as a clinical data source. *Briefings Bioinforma.* 22 (6), bbab347. doi:10.1093/bib/bbab347

Omar, N. E., Fahmy Soliman, A. I., Eshra, M., Saeed, T., Hamad, A., and Abou-Ali, A. (2021). Postmarketing safety of anaplastic lymphoma kinase (ALK) inhibitors: an analysis of the FDA Adverse Event Reporting System (FAERS). *ESMO open* 6 (6), 100315. doi:10.1016/j.esmoop.2021.100315

Pang, K. L., Low, N. Y., and Chin, K. Y. (2020). A review on the role of denosumab in fracture prevention. *Drug Des. Dev. Ther.* 14, 4029–4051. doi:10.2147/DDDT.S270829

Pariente, A., Gregoire, F., Fourrier-Reglat, A., Haramburu, F., and Moore, N. (2007). Impact of safety alerts on measures of disproportionality in spontaneous reporting databases: the notoriety bias. *Drug Saf.* 30 (10), 891–898. doi:10.2165/00002018-200730100-00007

Powles, T. J., Hickish, T., Kanis, J. A., Tidy, A., and Ashley, S. (1996). Effect of tamoxifen on bone mineral density measured by dual-energy x-ray absorptiometry in healthy premenopausal and postmenopausal women. *J. Clin. Oncol. official J. Am. Soc. Clin. Oncol.* 14 (1), 78–84. doi:10.1200/JCO.1996.14.1.78

Raje, N., Terpos, E., Willenbacher, W., Shimizu, K., García-Sanz, R., Durie, B., et al. (2018). Denosumab versus zoledronic acid in bone disease treatment of newly diagnosed multiple myeloma: an international, double-blind, double-dummy, randomised, controlled, phase 3 study. *Lancet Oncol.* 19 (3), 370–381. doi:10.1016/S1470-2045(18)30072-X

Rubin, K. H., Möller, S., Choudhury, A., Zorina, O., Kalsekar, S., Eriksen, E. F., et al. (2020). Cardiovascular and skeletal safety of zoledronic acid in osteoporosis observational, matched cohort study using Danish and Swedish health registries. *Bone* 134, 115296. doi:10.1016/j.bone.2020.115296

Shao, Q. H., Yin, X. D., Liu, H. X., Zhao, B., Huang, J. Q., and Li, Z. L. (2021). Kidney injury following ibuprofen and acetaminophen: A real-world analysis of post-marketing surveillance data. *Front. Pharmacol.* 12, 750108. doi:10.3389/fphar.2021.750108

Shapiro, C. L. (2020). Osteoporosis: a long-term and late-effect of breast cancer treatments. *Cancers* 12 (11), 3094. doi:10.3390/cancers12113094

She, G., Zhou, Z., Zha, Z., Wang, F., and Pan, X. (2017). Protective effect of zoledronic acid on articular cartilage and subchondral bone of rabbits with experimental knee osteoarthritis. *Exp. Ther. Med.* 14 (5), 4901–4909. doi:10.3892/etm.2017.5135

Smith, M. R., Eastham, J., Gleason, D. M., Shasha, D., Tchekmedyian, S., and Zinner, N. (2003). Randomized controlled trial of zoledronic acid to prevent bone loss in men receiving androgen deprivation therapy for nonmetastatic prostate cancer. *J. urology* 169 (6), 2008–2012. doi:10.1097/01.ju.0000063820.94994.95

Smith, M. R., McGovern, F. J., Zietman, A. L., Fallon, M. A., Hayden, D. L., Schoenfeld, D. A., et al. (2001). Pamidronate to prevent bone loss during androgen-deprivation therapy for prostate cancer. *N. Engl. J. Med.* 345 (13), 948–955. doi:10.1056/NEJMoa010845

Strum, S. B., Zukotynski, K., and Walker-Dilks, C. (2018). Bone health and bone-targeted therapies for nonmetastatic prostate cancer. *Ann. Intern. Med.* 168 (6), 459–460. doi:10.7326/L17-0702

Sverrisdóttir, A., Fornander, T., Jacobsson, H., von Schoultz, E., and Rutqvist, L. E. (2004). Bone mineral density among premenopausal women with early breast cancer in a randomized trial of adjuvant endocrine therapy. *J. Clin. Oncol. official J. Am. Soc. Clin. Oncol.* 22 (18), 3694–3699. doi:10.1200/JCO.2004.08.148

Tanaka, Y., Takeuchi, T., Soen, S., Yamanaka, H., Yoneda, T., Tanaka, S., et al. (2021). Effects of denosumab in Japanese patients with rheumatoid arthritis treated with conventional antirheumatic drugs: 36-month extension of a phase III study. *J. rheumatology* 48 (11), 1663–1671. doi:10.3899/jrheum.201376

Tang, S., Wu, Z., Xu, L., Wen, Q., and Zhang, X. (2022). Adverse reaction signals mining and hemorrhagic signals comparison of ticagrelor and clopidogrel: a pharmacovigilance study based on FAERS. *Front. Pharmacol.* 13, 970066. doi:10.3389/fphar.2022.970066

Tseng, O. L., Spinelli, J. J., Gotay, C. C., Ho, W. Y., McBride, M. L., and Dawes, M. G. (2018). Aromatase inhibitors are associated with a higher fracture risk than tamoxifen: a systematic review and meta-analysis. *Ther. Adv. Musculoskelet. Dis.* 10 (4), 71–90. doi:10.1177/1759720X18759291

Ubellacker, J. M., Haider, M. T., DeCristo, M. J., Allocca, G., Brown, N. J., Silver, D. P., et al. (2017). Zoledronic acid alters hematopoiesis and generates breast tumor-suppressive bone marrow cells. *Breast cancer Res. BCR* 19 (1), 23. doi:10.1186/s13058-017-0815-8

van Puijenbroek, E. P., Bate, A., Leufkens, H. G., Lindquist, M., Orre, R., and Egberts, A. C. G. (2002). A comparison of measures of disproportionality for signal detection in spontaneous reporting systems for adverse drug reactions. *Pharmacoepidemiol. drug Saf.* 11 (1), 3–10. doi:10.1002/pds.668

Verden, A., Dimbil, M., Kyle, R., Overstreet, B., and Hoffman, K. B. (2018). Analysis of spontaneous postmarket case reports submitted to the FDA regarding thromboembolic adverse events and JAK inhibitors. *Drug Saf.* 41 (4), 357–361. doi:10.1007/s40264-017-0622-2

Waqas, K., Lima Ferreira, J., Tsourdi, E., Body, J. J., Hadji, P., and Zillikens, M. C. (2021). Updated guidance on the management of cancer treatment-induced bone loss (CTIBL) in pre- and postmenopausal women with early-stage breast cancer. *J. bone Oncol.* 28, 100355. doi:10.1016/j.jbo.2021.100355

Wilson, C., Bell, R., and Hinsley, S. (2018). *Adjuvant zoledronic acid reduces fractures in breast cancer patients; an AZURE (BIG 01/04) study*, 94. Oxford, England: European journal of cancer, 70–78.

Zaman, K., Thürlimann, B., Huober, J., Schönenberger, A., Pagani, O., Lüthi, J., et al. (2012). Bone mineral density in breast cancer patients treated with adjuvant letrozole, tamoxifen, or sequences of letrozole and tamoxifen in the BIG 1-98 study (SAKK 21/07). *Ann. Oncol. official J. Eur. Soc. Med. Oncol.* 23 (6), 1474–1481. doi:10.1093/annonc/mdr448

frontiers | Frontiers in Pharmacology

TYPE Systematic Review
PUBLISHED 16 November 2023
DOI 10.3389/fphar.2023.1291164

Check for updates

OPEN ACCESS

EDITED BY
Robert L. Lins,
Independent Researcher, Antwerp, Belgium

REVIEWED BY
Louis Dron,
Cytel, Canada
Babak Mohit,
Merck Sharp & Dohme Corp, United States

*CORRESPONDENCE
Jamie Elvidge,
jamie.elvidge@nice.org.uk

RECEIVED 08 September 2023
ACCEPTED 30 October 2023
PUBLISHED 16 November 2023

CITATION
Elvidge J, Hopkin G, Narayanan N, Nicholls D and Dawoud D (2023), Diagnostics and treatments of COVID-19: two-year update to a living systematic review of economic evaluations. *Front. Pharmacol.* 14:1291164. doi: 10.3389/fphar.2023.1291164

COPYRIGHT
© 2023 Elvidge, Hopkin, Narayanan, Nicholls and Dawoud. This is an open-access article distributed under the terms of the Creative Commons Attribution License (CC BY). The use, distribution or reproduction in other forums is permitted, provided the original author(s) and the copyright owner(s) are credited and that the original publication in this journal is cited, in accordance with accepted academic practice. No use, distribution or reproduction is permitted which does not comply with these terms.

Diagnostics and treatments of COVID-19: two-year update to a living systematic review of economic evaluations

Jamie Elvidge[1]*, Gareth Hopkin[1], Nithin Narayanan[2], David Nicholls[1] and Dalia Dawoud[1]

[1]Science, Evidence and Analytics Directorate, National Institute for Health and Care Excellence, Manchester, United Kingdom, [2]Norwich Medical School, University of East Anglia, Norwich, United Kingdom

Objectives: As the initial crisis of the COVID-19 pandemic recedes, healthcare decision makers are likely to want to make rational evidence-guided choices between the many interventions now available. We sought to update a systematic review to provide an up-to-date summary of the cost-effectiveness evidence regarding tests for SARS-CoV-2 and treatments for COVID-19.

Methods: Key databases, including MEDLINE, EconLit and Embase, were searched on 3 July 2023, 2 years on from the first iteration of this review in July 2021. We also examined health technology assessment (HTA) reports and the citations of included studies and reviews. Peer-reviewed studies reporting full health economic evaluations of tests or treatments in English were included. Studies were quality assessed using an established checklist, and those with very serious limitations were excluded. Data from included studies were extracted into predefined tables.

Results: The database search identified 8,287 unique records, of which 54 full texts were reviewed, 28 proceeded for quality assessment, and 15 were included. Three further studies were included through HTA sources and citation checking. Of the 18 studies ultimately included, 17 evaluated treatments including corticosteroids, antivirals and immunotherapies. In most studies, the comparator was standard care. Two studies in lower-income settings evaluated the cost effectiveness of rapid antigen tests and critical care provision. There were 17 modelling analyses and 1 trial-based evaluation.

Conclusion: A large number of economic evaluations of interventions for COVID-19 have been published since July 2021. Their findings can help decision makers to prioritise between competing interventions, such as the repurposed antivirals and immunotherapies now available to treat COVID-19. However, some evidence gaps remain present, including head-to-head analyses, disease-specific utility values, and consideration of different disease variants.

Systematic Review Registration: [https://www.crd.york.ac.uk/prospero/display_record.php?ID=CRD42021272219], identifier [PROSPERO 2021 CRD42021272219].

KEYWORDS

cost-effectiveness, COVID-19, diagnostics, economic evaluation, health technology assessment, pharmacological, living review, cost-utility analysis

1 Introduction

The novel severe acute respiratory syndrome coronavirus 2 (SARS-CoV-2) and its associated disease (COVID-19) pandemic placed healthcare systems and wider economies under massive strain in 2020 and 2021. Decisions about diagnostic tests and treatments for the disease were made rapidly, forgoing traditional, rigorous health technology assessments (HTAs) that healthcare interventions are subjected to in many countries. Now that the early pandemic crisis has passed, HTA organisations will increasingly view COVID-19 as being equivalent to any other condition, and seek to understand the *cost* effectiveness of tests and treatments for it. Such evidence can support reimbursement decisions and the efficient allocation of scarce healthcare resources.

In July 2021, the first iteration of a systematic literature review to identify economic evaluations of tests and treatments for COVID-19 was conducted (Elvidge et al., 2022). Its objective was to identify up-to-date cost-effectiveness estimates for COVID-19 tests and treatments, and the methodological approaches, limitations and uncertainties present in published economic evaluations. Since then, the pandemic context, evidence base, and disease have evolved considerably. The present study reports a timely two-year update of the review, to provide a contemporary understanding of the cost-effectiveness evidence for COVID-19 tests and treatments.

This study has been supported by Next-Generation Health Technology Assessment (HTx), which is a Horizon 2020 project supported by the European Union, lasting for 5 years from January 2019. Its main aim is to create a framework for the next-generation of HTA to support patient-centred, societally oriented, real-time decision making on access to and reimbursement for health technologies throughout Europe.

2 Materials and methods

We performed an update of a previously published systematic literature review to identify full economic evaluations of diagnostics (e.g., tests) for SARS-CoV-2 and treatments (e.g., pharmaceuticals) for COVID-19 (Elvidge et al., 2021; Elvidge et al., 2022). The date range spanned the previous search date, 12th July 2021, to 3rd July 2023. The search strategy was consistent with the original search, including citation checking of included studies and efforts to identify relevant grey literature. Studies were included if they were full economic evaluations, comparing both the costs and health outcomes of 2 or more alternative tests for SARS-CoV-2 or treatments for COVID-19.

Every identified title and abstract was screened against the selection criteria by 2 reviewers (JE and NN/GH). For studies that were identified as potentially relevant, full-text articles were sought and assessed against the selection criteria by both reviewers. Studies that met the selection criteria were quality assessed by both reviewers, using the NICE economic evaluation quality checklist (National Institute for Health and Care Excellence NICE, 2012). Those judged to have very serious limitations were excluded. For each included study, data extraction was conducted by 1 reviewer using prespecified tables consistent with the original review. Extracted data for each study were checked and validated by another reviewer. At all stages, discrepancies between the reviewers were resolved through discussion or, if needed, adjudication by a senior reviewer (DD). Key study characteristics are presented in Tables 1, 2, and findings in Table 3. Due to extensive heterogeneity between studies, results were synthesised narratively (Shields and Elvidge, 2020).

3 Results

3.1 Included studies

Search strategies and results per database are provided in Supplementary Material. A total of 8,287 unique records were identified for initial screening of titles and abstracts (Figure 1). Of those, 8,233 were excluded, most commonly because they did not report a primary economic evaluation. Therefore, 54 studies proceeded to full-text review, with 28 meeting the inclusion criteria. Six studies were also identified through searches of grey literature: 1 through citation checking, which met our inclusion criteria, and 5 HTA reports, of which 2 met our criteria. Two reported on the same HTA and were considered to be duplicates, and 1 was not available in English. A total of 31 studies proceeded to quality assessment, of which 13 were excluded due to the presence of very serious limitations (Table 4). Finally, 18 studies of acceptable quality were included in this two-year update (Carta and Conversano, 2021; Congly et al., 2021; Ruggeri et al., 2022a; Ruggeri et al., 2022b; Dijk et al., 2022; Goswami et al., 2022; Kelton et al., 2022; Lau et al., 2022; Metry et al., 2022; Park et al., 2022; Rafia et al., 2022; Savinkina et al., 2022; Yeung et al., 2022; Alamer et al., 2023; Arwah et al., 2023; Kowal et al., 2023; Ruggeri et al., 2023; Shah et al., 2023).

Included studies evaluated interventions in community or outpatient settings (5/18) (Goswami et al., 2022; Park et al., 2022; Savinkina et al., 2022; Yeung et al., 2022; Ruggeri et al., 2023), where patients are at risk of admission to hospital, or an inpatient hospital setting (11/18) (Carta and Conversano, 2021; Congly et al., 2021; Ruggeri et al., 2022a; Ruggeri et al., 2022b; Dijk et al., 2022; Kelton et al., 2022; Lau et al., 2022; Rafia et al., 2022; Alamer et al., 2023; Kowal et al., 2023; Shah et al., 2023); one study included both settings (1/18) (Metry et al., 2022). One study evaluated point-of-care tests in an unspecified health facility (1/18) (Arwah et al., 2023). Of studies based in inpatient hospital settings, some were aimed at specific populations and places within the care pathway, namely moderate disease with non-invasive ventilation (3/12) (Ruggeri et al., 2022a; Ruggeri et al., 2022b; Rafia et al., 2022) or critical care (1/12) (Shah et al., 2023), though most had mixed or unspecified populations (8/12) (Carta and Conversano, 2021; Congly et al., 2021; Dijk et al., 2022; Kelton et al., 2022; Lau et al., 2022; Metry et al., 2022; Alamer et al., 2023; Kowal et al., 2023). Most studies (12/18) (Carta and Conversano, 2021; Congly et al., 2021; Dijk et al., 2022; Goswami et al., 2022; Lau et al., 2022; Metry et al., 2022; Rafia et al., 2022; Savinkina et al., 2022; Yeung et al., 2022; Alamer et al., 2023; Kowal et al., 2023; Ruggeri et al., 2023) took a healthcare system or payer perspective in their base-case analyses, while 2/18 took a provider (e.g., hospital) perspective (Ruggeri et al., 2022a; Shah et al., 2023), 2/18 took a partial societal perspective (Kelton et al., 2022; Arwah et al., 2023), and 2/18 did not explicitly report a perspective (Ruggeri et al., 2022b; Park et al.,

TABLE 1 General characteristics of included studies.

Study	Country (currency)	Population/Setting	Intervention(s) & comparator(s)	Type of evaluation	Quality assessment[b]
Alamer 2023 (Alamer et al., 2023)	Saudi Arabia (SAR)	415 patients with moderate to severe COVID-19 disease who were admitted to two Saudi COVID-19 referral hospitals	Favipiravir, standard of care (SoC)	CEA	Potentially serious limitations
Arwah 2023 (Arwah et al., 2023)	Kenya (USD)	Patients with suspected COVID-19 presenting at settings with access to point-of-care testing	Two comparisons	CUA	Potentially serious limitations
			Rapid tests with delayed confirmatory testing for negative, delayed testing		
			Rapid tests, clinical judgement		
Carta and Conversano 2021 (Carta and Conversano, 2021)	United States (USD)	Hospitalised COVID-19 patients (4 levels of respiratory support), aged 60	Remdesivir, dexamethasone, remdesivir + dexamethasone (R + D), SoC	CUA	Potentially serious limitations
Congly 2021 (Congly et al., 2021)	United States (USD)	Hospitalised patients, moderate (oxygen) & severe (ICU), aged 60	Combinations of SoC, redemsivir and dexamethasone, by severity	CUA	Potentially serious limitations
Dijk 2022 (Dijk et al., 2022)	United States (USD)	Hospitalised COVID-19 patients	Hydroxychloroquine, remdesivir, casirivimab + imdevimab (C + I), dexamethasone, baricitinib + remdesivir (B + R), tocilizumab, lopinavir + ritonavir (L + R), interferon b1a, SoC	CUA	Minor limitations
Goswami 2022 (Goswami et al., 2022)	United States (USD)	Outpatient adults with mild to moderate COVID-19 and 1 or more risk factor for severe disease	Molnupiravir, SoC	CUA	Minor limitations
Kelton 2021 (Kelton et al., 2022)	United States (USD)	Hospitalised COVID-19 patients	B + R, remdesivir	CUA	Potentially serious limitations
Kowal 2023 (Kowal et al., 2023)	United States (USD)	Hospitalised COVID-19 patients, stratified into equity-relevant subgroups by race/ethnicity and deprivation	Hypothetical treatment, SoC (per clinical trials in 2020)	DCUA	Potentially serious limitations
Lau 2022 (Lau et al., 2022)	Canada (CAD)	Adult, hospitalized patients with COVID-19	Remdesivir, SoC	CEA	Minor limitations
Metry 2022 (Metry et al., 2022)	United Kingdom (GBP)	In hospital or in community and at high risk of hospitalisation	Hospital setting	CUA	Minor limitations
			Baricitinib, B + R, C + I[a], lenzilumab[a], remdesivir, tocilizumab, SoC		
			Community setting		
			C + I[a], molnupiravir[a], nirmatrelvir + ritonavir (N + R), remdesivir, sotrovimab, SoC		
Park 2022 (Park et al., 2022)	Singapore (USD)	4 relevant scenarios of unvaccinated patients by age group	C + I, SoC	CEA	Potentially serious limitations
				CUA	
Rafia 2022 (Rafia et al., 2022)	United Kingdom (GBP)	Hospitalised COVID-19 patients requiring oxygen or non-invasive ventilation (NIV)	Remdesvir, SoC	CUA	Minor limtations
Ruggeri 2022 (Ruggeri et al., 2022a)	Portugal (EUR)	Hospitalised COVID-19 patients on low-flow oxygen	Remdesivir, SoC	CEA	Potentially serious limitations
Ruggeri 2022 (Ruggeri et al., 2022b)	Saudi Arabia (USD)	Hospitalised COVID-19 patients on low-flow oxygen	Remdesivir, SoC	CEA	Potentially serious limitations
Ruggeri 2023 (Ruggeri et al., 2023)	Italy (EUR)	Outpatients with COVID-19 not having low-flow oxygen	C + I, SoC	CEA	Potentially serious limitations
Savinkina 2022 (Savinkina et al., 2022)	United States (USD)	Newly diagnosed COVID-19 positive patients, including subgroups by high & low risk of severe disease and vaccination status (vaccine assumed to be 75% effective at reducing hospital risk)	N + R; SoC (no N + R); and 3 interim stratgies with different levels of N + R	CEA	Potentially serious limitations

(Continued on following page)

TABLE 1 (*Continued*) General characteristics of included studies.

Study	Country (currency)	Population/Setting	Intervention(s) & comparator(s)	Type of evaluation	Quality assessment[b]
Shah 2023 (Shah et al., 2023)	Tanzania (USD)	Hospitalised critically ill adult patients with COVID-19	Advanced critical care, essential critical care, district-level critical care, no critical care	CUA	Potentially serious limitations
Yeung 2022 (Yeung et al., 2022)	United States (USD)	Mild to moderate outpatients at high risk of progression to severe disease	Molnupiravir, N + R, fluvoxamine, SoC (pooled from key trials)	CUA	Minor limitations

[a]In Metry et al., the cost-effectiveness results for lenzilumab, molnupiravir and C + I were considered to be confidential and were therefore not made publicly available.
[b]Minor limitations indicates the study meets all quality assessment criteria, or fails 1 or more criteria but this is unlikely to change the conclusions about cost effectiveness. Potentially serious limitations indicates the study fails 1 or more quality assessment criteria and this has the potential to change the conclusions about cost effectiveness.
Abbreviations: B + R, baricitinib + remdesivir; CEA, cost-effectiveness analysis; CUA, cost—utility analysis; C + I, casirivimab + imdevimab; DCUA, distributional cost—utility analysis; ICU, intensive care unit; L + R, lopinavir + ritonavir; N + R, nirmatrelvir + ritonavir; NIV, non-invasive ventilation; SoC, standard of care.

2022). Multiple studies were conducted in the United States (8/18) (Carta and Conversano, 2021; Congly et al., 2021; Dijk et al., 2022; Goswami et al., 2022; Kelton et al., 2022; Savinkina et al., 2022; Yeung et al., 2022; Kowal et al., 2023), Saudi Arabia (2/18) (Ruggeri et al., 2022b; Alamer et al., 2023) and the United Kingdom (2/18) (Metry et al., 2022; Rafia et al., 2022), while single studies were conducted in each of Canada (Lau et al., 2022), Italy (Ruggeri et al., 2023), Kenya (Arwah et al., 2023), Portugal (Ruggeri et al., 2022a), Singapore (Park et al., 2022) and Tanzania (Shah et al., 2023). Most studies reported costs in US dollars (12/18), with 4/18 converting to US dollars from the local currency (Ruggeri et al., 2022b; Park et al., 2022; Arwah et al., 2023; Shah et al., 2023), while 6/18 reported costs in the local non-US currency (Ruggeri et al., 2022a; Lau et al., 2022; Metry et al., 2022; Rafia et al., 2022; Alamer et al., 2023; Ruggeri et al., 2023).

Included studies evaluated one or more of the following pharmacological treatments for COVID-19, usually in addition to standard care: remdesivir (9/18) (Carta and Conversano, 2021; Congly et al., 2021; Ruggeri et al., 2022a; Ruggeri et al., 2022b; Dijk et al., 2022; Kelton et al., 2022; Lau et al., 2022; Metry et al., 2022; Rafia et al., 2022), casirivimab + imdevimab (3/18) (Dijk et al., 2022; Park et al., 2022; Ruggeri et al., 2023), baricitinib + remdesivir (3/18) (Dijk et al., 2022; Kelton et al., 2022; Metry et al., 2022), dexamethasone (3/18) (Carta and Conversano, 2021; Congly et al., 2021; Dijk et al., 2022), nirmatrelivir + ritonivir (3/18) (Metry et al., 2022; Savinkina et al., 2022; Yeung et al., 2022), molnupiravir (2/18) (Goswami et al., 2022; Yeung et al., 2022), and tocilizumab (2/18) (Dijk et al., 2022; Metry et al., 2022). The following medicines were evaluated by single studies: baricitinib (Metry et al., 2022), favipiravir (Alamer et al., 2023), fluvoxamine (Yeung et al., 2022), hydroxychloroquine (Dijk et al., 2022), interferon beta-1a (Dijk et al., 2022), lopinavir + ritonivir (Dijk et al., 2022), remdesivir + dexamethasone (Carta and Conversano, 2021), sotrovimab (Metry et al., 2022). One study (1/18) evaluated lenzilumab alongside other treatments (lenzilumab, molnupiravir and casirivimab + imdevimab) but did not publish cost-effectiveness results for these other treatments due to confidentiality (Metry et al., 2022). One study (1/18) evaluated a hypothetical pharmacological treatment for COVID-19, with an efficacy profile derived from the ACTT-1 (remdesivir) and RECOVERY (dexamethasone) trials, and a price of $2,500 per course (Kowal et al., 2023). In all cases, standard care without the pharmacological intervention of interest was a comparator. One study (1/18) evaluated a test for SARS-CoV-2 (Arwah et al., 2023), and one (1/18) evaluated the cost effectiveness of different levels of critical care for the treatment of severe COVID-19 in a lower-middle income country setting (Shah et al., 2023). The comparators were not treating COVID-19 in critical care services; treating COVID-19 with basic critical care in district hospitals, reflecting standard care; essential critical care, defined as treating people with severe and critical disease with advanced care such as supplemental oxygen; and advanced critical care, where people with critical disease are treated with life-sustaining therapies such as mechanical ventilation.

Cost—utility analyses (CUAs) were reported by 12/18 studies, quantifying costs and a preference-based measure of health (Carta and Conversano, 2021; Congly et al., 2021; Dijk et al., 2022; Goswami et al., 2022; Kelton et al., 2022; Metry et al., 2022; Park et al., 2022; Yeung et al., 2022; Arwah et al., 2023; Kowal et al., 2023; Shah et al., 2023). In most cases (9/12), quality-adjusted life years (QALYs) were used (Carta and Conversano, 2021; Congly et al., 2021; Dijk et al., 2022; Goswami et al., 2022; Kelton et al., 2022; Metry et al., 2022; Rafia et al., 2022; Yeung et al., 2022; Kowal et al., 2023); the rest (3/12) used disability-adjusted life years (DALYs) (Park et al., 2022; Arwah et al., 2023; Shah et al., 2023). One of the QALY-based analyses was a *distributional* CUA (Kowal et al., 2023), and was a re-analysis of a study that was included in the initial review (Sheinson et al., 2021). Cost-effectiveness analyses (CEAs) were reported by 7/18 studies (Ruggeri et al., 2022a; Ruggeri et al., 2022b; Lau et al., 2022; Park et al., 2022; Savinkina et al., 2022; Alamer et al., 2023; Ruggeri et al., 2023). All CEA studies used deaths averted as their non-preference-based measure of health. One study (1/18) was a CEA conducted alongside a clinical trial (Lau et al., 2022). All other studies (17/18) reported model-based analyses, comprising decision trees (6/17) (Carta and Conversano, 2021; Congly et al., 2021; Metry et al., 2022; Park et al., 2022; Savinkina et al., 2022; Arwah et al., 2023); Markov models (6/17) (Ruggeri et al., 2022a; Ruggeri et al., 2022b; Dijk et al., 2022; Kowal et al., 2023; Ruggeri et al., 2023; Shah et al., 2023), some of which were nested within a disease epidemiology model (3/17) (Ruggeri et al., 2022a; Ruggeri et al., 2022b; Ruggeri et al., 2023); "hybrid" models, with a decision tree to model acute disease followed by a

TABLE 2 Economic evaluation characteristics of included studies.

Study	Analysis approach	Perspective	Time horizon	Cost categories	Cost year	Discounting	Health outcomes	Efficacy data source	Utility data source
Alamer 2023 (Alamer et al., 2023)	Patient-level simulation	Healthcare payer	5 months	Favipiravir, inpatient care, isolation room, personnel, laboratory, tests	2020	NR	Deaths averted	Retrospective comparative study using propensity score matching	NA
Arwah 2023 (Arwah et al., 2023)	Decision tree	Societal	Patient care episode	Testing, treatment, related healthcare services, isolation, travel, value of time, informal care, productivity loss	2021	3%	DALYs averted	Author assumptions and observational evidence	Non-COVID sources
Carta and Conversano 2021 (Carta and Conversano, 2021)	Decision tree	Healthcare	1 year	Treatments, inpatient care (by LoS), follow-up care	NR (appears to be 2020)	NA	QALY	RCTs. Remdesivir and dexamethasone effects assumed to be additive	Non-COVID sources
Congly 2021 (Congly et al., 2021)	Decision tree	Healthcare	1 year	Treatments, inpatient care (by DRG)	2020	NA	QALY	Meta analysis & RCT	Non-COVID sources
Dijk 2022 (Dijk et al., 2022)	Markov model	Healthcare	Lifetime	Treatments, inpatient care, rehabilitation	2020	3%	QALY	RCTs	Non-COVID sources
Goswami 2022 (Goswami et al., 2022)	Hybrid model: decision tree followed by Markov	Healthcare	Lifetime	Molnupiravir, inpatient care, outpatient care, emergency care	2021	3%	QALY	RCT	Primary vignettes study, EQ-5D-5L (United Kingdom n = 500) using US value set.
Kelton 2021 (Kelton et al., 2022)	Hybrid model: decision tree followed by Markov	Base case: partial societal (hospital plus indirect productivity costs). Scenario: hospital only	Base case: lifetime. Hospital scenario: hospitalisation duration	Treatments, inpatient care (base case: by LoS; hospital scenario: less DRG payments); post-discharge care; long-term all-cause care; lost work days	NR	3%	QALY	RCTs. "Data on file" cited for the ACTT-2 trial	Non-COVID sources
Kowal 2023 (Kowal et al., 2023)	Distributional reanalysis of Sheinson 2021 model	Healthcare (payer)	Same as Sheinson 2021						
Lau 2022 (Lau et al., 2022)	Trial-based	Healthcare public payer	To discharge or death	Remdesivir, ICU & ward stays, personnel, laboratory and radiology, procedures, surgeries	2020	None	Deaths averted	RCT	NA
Metry 2022 (Metry et al., 2022)	Hospital: partitioned survival model Community: decision tree model	Healthcare	Lifetime	Treatments, hospital care, outpatient monitoring, long COVID	NR	NR	QALY	Living NMAs (metaEvidence Initiative, 2022; The COVID-NMA Initiative, 2021)	Non-COVID sources

(Continued on following page)

10.3389/fphar.2023.1291164

TABLE 2 (*Continued*) Economic evaluation characteristics of included studies.

Study	Analysis approach	Perspective	Time horizon	Cost categories	Cost year	Discounting	Health outcomes	Efficacy data source	Utility data source
Park 2022 (Park et al., 2022)	Decision tree	NR	Duration of illness	C + I, hospital care	NR	NA	Deaths averted, DALYs averted	RCTs	Burden on illness study in Malta, derived from non-COVID disability weights
Rafia 2022 (Rafia et al., 2022)	Partitioned survival model	Healthcare	Lifetime	Treatments, hospital care	NR	3.5%	QALY	RCT	Non-COVID sources
Ruggeri 2022 (Ruggeri et al., 2022a)	Markov model nested within epidemiological model	Hospital	20 weeks	Remdesivir, inpatient costs including ICU	NR	NA	Deaths averted	RCT	NA
Ruggeri 2022 (Ruggeri et al., 2022b)	Markov model nested within epidemiological model with 3 scenarios: Static, decreasing and increasing infection rates	NR	20 weeks	Remdesivir, inpatient costs including ICU and IV	2020	NA	Deaths averted	RCT	NA
Ruggeri 2023 (Ruggeri et al., 2023)	Markov model nested within epidemiological model	Healthcare (payer)	20 weeks	C + I, inpatient costs including ICU	NR	NA	Deaths averted	RCT	NA
Savinkina 2022 (Savinkina et al., 2022)	Decision tree model	Healthcare	30 days	N + R, admission cost	NR	NA	Deaths averted	High risk, not vaccinated: RCT and observational data. Not high risk or high risk and vaccinated: RCT.	NA
Shah 2023 (Shah et al., 2023)	Markov model	Provider perspective	28 days	Cost of critical care (limited details)	2020	None	DALYs averted	Expert elicitation	Non-COVID sources
Yeung 2022 (Yeung et al., 2022)	Hybrid model: decision tree followed by Markov	Base case: healthcare Scenario:modified societal	Lifetime	Treatments, related care, age-adjusted other healthcare, productivity costs (scenario)	2021	3%	QALY	RCTs. Manufacturer press release cited for N + R trial	Non-COVID sources

Abbreviations: DALY, disability-adjusted life-year; DRG, diagnostic-related group; EQ-5D-5L, Euroqol 5 dimension 5 level; ICU, intensive care unit; LoS, length of stay; MV, mechanical (invasive) ventilation; NA, not applicable; NMA, network meta-analysis; NR, not reported; QALY, quality-adjusted life-year; RCT, randomised controlled trial.

10.3389/fphar.2023.1291164

TABLE 3 Results of included studies.

Study	Cost and health outcome results	ICER/net benefit of intervention(s) vs. comparator(s)	Cost-effectiveness threshold (if relevant)	Sensitivity & scenario analyses	Authors' conclusions regarding cost effectiveness	Authors' reported limitations and challenges
Alamer 2023 (Alamer et al., 2023)	Favipiravir: $17,197*; 0.97 probability survival	-$4,534* per death averted	No threshold reported	Analysis replicated with weighted model using propensity scores and PSA completed for weight and unweighted analyses	Favipiravir was associated with lower cost than SoC in both the unweighted and the weighted models. Favipiravir was also associated with a higher probability to be discharged alive	Limiting to deaths averted and time to discharge may miss important outcomes. Extending to CUA was not possible but would be desirable
	SoC: $35,331*; 0.93 probability survival		Report no agreed thresholds in Saudi but dominant	Favipiravir is less costly and more effective across all analyses		Study did not explore timing of treatment which may be important
Arwah 2023 (Arwah et al., 2023)	Scenario 1 (access to confirmatory testing)	Scenario 1	$1003 (stated as Kenyan threshold)	Deterministic sensitivity with key parameters varied and PSA	Using rapid testing as a first-line tool, and later confirmatory tests of negatives where available, was a cost-effective strategy. Otherwise, rapid testing is preferred to clinical judgement, although it is less costly and less effective	Limited by unavailable data on outcomes for false negative patients
	Rapid testing with delayed confirmatory testing: $1,336,231, 1999 DALYs	$965 per DALY averted (rapid testing more costly and more effective)		Cost-effectiveness was sensitive to changes in the prevalence, changes to sensitivity and specificity of rapid testing and confirmatory testing		
	Delayed testing: $1,107,118, 2236 DALYs	Scenario 2		Scenario 1:		
	Scenario 2 (no access to confirmatory testing)	$1490 per DALY averted (clinical judgement more costly and more effective)		Rapid testing had probability of 52.5% of being cost-effective at threshold		
	Rapid testing: $998,260.67, 2538 DALYs			Scenario 2:		
	Clinical judgement: $1,261,230, 2361 DALYs			Rapid testing had probability of 71% of being cost-effective at threshold		
Carter & Conversano 2021 (Carta and Conversano, 2021)	SoC: $33,370, 0.767 QALYs	R + D dominates both SoC and dexamethasone	$50K/QALY gained	OWSA results presented vs. SoC only. All ICERs vs. SoC robust except when remdesivir relative effect takes lower bound estimate (R + D: $24.4K/QALY, remdesivir: $261K/QALY)	This analysis supports the use of remdesivir and/or dexamethasone	Analysis is based on limited evidence of treatment effectiveness
	Remdesivir: $32,354, 0.773 QALYs	R + D vs. remdesivir: $5,222/QALY		PSA results consistent with base case		R + D does not have proven effectiveness
	Dexamethasone: $33,556, 0.803 QALYs	Excluding R + D:				Disease progression, long-COVID and different patient characteristics not explored
	R+D: $32,540, 0.809 QALYs	Remdesivir dominates SoC				Proxy utility data used
		Dexamethasone vs remdesivir: $40.6K/QALY				

(Continued on following page)

TABLE 3 (*Continued*) Results of included studies.

Study	Cost and health outcome results	ICER/net benefit of intervention(s) vs. comparator(s)	Cost-effectiveness threshold (if relevant)	Sensitivity & scenario analyses	Authors' conclusions regarding cost effectiveness	Authors' reported limitations and challenges
Congly 2021 (Congly et al., 2021)	Total costs & QALYs	All ICERs for remdesivir strategies are dominated by giving dexamethasone to all (moderate and severe) patients	$100K/QALY gained	Optimal strategy is not sensitive to OWSA	Dexamethasone for all patients was the most cost-effective strategy Dexamethasone for severe cases would be favoured at lower decision thresholds	Analysis is based on limited evidence of treatment effectiveness
	(Strategies denoted by treatment for moderate disease, treatment for severe disease.)	Dexamethasone for severe only, vs. SoC: $285/QALY		PSA: ICER for dexamethasone (all patients) is below £100K/QALY with 98% probability		Fixed DRG costs do not account for different hospital stay durations
	1. SoC, SoC: $11.1K, 0.716	Dexamethasone for all patients vs. severe only: $1,718/QALY				No long-term health outcomes
	2. SoC, Dex: $11.1K, 0.726					Proxy utility data used
	3. Dex, Dex: $11.1K, 0.735					
	4. SoC, Rem: $11.8K, 0.710					
	5. Rem, SoC: $13.1K, 0.725					
	6. Rem, Dex: $13.1K, 0.734					
	7. Rem, Rem: $13.7K, 0.719					
Dijk 2022 (Dijk et al., 2022)	Incremental vs. SoC	ICERs vs. SoC	$100K/QALY gained	Value of information analysis	At a threshold of $100K/QALY gained, treatment with remdesivir, C + I, dexamethasone, B + R and tocilizumab are cost effective versus SoC	Some parameters were estimated from non-COVID studies. Effectiveness estimates drawn from single trials
	Hydroxychloroquine (Hyd): -$12,227, -0.263 QALYs	Hyd: $46,427 (SWQ)		Decisions about Dex, C + I, B + R, L + R and IF would not change with		Analysis focuses on the research and approval health policy questions, not comparisons
	Remdesivir (Rem): -$5, +0.252 QALYs	Rem: Dominant		further evidence		
	C + I: $696, +0.171 QALYs	C + I: $4,075		For Rem and Toc, the value of further evidence would not outweigh the cost of research		
	Dexamethasone (Dex): $6856, +0.614 QALYs	Dex: $11,619		For Hyd, further evidence to investigate decremental cost effectiveness may be worthwhile		
	B + R: $10,673, +0.775 QALYs	B + R: $13,772				
	Tocilizumab (Toc): $35,849, +0.882 QALYs	Toc: $40,633				
	Interferon b1 (IF): -$2,538, -0.472 QALYs	IF: $5,377 (SWQ)				
	L + R: -$1,404, -0.091 QALYs	L + R: $15,418 (SWQ)				
		Fully incremental NR due to heterogeneous SoC arms across trials				

(Continued on following page)

TABLE 3 (*Continued*) Results of included studies.

Study	Cost and health outcome results	ICER/net benefit of intervention(s) vs. comparator(s)	Cost-effectiveness threshold (if relevant)	Sensitivity & scenario analyses	Authors' conclusions regarding cost effectiveness	Authors' reported limitations and challenges
Goswami 2022 (Goswami et al., 2022)	Molnupiravir: $8,795, 17.721 QALYs	Molnupiravir is dominant compared with SoC	$100K per QALY gained	Results robust to scenario and one-way sensitivity analyses. PSA: Molnupiravir 100% likely to have an ICER below threshold	Compared with SoC, treatment with molnupiravir can be considered a cost-effective option in the management of outpatients with COVID-19 at risk of progression to severe disease in the US	Other outpatient treatments for COVID-19 not included. Appropriate utility data were unavailable and required primary research
	SoC: $9,690, 17.512 QALYs					
Kelton 2022 (Kelton et al., 2022)	Partial societal perspective	Partial societal perspective	$50K/QALY gained	All ICERs robust to OWSA, including oxygen/NIV subgroup	B + R is more cost effective than remdesivir alone for patients hospitalised because of COVID-19 in the US	Lack of data to inform long-term burden of COVID-19
	Remdesivir: $372K, 11.7 QALYs	B + R vs. remdesivir: $22.3K/QALY, $17.9K/LYG		B + R more cost effective if no survival benefit (due to future unrelated medical costs avoided)		Analysis does not capture potential readmissions or resource capacity constraints
	B + R: $380K, 12.1 QALYs	Hospital perspective		PSA: consistent with deterministic		Data informing utility values are limited
	Note: >80% of costs composed of other long-term medical costs	B + R dominates remdesivir				National average DRG costs may lack generalisability
	Hospital perspective					
	B + R vs. remdesivir: -$1,778, +0.0018 QALYs					
Kowal 2023 (Kowal et al., 2023)	Deterministic incremental results	Deterministic results	$150,000 per QALY gained ($50K & $100K in sensitivity analyses)	Population NHB by threshold:	Funding COVID-19 treatments reduced the population-level burden of health inequality by 0.234% (or 130,000 QALYs)	Underreporting of COVID-19 cases, hospitalisations deaths, and potentially variable reporting across equity subgroups. 20% of the population was not captured by the DCUA
	Average:	Average: $28,600 per QALY gained		$50K: 391,114 QALYs	Distributional CUA of inpatient COVID-19 treatments may improve overall health while reducing health inqualities	No trial data were identified that reported subgroup effects
	Costs $12,741, QALYs +0.445	Highest deprivation: $28,000 per QALY gained		$100K: 649,456 QALYs		Lack of quality-of-life data at the subgroup level
		Lowest deprivation: $29,800/QALY gained		Population NHB remain positive up to inpatient treatment cost of $60,100 per patient		
		Including inequitable opportunity costs (NHB)				
		Total: 735,569 QALYs				
		Hispanic, highest deprivation: 72,083; lowest deprivation: 1,106				
		Black, highest deprivation: 47,342; lowest deprivation: 1,622				

(Continued on following page)

TABLE 3 (*Continued*) Results of included studies.

Study	Cost and health outcome results	ICER/net benefit of intervention(s) vs. comparator(s)	Cost-effectiveness threshold (if relevant)	Sensitivity & scenario analyses	Authors' conclusions regarding cost effectiveness	Authors' reported limitations and challenges
		White, highest deprivation: 113,982; lowest deprivation: 27,450				
Lau 2022 (Lau et al., 2022)	Remdesivir: $28,276*, 0.809 deaths averted	Dominant	Thresholds of $0, $14,914*, $37,286* and $74,571* used for interpreting PSA	Results similar across deterministic scenarios. Major drivers of cost effectiveness were inpatient care and remdesivir costs	Remdesivir plus SoC is likely the preferred treatment strategy compared with usual care alone, for hospitalised adults with COVID-19	Short time horizon may miss downstream costs and later events
	Placebo: $28,357*, 0.771 deaths averted			Remdesivir dominant in 58% of PSA simulations and below $74,571* in 82%		Data from RCT may not reflect routine clinical practice
Metry 2022 (Metry et al., 2022)	Total costs* & QALYs	In hospital, on oxygen:	$27.5K* /QALY gained	Treatments are more cost effective when duration of long COVID was shorter, and in younger patients. In the community setting, a higher risk of hospitalisation makes early treatment more cost effective	In hospital, all treatments evaluated had scenarios where the ICER vs. SoC was below the threshold	The decision problem has evolved, so studies do not reflect the current conditions. Therefore, many assumptions were required. No head-to-head studies of interventions were identified. Confidential results not published for lenzilumab, molnupiravir or casirivimab + imdevimab
	In hospital, on oxygen:	SoC: reference			In the community setting, N + R may be cost effective compared with SoC	
	SoC: $30,436, 4.61	T: $9,254*				
	Toc: $35,146, 5.12	Rem: dominated				
	Rem: $38,202, 5.08	Bar: $18,812*				
	Baricitinib (Bar): $41,572, 5.46	B + R: dominated				
	B + R: $41,974, 5.32	In hospital, no oxygen:				
	In hospital, no oxygen:	SoC: reference				
	SoC: $13,316, 5.79	Bar: $7,564*				
	Bar: $16,073, 6.29	Rem: dominated				
	Rem: $16,487, 6.07	B + R: dominated				
	B+R: $17,509, 6.21	In the community, high risk:				
	In the community, high risk:	SoC: reference				
	SoC: $1,448, 13.42	N + R: $8,484*				
	N+R: $2,483, 13.53	Sot: dominated				
	Sot: $4,924, 13.48	Rem: dominated				
	Rem: $6,039, 13.45					
Park 2022 (Park et al., 2022)	Incremental results (costs and DALYs averted)	Treatment with C + I vs. SoC:	1.15 gross national income per DALY = $74K in 2021	Results were robust to sensitivity analyses setting the relative risk reduction to the 95% CI bounds	All strategies considered were cost effective using the specified threshold	Study prior to widespread circulation of delta and omicron disease variants. Efficacy and cost effectiveness of C + I may differ by variant
	Treatment with C+I vs SoC:	Dominant			Treating people aged ≥60 was the most cost saving strategy	

(Continued on following page)

TABLE 3 (*Continued*) Results of included studies.

Study	Cost and health outcome results	ICER/net benefit of intervention(s) vs. comparator(s)	Cost-effectiveness threshold (if relevant)	Sensitivity & scenario analyses	Authors' conclusions regarding cost effectiveness	Authors' reported limitations and challenges
	Treat ages ≥80: -$0.08 m, 38	Dominant				
	Treat ages ≥70: -$0.1 m, 66	Dominant				
	Treat ages ≥60: -$0.34 m, 161	$800/DALY averted				
	Treat ages ≥50: +$0.17 m, 198					
Rafia 2022 (Rafia et al., 2022)	Total costs* and QALYs (probabilistic)	If remdesivir has a survival effect:	$27.5K* /QALY gained	ICERs most affected by time horizon, baseline survival with SoC, and inclusion of unrelated costs. At analysis price, remdesivir mortality HR must be 0.915 or higher to be cost effective	Remdesivir is likely to be cost effective only if it prevents death, and this is highly uncertain within the supplemental oxygen population	Rapidly changing context means some parameter estimates and assumptions out of date
	If remdesivir has a survival effect:	ICER vs. SoC: $17,056*		PSA: ICER below threshold with 74% probability if it confers a survival benefit, else 0%		Model cannot track individual patients
	SoC: $12,920, 6.35	If remdesivir has no survival effect:				Analyses conducted at list prices, may not reflect true prices paid
	Remdesivir: $17,549, 6.62	ICER vs. SoC: >$1M.				Potentially some double counting of COVID-19 disutility
	If remdesivir has no survival effect:					
	SoC: $14,190, 6.35					
	Remdesivir: $16,481, 6.35					
Ruggeri 2022 (Ruggeri et al., 2022a)	Incremental results	NR	NR	Results sensitive to Rt; admission, ICU and mortality rates; remdesivir treatment effect. However, conclusions remain the same	The ability of remdesivir to decrease ward LoS and ICU admissions would produce signifcant cost savings for hospitals, a more manageable hospital capacity in a public health emergency, and a faster recovery for hospitalised patients who require supplemental oxygen	Infection forecasts were informed by various sources, including historical data and expert opinion, and are therefore uncertain. Potential side effects of remdesivir were not included
	23,579 cases:			PSA: results not reported in detail, but remdesivir appears to be cost-incurring (i.e., not dominant) in a significant proportion of PSA results		
	Costs -$27.8 m*					
	Deaths averted 165.9					
	Calibrated to 1,000 cases:					
	Costs -$1.2 m*					
	Deaths averted 7.0					

(Continued on following page)

TABLE 3 (*Continued*) Results of included studies.

Study	Cost and health outcome results	ICER/net benefit of intervention(s) vs. comparator(s)	Cost-effectiveness threshold (if relevant)	Sensitivity & scenario analyses	Authors' conclusions regarding cost effectiveness	Authors' reported limitations and challenges
Ruggeri 2022 (Ruggeri et al., 2022b)	Incremental results	NR	NR	Results sensitive to Rt values, ICU and mortality rates, baseline hospitalisation and remdesivir mortality effect, but conclusions remain the same	In Saudi Arabia, remdesivir plus standard of care has the potential to reduce healthcare resource use, mortality, and costs when compared with	Some infection forecasts were informed by expert opinion, and are therefore uncertain. Many inputs informed by targeted, rather than systematic, literature review, including only 1 RCT. Treatment-related adverse eventsnot captured
	Static infection rate (178,405 cases):			Rt = 0.8 (decreasing; 109,087 cases): costs -$154.7 m[1], DA 815	standard of care alone across a range of plausible local epidemiological scenarios	
	Costs -$174.81 m			Calibrated to 1,000 cases: costs -$1.4 m, DA 7.5		
	Deaths averted (DA) 1.2			Rt = 1.2 (increasing; 247,724 cases): costs -$377.3 m, DA 1,582		
	Calibrated to 1,000 cases:			Calibrated to 1,000 cases: costs -$1.5 m, DA 6.4		
	Costs -$979,836			PSA: remdesivir is dominant in 93% of simulations		
	Deaths averted 6.7					
Ruggeri 2023 (Ruggeri et al., 2023)	Incremental results	NR	NR	Results sensitive to Rt; admission, ICU and mortality rates; C + I effect on admissions. However, conclusions remain the same	[With C + I] hospitals can achieve important cost savings while patients can experience a more favourable disease course [including reduction in death]	Epidemiological model based on estimated parameters, including Rt. Limited clinical evidence about C + I (1 RCT). True price of C + I in Italy is not known, therefore this analysis uses the US price. Dominant COVID-19 variants at the time of publication (alpha and delta) are not the variant that C + I is likely to be active against (omicron; prevalence 4.76%)
	194,451 cases:			PSA: C + I dominant in more than 90% of simulations		
	Costs -$82.4 m*					
	Deaths averted 1,535					
	Calibrated to 1,000 cases:					
	Costs -$423,730*					
	Deaths averted 7.9					
Savinkina 2022 (Savinkina et al., 2022)	Base-case (high) effect scenario, calibrated to 1,000 patients:	Base-case (high) effect scenario:	$10,000 to $5 m per DA	ICERs, low-effect scenario:	For almost every scenario prescribing N + R to unvaccinated patients at high risk of severe COVID-19 was cost saving. This group should almost always be treated if treatment is available	Analysis does not consider drug supply, budgetary constraints, non-adherence, contraindications to N + R, other active treatments, differential costs in different vacc and risk groups, or transmission dynamics
	No N + R: $221K, 0.77 deaths	No N + R: baseline		No N + R: baseline		
	N + R for unvacc high risk: $182K, 0.51 deaths	N + R for unvacc high risk: dominant		N + R for unvacc high risk: $319K per DA		
	N + R for all high risk: 0.29 deaths, $273K	N + R for all high risk: $397K per DA		N + R for all high risk: $2.6 m per DA		

(Continued on following page)

TABLE 3 (*Continued*) Results of included studies.

Study	Cost and health outcome results	ICER/net benefit of intervention(s) vs. comparator(s)	Cost-effectiveness threshold (if relevant)	Sensitivity & scenario analyses	Authors' conclusions regarding cost effectiveness	Authors' reported limitations and challenges
	N + R for all high risk and unvacc low risk: 0.22 deaths, $348K	N + R for all high risk and unvacc low risk: $1.0 m per DA		N + R for all high risk and unvacc low risk: $5.3 m per DA		
	N + R for all: 0.18 deaths, $566K	N + R for all: $5.0m per DA		N + R for all: $22.1m per DA		
				Cost results reported for various OWSA values (but ICERs NR)		
Shah 2023 (Shah et al., 2023)	Not reported (incremental only)	Advanced CC vs none:	$101 per DALY averted (conservative threshold for Tanzania)	Probability of essential and emergency care being cost effective is 96% and 99% compared to no care and district level care at Tanzanian threshold	Essential and emergency critical care is likely to be highly cost effective in low-resource settings	Analysis relies on low quality sources for parameters due to scarcity of data, does not include needs of moderate patients, and did not reflect availability of regional and referral hospitals
		$186 per DALY averted		In deterministic analyses, results were most sensitive to effectiveness of essential and emergency care in preventing severe cases becoming critical, unit costs of advanced care		Markov model cannot capture pace of change of treatment even within 24 h cycle
		Essential CC vs none:				Triangular distributions used may be less appropriate but reflect uncertain nature of data
		$37 per DALY averted				
		Advanced CC vs district:				
		$144 per DALY averted				
		Essential CC vs. district:				
		$14 per DALY averted				
Yeung 2022 (Yeung et al., 2022)	Costs and QALYs	ICERs vs. SoC	$50K-150K per QALY gained	PSA (healthcare perspective), probability ICER < $50K, $100K, $150K:	At their current prices, each intervention is estimated to meet standard cost-effectiveness levels in the US healthcare system, even under a scenario with a lower hospitalisation risk that may reflect the Omicron wave	Analysis underpinned by immature evidence base and heterogenous trial designs, including non-US settings and different prevalent COVID-19 variants
	Healthcare perspective	Healthcare perspective		Mol: 31%, 69%, 84%		Modified societal perspective has limited scope
	Molnupiravir (Mol): $298.5K, 15.938	Mol: $61K		N + R: 97%, 100%, 100%		
	N + R: $298.5K, 15.964	N + R: $21K		Flu: 100%, 100%, 100%		
	Fluvoxamine (Flu): $297.8K, 15.939	Flu: $8K		Key scenarios, (ICERs vs. SoC):		
	SoC: $297.7K, 15.925	Modified societal perspective (approx.)		Unvaccinated population:		

(Continued on following page)

TABLE 3 (*Continued*) Results of included studies.

Study	Cost and health outcome results	ICER/net benefit of intervention(s) vs. comparator(s)	Cost-effectiveness threshold (if relevant)	Sensitivity & scenario analyses	Authors' conclusions regarding cost effectiveness	Authors' reported limitations and challenges
	Modified societal perspective	Mol: $43K		Mol: $48K		
	Mol: $301.4K, 15.952	N + R: $26K		N + R: $15K		
	N + R: $302.3K, 16.006	Flu: $20K		Flu: $4K		
	Flu: $300.8K, 15.954			Lower hospitalisation risk (e.g., Omicron variant):		
	SoC: $300.2K, 15.925			Mol: $74K		
				N + R: $34K		
				Flu: $21K		

Abbreviations: Bar, baricitinib; B + R, baricitinib and remdesivir; CC, critical care; CUA, cost—utility analysis; C + I, casirivimab + imdevimab; DALY, disability-adjusted life-year; DA, DALY averted; DCUA, distributional cost—utility analysis; Dex, dexamethasone; DRG, diagnostic-related group; Flu, fluvoxamine; HR, hazard ratio; Hyd, hydroxychloroquine; ICER, incremental cost-effectiveness ratio; ICU, intensive care unit; IF, interferon beta-1a; L + R, lopinavir + ritonivir; Mol, molnupiravir; N + R, nirmatrelvir + ritonavir; NR, not reported; QALY, quality-adjusted life-year; RCT, randomised controlled trial; Rt, disease reproduction rate; R + D, remdesivir and dexamethasone; SoC, standard of care; Sot, sotrovimab; SWQ, south-west quadrant (of the cost-effectiveness plane, i.e., lower cost and lower effectiveness); Toc, tocilizumab.

Notes: (Elvidge et al., 2022) This study (Ruggeri et al., 2022b) reports the cost results for "static" and "decreasing" infection rate scenarios the other way around, such that the cost in the "static" scenario is lower than the cost under decreasing infection rates. This appears to be an error, therefore we have swapped the cost results.

*Cost conversions to USD listed below. The OECD exchange rate for the reported price year is used (Exchange rates indicator, 2023). Where no price-year is explicitly reported, we have assumed the relevant exchange rate is the year prior to the year of publication.

- Alamer 2023 (Alamer et al., 2023): 1 USD = 3.750 SAR (2020).
- Lau 2022 (Lau et al., 2022): 1 USD = 1.341 CAD (2020).
- Metry 2022 (Metry et al., 2022), Rafia 2022 (Rafia et al., 2022): 1 USD = 0.727 GBP (2021).
- Ruggeri 2022 (Ruggeri et al., 2022a): 1 USD = 0.845 EUR (2021).
- Ruggeri 2023 (Ruggeri et al., 2023): 1 USD = 0.950 EUR (2022).

long-term Markov component (3/17) (Goswami et al., 2022; Kelton et al., 2022; Yeung et al., 2022); partitioned survival models (2/17) (Metry et al., 2022; Rafia et al., 2022); and a patient-level simulation (1/17) (Alamer et al., 2023). A potential financial conflict of interest in favour the intervention under evaluation was reported by 6/18 included studies (Ruggeri et al., 2022b; Goswami et al., 2022; Kelton et al., 2022; Lau et al., 2022; Ruggeri et al., 2023).

3.2 Cost effectiveness

3.2.1 Treatments: inpatient hospital setting

For evaluations based in inpatient hospital populations, with or without supplemental oxygen, 8/12 studies were CUAs that specified one or more willingness-to-pay thresholds to determine cost effectiveness of evaluated interventions. Dexamethasone was found to be cost effective compared with standard care (Carta and Conversano, 2021; Congly et al., 2021; Dijk et al., 2022); this conclusion was robust to sensitivity analyses, and a value of information analysis indicated there would be no value in further research (Dijk et al., 2022). Remdesivir was also generally cost effective *versus* standard care (Carta and Conversano, 2021; Congly et al., 2021; Dijk et al., 2022; Rafia et al., 2022), though 1 study noted that this result was highly sensitive to whether it confers a survival benefit or not (Rafia et al., 2022). If it does, the reported ICER was around $17,000 per QALY gained, rising to over $1 million per QALY gained if it does not reduce the risk of death. Its mortality hazard ratio should be at least 0.92 for it to be cost effective. Further, 1 study found that using dexamethasone for all hospitalised patients dominated any strategy that involved remdesivir (Congly et al., 2021). Baricitinib in addition to remdesivir was found to be cost effective *versus* standard care in 2 studies (Dijk et al., 2022; Metry et al., 2022), and this was robust to sensitivity analyses. In a US study (Kelton et al., 2022), it had an ICER of around $22,000 per QALY gained compared with using remdesivir alone, in a partial societal analysis, and was dominant from a hospital perspective. However, this was in conflict with a fully incremental analysis from a United Kingdom healthcare perspective (Metry et al., 2022), that suggested barcitinib monotherapy was the most cost effective treatment for hospitalised patients, with ICERs of around $7,500–19,000 per QALY gained depending on the patient's need for oxygen. The conflicting results may be explained by healthcare resource cost differences between the US and United Kingdom, or the studies' different sources for relative effectiveness data; one (Kelton et al., 2022) made use of data on file from a single trial (ACTT-2), while the other (Metry et al., 2022) used outputs from published "living" network meta analyses (metaEvidence Initiative, 2022; The COVID-NMA Initiative, 2021). In a study that compared treatments with standard care but not with each other (Dijk et al., 2022), casirivimab + imdevimab and tocilizumab were estimated to have ICERs of around $13,000 and $40,000 per QALY gained, respectively, while hydroxychloroquine, interferon beta-1a and

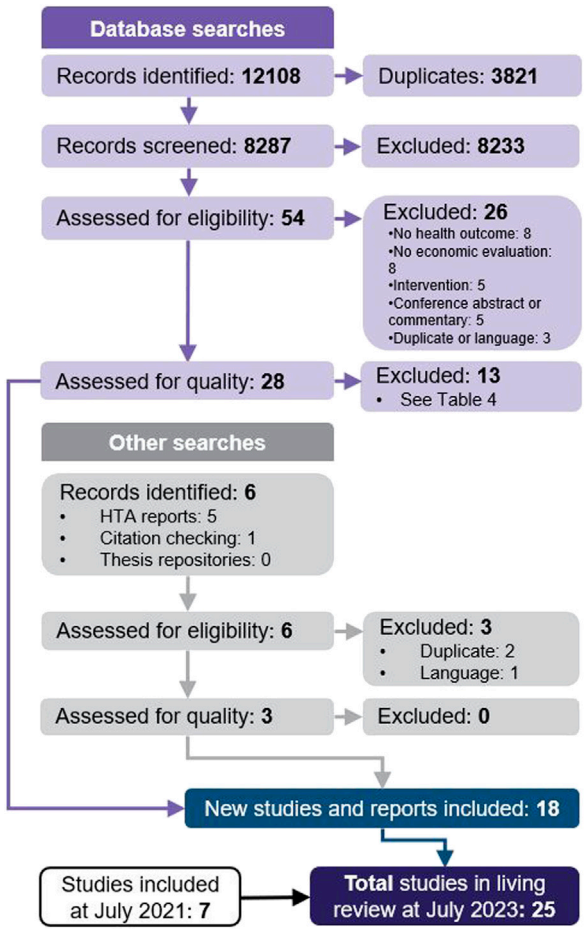

FIGURE 1
PRISMA diagram for study selection.

lopinavir + ritonivir were found to be cost saving but detrimental to health outcomes (QALYs). The resulting southwest-qaudrant ICERs were around $46,000, $5,500 and $15,000, respectively, which, at the specified threshold of $100,000 per QALY gained, imply the cost savings would not be sufficient to offset the health outcomes foregone. Value of information analysis identified that further research to examine disinvestment in hydroxychloroquine may be worthwhile, though it is not widely used.

One QALY-based evaluation of inpatient pharmacological treatment (Kowal et al., 2023) was a distributional re-analysis of a study previously included in this review (Sheinson et al., 2021). Treatment was found to be more cost effective in more deprived populations, with an ICER of $28,000 per QALY gained in the most deprived group and $29,800 in the least deprived group. Including the existing inequitable distribution of opportunity costs in the US health system, population-level net health benefits would remain positive up to a treatment cost of $60,100 per patient.

The one other CUA in the inpatient setting (Shah et al., 2023) evaluated using different levels of critical care to treat people with COVID-19, and used DALYs averted as the health outcome. At a specified conservative threshold in Tanzania of $101 per DALY averted, using essential critical care (e.g., supplemental oxygen) for people with COVID-19 was cost effective compared with no critical care, with an ICER of $37 per DALY averted, and district-level critical care ($14 per DALY averted). The equivalent ICERs for using advanced critical care (e.g., mechanical ventilation critical disease) were above the threshold, at $186 and $144, respectively.

There were 4 CEAs evaluating treatments in the inpatient setting, with all using deaths averted as the health outcome. Two studies used the same economic model with country-specific input data, and found remdesivir to reduce deaths and costs compared with standard care in Portugal [7 deaths averted and $1.2 m saved per 1,000 cases (Ruggeri et al., 2022a)] and Saudi Arabia [6.7, $980,000 (Ruggeri et al., 2022b)]. One trial-based analysis reached similar results in the setting of Canada (Lau et al., 2022); it was dominant in 58% of probabilistic sensitivity analysis (PSA) simulations, and the ICER was below $100,000 for death averted in 82%. Favipiravir was evaluated by 1 study (Alamer et al., 2023), and was also found to reduce deaths and costs in Saudi Arabia, with a saving of $4,500 per death averted.

3.2.2 Treatments: outpatient and community setting

For patients treated in the outpatient and community setting, at risk of progressing to severe disease requiring inpatient or critical care, 4 studies compared active treatments with standard care only. Among them, casirivimab + imdevimab was found to have an ICER of $800 per DALY averted in 1 study (Park et al., 2022)–below the specified $74,000 per DALY threshold in Singapore–and to reduce deaths and save costs in Italy in another study (Ruggeri et al., 2023). The conclusions of both studies were robust to sensitivity analyses undertaken; in the latter case, casirivimab + imdevimab was dominant in over 90% of simulations. One study estimated that molnupiravir would dominate standard care, generating incremental QALYs and reducing costs (Goswami et al., 2022).

In the only identified study that explicitly compared different strategies for using a treatment in subgroups defined by vaccination status (Savinkina et al., 2022), base-case results suggested that nirmatrelivir + ritonivir would dominate standard care for unvaccinated high-risk groups. However, that result was sensitive to the relative effect estimate, with a plausible lower-bound effect leading to an ICER of over $300,000 per death averted. Nirmatrelivir + ritonivir would be less cost effective if used in vaccinated and low-risk groups, with the ICER rising to $5 m per death averted if used for all patients. In a US HTA analysis (Yeung et al., 2022), it had an ICER of $21–26,000 per QALY gained *versus* standard care, depending on the perspective chosen, and in a United Kingdom fully incremental analysis conducted for HTA (Metry et al., 2022), the ICER was around $8,500 per QALY gained (remdesivir and sotrovimab were dominated). The US HTA study also found fluvoxamine would be cost effective compared with standard care relative to typical US thresholds (ICER: $8–20,000 per QALY gained), while molnupiravir had an ICER of $61,000 per QALY gained from a healthcare perspective, and $43,000 per QALY gained from a partial societal perspective that captured productivity costs. Therefore, the perspective would be a relevant factor for healthcare decision makers using the specified conservative threshold of $50,000 per QALY. In a scenario focusing on an unvaccinated

TABLE 4 Studies excluded due to having very serious limitations.

Study	Comparators & type of evaluation	Summary of very serious limitations that affect reliability of study conclusions
Beshah 2023 (Beshah et al., 2023)	Non-invasive ventilation, mechanical ventilation CUA	Lifetime horizon, but no long-term outcomes included. Insufficient information about source of baseline outcomes. Relative effectiveness and resource use inputs sourced from a single non-randomised study without adjustment for selection bias. Limited analysis of uncertainty
Chow 2022 (Chow et al., 2022)	Statins, SoC CUA	Health outcome is not defined and cannot be accurately inferred from the information reported. Relative effectiveness and resource use inputs sourced from non-randomised studies only. Unit costs are informed by World Health Organisation cost codes, generalisability to the study setting is unclear. No analysis of uncertainty reported
Gandjour 2022 (Gandjour, 2022)	Off-the-shelf self tests, personal protective measures + no testing CEA	The design and appropriateness of the model structure are unclear (no schematic). Downstream effects of test results are not considered. Only costs associated with the test appear to have been included, and the unit price source is not reported. An ICER is reported, but the component incremental costs and QALYs are not. Many parameters are not subjected to uncertainty analysis
Jovanoski 2022 (Jovanoski et al., 2022)	Casirivimab + imdevimad, SoC CUA	Some modelling assumptions and cost data sources may overstate the impact of the intervention. ICERs are reported, but the component incremental costs and QALYs are not. Minimal analysis of uncertainty is reported. There is a potential conflict of interest
Kilcoyne 2022 (Kilcoyne et al., 2022a)	Lezilumab, SoC CEA.	Time horizon (28 days) is too short to capture all relevant cost and health outcomes. Hospitalisation costs are sourced from a modelling study, when national schedules of costs are available. Estimates of clinical effectiveness are drawn from a single manufacturer funded RCT. Absolute values of intervention effect are used, which imposes the assumption of independent prior distributions in the treatment and comparator arms, which is unlikely. A joint measure of cost-effectiveness is not presented. A probabilistic sensitivity analysis is not performed. There is a potential conflict of interest
Kilcoyne 2022 (Kilcoyne et al., 2022b)	Lezilumab, SoC CEA.	Time horizon (28 days) is too short to capture all relevant cost and health outcomes, though a scenario analysis extends the time horizon to 1 year. Absolute values of intervention effect are used, which imposes the assumption of independent prior distributions in the treatment and comparator arms. A joint measure of cost-effectiveness is not presented. A probabilistic sensitivity analysis is not performed. There is a potential conflict of interest
Krylova 2021 (Krylova et al., 2021)	Favipiravir, umifenovir CEA	No time horizon; clinical outcomes are directly dependent on the source of effectiveness evidence used (17 or 28 days). No long-term or downstream outcomes included. For one comparison, different RCTs are used to inform effectiveness, without adjustment for potential confounding. Substantial use of assumptions to inform resource use parameters. No uncertainty analysis
Ohsfeld 2021 (Ohsfeldt et al., 2021)	Baricitinib, SoC CEA, CUA.	Absolute values of intervention effect are used, which imposes the assumption of independent prior distributions in the treatment and comparator arms. Extensive use of unpublished "data on file", assumptions and other various sources to inform resources use and cost parameters. Limited justification for ranges used in sensitivity analysis. There is a potential conflict of interest
Oksuz 2021 (Oksuz et al., 2021)	Remdesivir, SoC CUA	Time horizon ("a COVID-19 episode") is unclear but is likely to be too short to capture all relevant cost and health outcomes. Baseline, relative effectiveness and resource use outcomes from unadjusted real-world data ($n = 78$). PSA distributions are not reported. There is a potential conflict of interest
Petrov 2022 (Petrov et al., 2022)	Anakinra, baricitinib, kanakinumab, levilimab, olokizumab, netakimab, sarilumab, secukinumab, tocilizumab, tofacitinib CBA	The cited clinical evidence is insufficient to justify the assumption of equal effectiveness required by the chosen cost-minimisation approach. It is inappropriate to compare the studied interventions because are intended for different patient populations
Schallner 2022 (Schallner et al., 2022)	Intensive care, non-intensive care CUA	Baseline outcomes for the standard care comparator arm (instant death) were arbitrary researcher assumptions, with no support from experts or data reported. Relative effectiveness inputs sources from a single, small study with a historical control. Costs associated with standard care not

(Continued on following page)

TABLE 4 (*Continued*) Studies excluded due to having very serious limitations.

Study	Comparators & type of evaluation	Summary of very serious limitations that affect reliability of study conclusions
		considered (instand death would not be costless). Limited analysis of uncertainty reported
Subhi 2023 (Subhi et al., 2023)	Remdesivir, favipiravir, SoC CEA	No long-term outcomes are included. Source of baseline outcomes appears to be a trial conducted in a different setting. Relative effectiveness estimates for favipiravir rely on researcher assumptions. Indirect comparison between remdesivir and favipiravir is a naïve comparison. Resource use inputs informed by expert elicitation. Details of the experts and elicitation process are not reported. There is a potential conflict of interest
Wai 2023 (Wai et al., 2023)	Molnupiravir, nirmatrelvir + ritonavir, SoC CEA	Time horizon (28 days) is too short to capture all relevant cost and health outcomes. Baseline outcomes and relative effectiveness outcomes sources from a single non-randomised study. Unclear how resource use inputs were recorded and how cost inputs were sourced. No analysis of uncertainty reported

Abbreviations: CBA, cost-benefit analysis; CEA, cost-effectiveness analysis; CUA, cost—utility analysis; ICER, incremental cost-effectiveness ratio; QALY, quality-adjusted life-year; RCT, randomised controlled trial; SoC, standard of care.

patient population, in whom the effects of COVID-19 may be more severe, the US HTA (Yeung et al., 2022) found that nirmatrelivir + ritonivir, fluvoxamine and molnpiravir would be more cost effective *versus* standard care (with healthcare perspective ICERs of $15,000, $4,000 and $48,000 per QALY gained, respectively).

3.2.3 Diagnostic tests

The single study that evaluated a diagnostic test (Arwah et al., 2023) found that rapid antigen tests, plus a delayed nucleic acid amplifying test (NAAT) used in a confirmatory way, had an ICER of $965 per DALY averted compared with typical standard care in Kenya: delayed NAAT alone. This would be cost effective by a close margin relative to the specified local threshold of $1,003 per DALY; in PSA, the probability of the ICER being below the threshold was 53%. The ICER was sensitive to the underlying disease prevalence and the accuracy of rapid and confirmatory tests. In a scenario where NAAT is not available, rapid testing was estimated to dominate a "no testing" strategy that relies on clinical judgement.

4 Discussion

4.1 Principal findings

This updated systematic review indicates that pharmacological treatments that have been repurposed for to treat COVID-19 in recent years have the potential to be cost effective. In particular, the use of the low-cost corticosteroid dexamethasone—which has become routine practice to treat severe COVID-19 in an inpatient setting—appears to be clearly cost effective. Remdesivir and baricitinib, potentially in combination, appear to be promising candidates to treat severe disease, too. Limited cost-effectiveness evidence in the inpatient setting for casirivimab + imdevimab, tocilizumab, hydroxychloroquine, interferon beta-1a and lopinavir + ritonivir suggests the latter 3 treatments may produce worse health outcomes than standard care without a commensurate cost saving to be considered by decision makers.

In the outpatient and community setting, there is some evidence that casirivimab + imdevimab, fluvoxamine and molnupiravir may be cost effective over standard care. Results from 3 studies indicate that nirmatrelivir + ritonivir may be a cost effective treatment in the community setting among patients at high risk of hospitalisation (such as unvaccinated people), though the one fully incremental analysis among them does not include casirivimab + imdevimab, fluvoxamine or molnupiravir in the published results. The 2 studies that evaluated non-pharmacological interventions were both in lower-income settings. They suggested that rapid antigen tests may be cost effective where there is slow existing testing infrastructure, and certainly where there is none; and using the most advanced forms of critical care to treat COVID-19 might be difficult to justify, on cost-effectiveness grounds, in a resource-limited setting.

Compared with the first iteration of this review (Elvidge et al., 2022) conducted in July 2021, this update has identified economic evaluations of a much larger set of interventions for COVID-19. Previously, the evidence base was limited to evaluations of monotherapy and combination therapy use of dexamethasone and remdesivir, which were early candidate interventions for the treatment of COVID-19 in hospital. While we have identified additional cost-effectiveness evidence regarding these treatments, other pharmacological interventions have received marketing authorisation to treat the disease since 2021, and it is logical that healthcare decisions makers will be interested in understanding which of them offer value for money. Here, we have identified such evidence for antiviral therapies (casirivimab + imdevimab, favipiravir, lopinavir + ritonivir, molnupiravir, nirmatrelivir + ritonivir), immunotherapies (baricitinib, sotrovimab, tocilizumab) and various other repurposed medicines (fluvoxamine, hydroxychloroquine, interferon beta-1a).

We identified 1 study evaluating the cost effectiveness of a test for SARS-CoV-2, representing 6% of our included studies (1/18) compared with 29% in the initial review (2/7). It is likely that this reflects the changing pandemic context over time, such that comparing alternative testing strategies is no longer a prominent concern, relative to assessing the value of the growing number of available treatment options. The study evaluating rapid antigen tests was one of 2 included studies evaluating non-pharmacological interventions; the other estimated the value of using scarce

critical care resources to treat people with COVID-19. Notably, both studies were conducted in lower-income settings (Kenya and Tanzania, respectively). This suggests testing strategies and the allocation of scarce hospital resources for COVID-19 remain a prominent concern for healthcare decision makers in settings where the most effective and newly licensed pharmacological interventions are less likely to be available.

The context around COVID-19 has changed substantially since the first iteration of this review. However, it does not appear that economic evaluations have adapted their methods to reflect these changes. This is likely to limit their scope to inform decision-making. First, vaccination programmes have been successful across the world and the vast majority of people, particularly in high and middle income countries, have now received at least one dose of a vaccine (Mathieu et al., 2020). Despite this, there appears to be limited consideration within economic evaluations of the impact of vaccination on disease severity and implications for cost effectiveness. Some identified studies did report scenario analyses in unvaccinated populations who are more likely to experience severe symptoms, and one (Savinkina et al., 2022) study from the US explicitly compared alternative strategies for using nirmatrelivir + ritonivir in different patient subgroups defined by their vaccination status and risk of hospitalisation. This approach is likely required to properly define who would benefit from treatments, but has not become widespread. Second, there are several reasons that parameters derived from studies completed at different stages of the pandemic may not be generalisable to the present day. These include pressures and constraints on hospital care during acute phases of the pandemic, differing approaches to standard care and changing disease variants. One study (Yeung et al., 2022) reported a scenario analysis in the context of a hypothetical variant with lower baseline risk of hospitalisation, though this did not consider differential treatment effectiveness for different variants. Differential efficacy between variants has been observed in practice, and means cost-effectiveness evaluations may need to increasingly examine value for money in specific subpopulations (Coronavirus COVID-19 Update, 2022). While some studies did note that the changing composition of COVID-19 over time may limit the generalisability of their cost-effectiveness conclusions (Metry et al., 2022; Park et al., 2022; Ruggeri et al., 2023), this appears to be an underconsidered issue. Third, there are now treatments that are established as best-practice due to the emergence of results from large-scale platform trials. Indeed, within this review dexamethasone is highlighted as a low-cost option that is effective for patients with respiratory support. However, there have been limited attempts to compare established treatments with *each other*. There is a need for more fully incremental cost-effectiveness analyses that compare alternative options simultaneously, rather than indirectly through how they compare with standard care. This may require measures of relative effectiveness to be derived from network meta analyses, such as used Metry et al. (2022), rather than data from local sites or individual trials.

In terms of the analytical methods used, most included studies were model-based analyses, which is consistent with the first iteration of this review. The modelling methods used remained similar. Time horizons varied from short term to lifetime; decision tree, Markov, and hybrid model structures were prominent; the utility weights used by CUAs were often generalised from non-COVID sources; and the known long-terms effects of disease ("long COVID") were generally not captured. However, it is notable that some models may be overrepresented in the evidence base, due to repeated adaptations or reanalyses of the same model. In particular, remdesivir was evaluated in settings of Portugal (Ruggeri et al., 2022a) and Saudia Arabia (Ruggeri et al., 2022b) using a common model with country-specific inputs; casirivimab + imdevimab was evaluated in Italy (Ruggeri et al., 2023) using essentially the same model; and an existing CUA of a hypothetical treatment was reanalysed through a distributional lens (Kowal et al., 2023).

4.2 Strengths and limitations

This update to our "living" systematic review is methodologically consistent with the first publication (Elvidge et al., 2022), and so the same issues concerning the search strategy and generalisability of findings apply here. Our review aimed to provide a comprehensive account of available evidence and as such, a large number of unique records were identified by the database search (8,287); this is due to the known sensitivity of search terms used to identify economic evaluations (Hubbard et al., 2022). This increases the sensitivity of the search, reducing the likelihood of missing relevant studies, but it also means future updates will continue to require a labour-intensive screening process. In addition, like before, we chose to exclude studies that met our selection criteria but were judged to have very serious limitations that may materially affect the conclusions about cost effectiveness. This follows the process used in NICE clinical guideline development (National Institute for Health and Care Excellence NICE, 2012), to avoid conflating results across studies of varying quality, and was predefined in our study protocol (Elvidge et al., 2021). In this update, it led to the exclusion of 13 potentially relevant studies (Table 4). While 2 reviewers conducted this quality assessment, discussing and resolving any disagreements, we recognise that this is necessarily subjective; other reviewers may have reached different quality assessment decisions, or simply included all studies that met the selection criteria. There may also be included studies which did not meet the bar for exclusion, but which have analytical flaws or are based on parameters that are biased or do not reflect best available evidence and this may impact their findings. Further, we excluded public health interventions (e.g., lockdowns, face masks) and vaccinations from our review. The economic value of such interventions may still be of interest to some healthcare decision makers, for example where vaccine coverage is relatively low. Finally, our review is, like any, at risk of publication bias. Several of the included studies exhibit a potential conflict of interest. We cannot know whether those analyses would have reached publication if they had demonstrated negative conclusions about the intervention, nor how many such analyses exist and were not published.

4.3 Implications for future research

After 3 years' worth of cost-effectiveness research for COVID-19 interventions, there are some new and some

persistent evidence gaps that would benefit from further thought. Head-to-head economic evaluations of interventions are in the minority of the identified studies. There are now several treatments available in both the pre-hospital and hospital settings, and fully incremental analyses that compare options simultaneously may be valuable for decision making. The cost effectiveness of tests and treatments may be influenced by what happens later in the clinical pathway, and so a whole-disease model that reports fully incremental results would be a valuable resource. In the context of a continuously evolving disease, and with variable standard care and vaccination uptake in different settings, a whole-disease model could allow for rapid re-analyses in light of new evidence or changes in the prevailing conditions. In general, researchers should routinely reflect on the potential implications of vaccination status and disease variants for the generalisability of their conclusions. Further, these authors recommend that researchers routinely conduct detailed sensitivity analyses examining potential cost effectiveness under a wide range of baseline outcomes and relative effectiveness. Such analyses may help to 'future-proof' their studies to ensure they remain useful under different prevailing coditions. There remains a need for robust evidence about the health-related quality of life impact of COVID-19 in the short and long term, to support the conduct of CUAs. Finally, several identified studies were CEAs that used deaths averted as their health outcome with relatively short time horizons. This may be reasonable in some circumstances, and particularly if one intervention appears both more effective and have lower costs. However, in general, decision makers should be aware that where a treatment has an effect on survival, short-term analyses will not fully capture all relevant differences in outcomes between it the comparator.

5 Conclusion

This updated review of economic evaluations of tests and treatments for COVID-19, covering the period from July 2021 and July 2023, provides a contemporary summary of the cost-effectiveness evidence. Compared with the first iteration of the review (up to July 2021), we have identified 18 additional studies of acceptable quality that healthcare decision makers, such as HTA and payer organisations, may consider to inform their COVID-related decision making. In particular, the evidence may support prioritisation between the numerous antiviral therapies and immunotherapies. Conclusions about some treatments, such as dexamethasone (cost effective) and hydroxychloroquine (not cost effective), support current standard practice. An evidence gap remains for a whole-disease model that can support holistic decision making about multiple tests and treatments at linked decision points in a fully incremental way.

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

JE: Conceptualization, Data curation, Formal Analysis, Methodology, Project administration, Supervision, Writing–original draft, Writing–review and editing. GH: Formal Analysis, Writing–original draft, Writing–review and editing. NN: Formal Analysis, Writing–original draft. DN: Conceptualization, Data curation, Validation, Writing–review and editing. DD: Formal Analysis, Methodology, Supervision, Writing–review and editing.

Funding

The authors declare financial support was received for the research, authorship, and/or publication of this article. The HTx project has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No 825162. This dissemination reflects only the author's view and the European Commission is not responsible for any use that may be made of the information it contains.

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.

Supplementary material

The Supplementary Material for this article can be found online at: https://www.frontiersin.org/articles/10.3389/fphar.2023.1291164/full#supplementary-material

References

Alamer, A., Almutairi, A. R., Halloush, S., Al-jedai, A., Alrashed, A., AlFaifi, M., et al. (2023). Cost-effectiveness of Favipiravir in moderately to severely ill COVID-19 patients in the real-world setting of Saudi arabian pandemic referral hospitals. *Saudi Pharm. J.* 31 (4), 510–516. doi:10.1016/j.jsps.2023.02.003

Arwah, B., Mbugua, S., Ngure, J., Makau, M., Mwaura, P., Kamau, D., et al. (2023). Cost & cost-effectiveness of implementing SD biosensor antigen detecting SARs-CoV-2 rapid diagnostic tests in Kenya. Preprint. *Health Econ*. doi:10.1101/2023.01.05.23284225

Beshah, S. A., Zeru, A., Tadele, W., Defar, A., Getachew, T., and Fekadu Assebe, L. (2023). A cost-effectiveness analysis of COVID-19 critical care interventions in Addis Ababa, Ethiopia: a modeling study. *Cost Eff. Resour. Allocation* 21 (1), 40. doi:10.1186/s12962-023-00446-8

Carta, A., and Conversano, C. (2021). Cost utility analysis of Remdesivir and Dexamethasone treatment for hospitalised COVID-19 patients - a hypothetical study. *BMC Health Serv. Res.* 21 (1), 986. doi:10.1186/s12913-021-06998-w

Chow, R., Simone, C. B., Ii, Prsic, E. H., and Shin, H. J. (2022). Cost-effectiveness analysis of statins for the treatment of hospitalized COVID-19 patients. *Ann. Palliat. Med.* 11 (7), 2285–2290. doi:10.21037/apm-21-2797

Congly, S. E., Varughese, R. A., Brown, C. E., Clement, F. M., and Saxinger, L. (2021). Treatment of moderate to severe respiratory COVID-19: a cost-utility analysis. *Sci. Rep.* 11 (1), 17787. doi:10.1038/s41598-021-97259-7

Coronavirus (COVID-19) Update (2022). *Coronavirus (COVID-19) update: FDA limits use of certain monoclonal antibodies to treat COVID-19 due to the omicron variant [press release]*.

Dijk, S. W., Krijkamp, E. M., Kunst, N., Gross, C. P., Wong, J. B., and Hunink, M. G. M. (2022). Emerging therapies for COVID-19: the value of information from more clinical trials. *Value Health* 25 (8), 1268–1280. doi:10.1016/j.jval.2022.03.016

Elvidge, J., Summerfield, A., Nicholls, D., and Dawoud, D. (2021). *Diagnosis and treatment of COVID-19: a systematic review of economic evaluations*. PROSPERO review protocol: CRD42021272219.

Elvidge, J., Summerfield, A., Nicholls, D., and Dawoud, D. (2022). Diagnostics and treatments of COVID-19: a living systematic review of economic evaluations. *Value Health* 25 (5), 773–784. doi:10.1016/j.jval.2022.01.001

Exchange rates (indicator) (2023). Exchange rates (indicator). Available from: https://data.oecd.org/conversion/exchange-rates.htm (Accessed September 08, 2023).

Gandjour, A. (2022). Benefits, risks, and cost-effectiveness of COVID-19 self-tests from a consumer's perspective. *BMC Health Serv. Res.* 22 (1), 47. doi:10.1186/s12913-021-07277-4

Goswami, H., Alsumali, A., Jiang, Y., Schindler, M., Duke, E. R., Cohen, J., et al. (2022). Cost-effectiveness analysis of molnupiravir versus best supportive care for the treatment of outpatient COVID-19 in adults in the US. *PharmacoEconomics* 40 (7), 699–714. doi:10.1007/s40273-022-01168-0

Hubbard, W., Walsh, N., Hudson, T., Heath, A., Dietz, J., and Rogers, G. (2022). Development and validation of paired MEDLINE and Embase search filters for cost-utility studies. *BMC Med. Res. Methodol.* 22 (1), 310. doi:10.1186/s12874-022-01796-2

Jovanoski, N., Kuznik, A., Becker, U., Hussein, M., and Briggs, A. (2022). Cost-effectiveness of casirivimab/imdevimab in patients with COVID-19 in the ambulatory setting. *J. Manag. Care & Specialty Pharm.* 28 (5), 555–565. doi:10.18553/jmcp.2022.21469

Kelton, K., Klein, T., Murphy, D., Belger, M., Hille, E., McCollam, P. L., et al. (2022). Cost-effectiveness of combination of baricitinib and remdesivir in hospitalized patients with COVID-19 in the United States: a modelling study. *Adv. Ther.* 39 (1), 562–582. doi:10.1007/s12325-021-01982-6

Kilcoyne, A., Jordan, E., Thomas, K., Pepper, A. N., Zhou, A., Chappell, D., et al. (2022a). Clinical and economic benefits of lenzilumab plus standard of care compared with standard of care alone for the treatment of hospitalized patients with coronavirus disease 19 (COVID-19) from the perspective of national health service england. *Clin. Outcomes Res.* 14, 231–247. doi:10.2147/CEOR.S360741

Kilcoyne, A., Jordan, E., Zhou, A., Thomas, K., Pepper, A. N., Chappell, D., et al. (2022b). Clinical and economic benefits of lenzilumab plus standard of care compared with standard of care alone for the treatment of hospitalized patients with COVID-19 in the United States from the hospital perspective. *J. Med. Econ.* 25 (1), 160–171. doi:10.1080/13696998.2022.2030148

Kowal, S., Ng, C. D., Schuldt, R., Sheinson, D., and Cookson, R. (2023). The impact of funding inpatient treatments for COVID-19 on health equity in the United States: a distributional cost-effectiveness analysis. *Value Health* 26 (2), 216–225. doi:10.1016/j.jval.2022.08.010

Krylova, O., Krasheninnikov, A., Mamontova, E., Tananakina, G., and Belyakova, D. (2021). Pharmacoeconomic analysis of treatment regimens for coronavirus infection coronavirus disease-19. *Open Access Macedonian J. Med. Sci.* 9 (E), 1182–1189. doi:10.3889/oamjms.2021.7015

Lau, V. I., Fowler, R., Pinto, R., Tremblay, A., Borgia, S., Carrier, F. M., et al. (2022). Cost-effectiveness of remdesivir plus usual care versus usual care alone for hospitalized patients with COVID-19: an economic evaluation as part of the Canadian Treatments for COVID-19 (CATCO) randomized clinical trial. *CMAJ Open* 10 (3), E807–E817. doi:10.9778/cmajo.20220077

Mathieu, E., Ritchie, H., Guirao-Rodés, L., Appel, C., Giattino, C., Hassell, J., et al. (2020). *Coronavirus pandemic (COVID-19)*.

metaEvidence Initiative (2022). *Living meta-analysis and evidence synthesis of therapies for COVID19*.

Metry, A., Pandor, A., Ren, S., Shippam, A., Clowes, M., Dark, P., et al. (2022). *Therapeutics for people with COVID-19 [ID4038]*. Assessment report, 2022.

National Institute for Health and Care Excellence (NICE) (2012). *The guidelines manual—appendix G: methodology checklist: economic evaluations (Section 2: study limitations)*.

Ohsfeldt, R., Kelton, K., Klein, T., Belger, M., Mc Collam, P. L., Spiro, T., et al. (2021). Cost-effectiveness of baricitinib compared with standard of care: a modeling study in hospitalized patients with COVID-19 in the United States. *Clin. Ther.* 43 (11), 1877–1893.e4. doi:10.1016/j.clinthera.2021.09.016

Oksuz, E., Malhan, S., Gonen, M. S., Kutlubay, Z., Keskindemirci, Y., Jarrett, J., et al. (2021). Cost-effectiveness analysis of remdesivir treatment in COVID-19 patients requiring low-flow oxygen therapy: payer perspective in Turkey. *Adv. Ther.* 38 (9), 4935–4948. doi:10.1007/s12325-021-01874-9

Park, M., Tan, K. B., Vasoo, S., Dickens, B. L., Lye, D., and Cook, A. R. (2022). Estimated health outcomes and costs associated with use of monoclonal antibodies for prevention or mitigation of SARS-CoV-2 infections. *JAMA Netw. Open* 5 (4), e225750. doi:10.1001/jamanetworkopen.2022.5750

Petrov, V. I., Ryazanova, N. Y., Ponomareva, A. V., Shatalova, O. V., and Levina, Y. V. (2022). CLINICAL AND ECONOMIC ANALYSIS OF GENETICALLY ENGINEERED BIOLOGICS CONSUMPTION BY PATIENTS WITH COVID-19. *Pharm. Pharmacol.* 10 (2), 198–206. doi:10.19163/2307-9266-2022-10-2-198-206

Rafia, R., Martyn-St James, M., Harnan, S., Metry, A., Hamilton, J., and Wailoo, A. (2022). A cost-effectiveness analysis of remdesivir for the treatment of hospitalized patients with COVID-19 in england and wales. *Value Health* 25 (5), 761–769. doi:10.1016/j.jval.2021.12.015

Ruggeri, M., Signorini, A., and Caravaggio, S. (2023). Casirivimab and imdevimab: costeffectiveness analysis of the treatment based on monoclonal antibodies on outpatients with Covid-19. *PLOS ONE* 18 (2), e0279022. doi:10.1371/journal.pone.0279022

Ruggeri, M., Signorini, A., Caravaggio, S., Alraddadi, B., Alali, A., Jarrett, J., et al. (2022b). Modeling the potential impact of remdesivir treatment for hospitalized patients with COVID-19 in Saudi Arabia on healthcare resource use and direct hospital costs: a hypothetical study. *Clin. Drug Investig.* 42, 669–678. doi:10.1007/s40261-022-01177-z

Ruggeri, M., Signorini, A., Caravaggio, S., Rua, J., Luís, N., Braz, S., et al. (2022a). Estimation model for healthcare costs and intensive care units access for covid-19 patients and evaluation of the effects of remdesivir in the Portuguese context: hypothetical study. *Clin. Drug Investig.* 42 (4), 345–354. doi:10.1007/s40261-022-01128-8

Savinkina, A., Paltiel, A. D., Ross, J. S., and Gonsalves, G. (2022). Population-level strategies for nirmatrelvir/ritonavir prescribing—a cost-effectiveness analysis. *Open Forum Infect. Dis.* 9 (12), ofac637. doi:10.1093/ofid/ofac637

Schallner, N., Lieberum, J., Kalbhenn, J., Bürkle, H., and Daumann, F. (2022). Intensive care unit resources and patient-centred outcomes in severe COVID -19: a prospective single-centre economic evaluation. *Anaesthesia* 77 (12), 1336–1345. doi:10.1111/anae.15844

Shah, H. A., Baker, T., Schell, C. O., Kuwawenaruwa, A., Awadh, K., Khalid, K., et al. (2023). Cost effectiveness of strategies for caring for critically ill patients with COVID-19 in Tanzania. *PharmacoEconomics - Open* 7 (4), 537–552. doi:10.1007/s41669-023-00418-x

Sheinson, D., Dang, J., Shah, A., Meng, Y., Elsea, D., and Kowal, S. (2021). A cost-effectiveness framework for COVID-19 treatments for hospitalized patients in the United States. *Adv. Ther.* 38 (4), 1811–1831. doi:10.1007/s12325-021-01654-5

Shields, G., and Elvidge, J. (2020). Challenges in synthesising cost-effectiveness estimates. *Syst. Rev.* 9 (1), 289. doi:10.1186/s13643-020-01536-x

Subhi, A., Shamy, A. M. E., Hussein, S. A. M., Jarrett, J., Kozma, S., Harfouche, C., et al. (2023). Use of anti-viral therapies in hospitalised COVID-19 patients in the United Arab Emirates: a cost-effectiveness and health-care resource use analysis. *BMC Health Serv. Res.* 23 (1), 383. doi:10.1186/s12913-023-09376-w

The COVID-NMA Initiative (2021). *A living mapping and living systematic review of Covid-19 trials*.

Wai, A. K.-C., Chan, C. Y., Cheung, A. W.-L., Wang, K., Chan, S. C.-L., Lee, T. T.-L., et al. (2023). Association of Molnupiravir and Nirmatrelvir-Ritonavir with preventable mortality, hospital admissions and related avoidable healthcare system cost among high-risk patients with mild to moderate COVID-19. *Lancet Regional Health - West. Pac.* 30, 100602. doi:10.1016/j.lanwpc.2022.100602

Yeung, K., Whittington, M., Beinfeld, M., Mohammed, R., Wright, A., Nhan, E., et al. (2022). *Special assessment of outpatient treatments for COVID-19; evidence report*, 2022.

frontiers | Frontiers in Pharmacology

TYPE Methods
PUBLISHED 24 November 2023
DOI 10.3389/fphar.2023.1249998

Check for updates

OPEN ACCESS

EDITED BY
Blythe Adamson,
Flatiron Health, United States

REVIEWED BY
Osman N Yogurtcu,
United States Food and Drug Administration, United States
Paul Zarogoulidis,
Euromedica General Clinic, Greece

*CORRESPONDENCE
Koen Degeling,
info@koendegeling.nl

RECEIVED 29 June 2023
ACCEPTED 10 November 2023
PUBLISHED 24 November 2023

CITATION
Zebrowska K, Banuelos RC, Rizzo EJ, Belk KW, Schneider G and Degeling K (2023), Quantifying the impact of novel metastatic cancer therapies on health inequalities in survival outcomes. *Front. Pharmacol.* 14:1249998. doi: 10.3389/fphar.2023.1249998

COPYRIGHT
© 2023 Zebrowska, Banuelos, Rizzo, Belk, Schneider and Degeling. This is an open-access article distributed under the terms of the Creative Commons Attribution License (CC BY). The use, distribution or reproduction in other forums is permitted, provided the original author(s) and the copyright owner(s) are credited and that the original publication in this journal is cited, in accordance with accepted academic practice. No use, distribution or reproduction is permitted which does not comply with these terms.

Quantifying the impact of novel metastatic cancer therapies on health inequalities in survival outcomes

Karolina Zebrowska[1], Rosa C. Banuelos[2], Evelyn J. Rizzo[2], Kathy W. Belk[2], Gary Schneider[2] and Koen Degeling[1]*

[1]Healthcare Consultancy Group, London, United Kingdom, [2]Healthcare Consultancy Group, New York, NY, United States

Background: Novel therapies in metastatic cancers have contributed to improvements in survival outcomes, yet real-world data suggest that improvements may be mainly driven by those patient groups who already had the highest survival outcomes. This study aimed to develop and apply a framework for quantifying the impact of novel metastatic cancer therapies on health inequalities in survival outcomes based on published aggregate data.

Methods: Nine (N = 9) novel therapies for metastatic breast cancer (mBC), metastatic colorectal cancer (mCRC), and metastatic non–small cell lung cancer (mNSCLC) were identified, 3 for each cancer type. Individual patient data (IPD) for overall survival (OS) and progression-free survival (PFS) were replicated from published Kaplan-Meier (KM) curves. For each cancer type, data were pooled for the novel therapies and comparators separately and weighted based on sample size to ensure equal contribution of each therapy in the analyses. Parametric (mixture) distributions were fitted to the weighted data to model and extrapolate survival. The inequality in survival was defined by the absolute difference between groups with the highest and lowest survival for 2 stratifications: one for which survival was stratified into 2 groups and one using 5 groups. Additionally, a linear regression model was fitted to survival estimates for the 5 groups, with the regression coefficient or slope considered as the inequality gradient (IG). The impact of the pooled novel therapies was subsequently defined as the change in survival inequality relative to the pooled comparator therapies. A probabilistic analysis was performed to quantify parameter uncertainty.

Results: The analyses found that novel therapies were associated with significant increases in inequalities in survival outcomes relative to their comparators, except in terms of OS for mNSCLC. For mBC, the inequalities in OS increased by 13.9 (95% CI: 1.4; 26.6) months, or 25.0%, if OS was stratified in 5 groups. The IG for mBC increased by 3.2 (0.3; 6.1) months, or 24.7%. For mCRC, inequalities increased by 6.7 (3.0; 10.5) months, or 40.4%, for stratification based on 5 groups; the IG increased by 1.6 (0.7; 2.4) months, or 40.2%. For mNSCLC, inequalities decreased by 14.9 (–84.5; 19.0) months, or 12.2%, for the 5-group stratification; the IG decreased by 2.0 (–16.1; 5.1) months, or 5.5%. Results for the stratification based on 2 groups demonstrated significant increases in OS inequality for all cancer types. In terms of PFS, the increases in survival inequalities were larger in a relative sense compared with OS. For mBC, PFS inequalities increased by 8.7 (5.9; 11.6) months,

or 71.7%, for stratification based on 5 groups; the IG increased by 2.0 (1.3; 2.6) months, or 67.6%. For mCRC, PFS inequalities increased by 5.4 (4.2; 6.6) months, or 147.6%, for the same stratification. The IG increased by 1.3 (1.1; 1.6) months, or 172.7%. For mNSCLC, inequalities increased by 18.2 (12.5; 24.4) months, or 93.8%, for the 5-group stratification; the IG increased by 4.0 (2.8; 5.4) months, or 88.1%. Results from the stratification based on 2 groups were similar.

Conclusion: Novel therapies for mBC, mCRC, and mNSCLC are generally associated with significant increases in survival inequalities relative to their comparators in randomized controlled trials, though inequalities in OS for mNSCLC decreased nonsignificantly when stratified based on 5 groups. Although further research using real-world IPD is warranted to assess how, for example, social determinants of health affect the impact of therapies on health inequalities among patient groups, the proposed framework can provide important insights in the absence of such data.

KEYWORDS

oncology, inequality, overall survival, progression-free survival, health disparities, colorectal cancer, non-small cell lung cancer, breast cancer

Introduction

Disparities and inequalities in cancer survival outcomes exist, and they have been well-documented in equity-informed literature. Studies that examine survival disparities in patients undergoing oncology care have found that treatment improved overall survival (OS); however, social determinants of health (SDOH), such as Black race, low income, lack of insurance, and low educational attainment, have been associated with poorer OS outcomes (Acharya et al., 2016; Austin et al., 2016; Cui and Finkelstein, 2022; Fabregas et al., 2022; Lee and Singh, 2022; Namburi et al., 2022; Tran et al., 2022; Alnajar et al., 2023). For example, one study found that the percentage of individuals with survival <1 year after diagnosis in Black individuals and White individuals was 41.4% and 22.2% for lung cancer, 9.8% and 7.1% for colorectal cancer, and 2.9% and 0.7% for breast cancer, respectively (Cui and Finkelstein, 2022). Another study showed that patients with advanced lung cancer living in the most materially deprived areas had the shortest median survival time (Qureshi et al., 2023). Hamers et al. (2020) found that, out of all patients diagnosed with stage IV colorectal cancer between 2008 and 2016 in the Netherlands Cancer Registry, OS improved only for those patients who were already doing well compared with others. Further, Asaria et al. (2015) demonstrated that, compared with no screening, a UK bowel cancer screening program improved health across the distribution but widened health inequality between the healthiest and least healthy participants.

There are several methodologic approaches to quantifying inequalities within healthcare from a health economic perspective. These include distributional cost-effectiveness analysis, extended cost-effectiveness analysis, equity-based weighting, multiple criteria decision analysis, and mathematical programming (Ward et al., 2022). A challenge in the use of these methods is that they are mostly informed by granular patient-level data on the relationship between SDOH and health outcomes. SDOH operate at individual, community, and population levels to impact health outcomes (Sengupta and Honey, 2020) and include but are not limited to socioeconomic factors, clinical factors, behavioral factors, environmental factors, and biological factors (American Association for Cancer Research, 2022). However, such data are often not available, which limits the feasibility of performing these types of equity-informed health economic analyses.

To facilitate equity-informed analyses in the absence of individual patient data (IPD), this study aimed to develop and apply a framework for quantifying the impact of novel metastatic cancer therapies on health inequalities in survival outcomes based on aggregate data. The framework was applied to estimate the impact of novel therapies on OS and progression-free survival (PFS) outcomes in metastatic breast cancer (mBC), metastatic colorectal cancer (mCRC), and metastatic non–small cell lung cancer (mNSCLC).

Materials and methods

Framework

The proposed framework defines the distribution of health in terms of survival of the different patient groups that can be stratified in Kaplan-Meier (KM) curves. This analysis focuses on 2 stratifications: a distribution based on 2 groups and a distribution based on 5 groups of survival. Although the number of groups in which survival will be stratified is a somewhat arbitrary choice, the 5-group stratification was used here because this number of groups is often used to define distributions across populations, for example, based on socioeconomic quintiles (Cookson et al., 2017). The 2-group stratification was additionally applied to investigate and demonstrate that results may change when a different number of groups is used, and to illustrate that even this most basic stratification can provide meaningful insights. Figure 1A illustrates the stratification based on 5 groups. The distribution of health can subsequently be obtained from the median survival within each group, as illustrated in Figure 1B, C. Given that most survival data are censored, this step may involve parametric survival modeling to extrapolate survival curves.

For both the 2-group and 5-group stratifications, the inequality in survival for a certain treatment was defined by the absolute

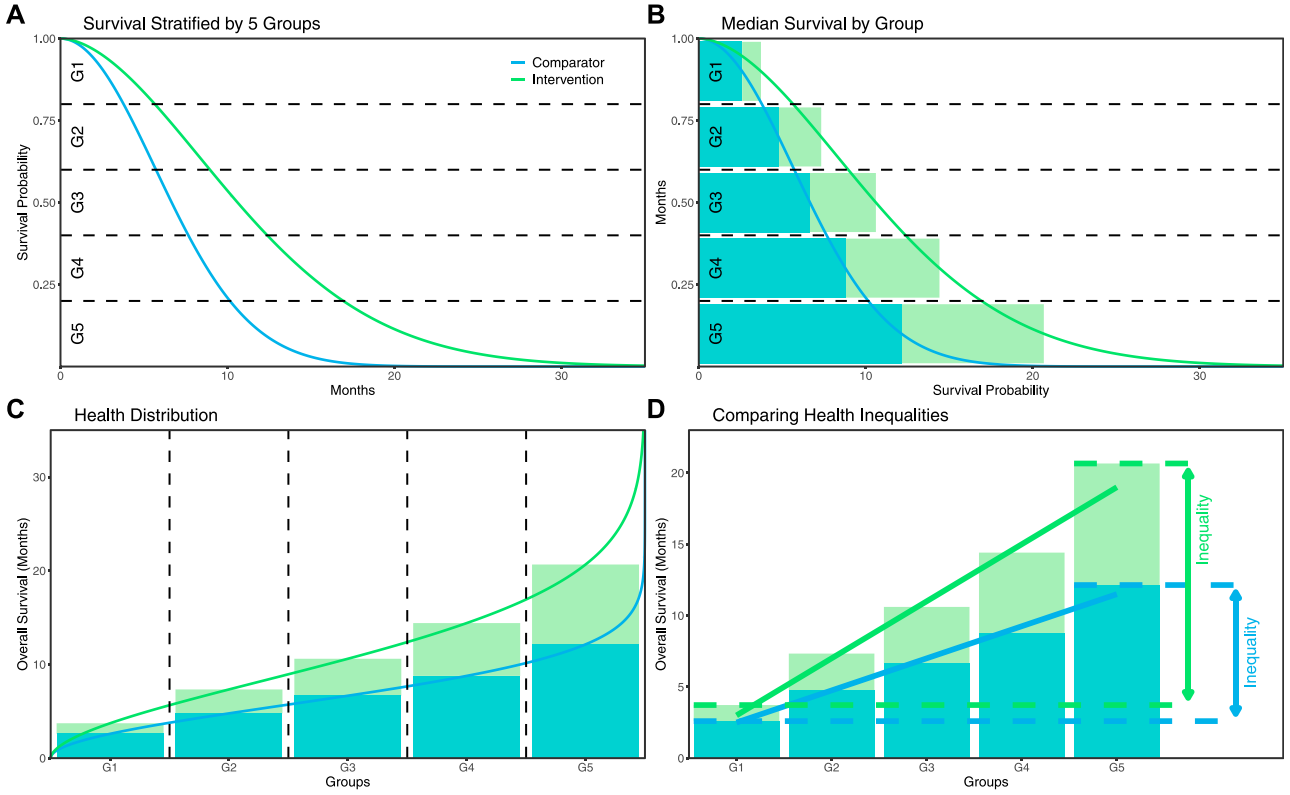

FIGURE 1
(A) illustration of the definition of the patient groups for the 5-group stratification of survival; **(B)** illustration of the median survival in each group; **(C)** illustration of the resulting health distribution based on the median survival in each group; **(D)** illustration of the definition of the health inequalities.

TABLE 1 Intervention and comparator combinations included in the analysis.

Novel therapy (intervention)	Comparator	Clinical trial	References
mBC			
Neratinib + capecitabine	Lapatinib + capecitabine	NALA	Saura et al. (2020)
Tucatinib, trastuzumab, and capecitabine	Placebo, trastuzumab, and capecitabine	HER2CLIMB	Curigliano et al. (2022)
Margetuximab + chemotherapy	Trastuzumab + chemotherapy	SOPHIA	Rugo et al. (2021)
mCRC			
Encorafenib + cetuximab + binimetinib	Investigator's choice - either cetuximab + irinotecan or cetuximab + FOLFIRI (control)	BEACON	Tabernero et al. (2021)
Regorafenib	Placebo	CORRECT	Grothey et al. (2013)
Trifluridine/tipiracil	Placebo	TERRA	Xu et al. (2018)
mNSCLC			
Osimertinib[a]	First-generation or second-generation EGFR-TKI	FLAURA	PFS: (Soria et al., 2018)
			OS: (Ramalingam et al., 2020)
Nivolumab	Docetaxel	CHECKMATE 078	Wu et al. (2019)
Pembrolizumab[b] + ipilimumab	Chemotherapy	KEYNOTE-042	Mok et al. (2019)

[a]Although Soria et al. (2018) included OS, data, the follow-up period was not sufficient (35% survival not reached), hence Ramalingam et al. (2020) was used.
[b]PD-L1 TPS, of >50%.

difference in survival between the highest and lowest groups. This is shown visually in Figure 1D. To consider the health distribution across all patient groups for the 5-group stratification, the survival inequality was additionally defined based on the regression slope of a simple linear regression model fitted to the outcomes of all 5 groups, referred to as the inequality gradient (IG).

The impact of novel therapies on survival inequalities can then be defined by the absolute and relative change in the survival inequality (i.e., absolute difference in survival between the lowest and highest groups and the IG) relative to their comparators.

Application of the framework

To quantify the impact of novel metastatic cancer therapies on health inequalities through the above framework, it was applied to novel therapies for mBC, mCRC, and mNSCLC. These metastatic cancer types were selected based on their incidence and the availability of novel therapies that met the inclusion criteria. For each cancer type, 3 novel drugs were identified based on the following 5 criteria: a) US Food and Drug Administration drug approval between January 2015 and January 2023; b) availability of results from a Phase III randomized controlled trial (RCT); c) at least 100 patients in each arm of the RCT; d) published KM curves for OS and PFS; and e) sufficient follow-up such that the OS and PFS were lower than 35% at the end of follow-up, to reduce the impact of structural assumptions in any survival extrapolations. Table 1 provides an overview of the novel therapies that were selected and their comparators.

For each treatment, IPD for OS and PFS were replicated from the KM curves and summary statistics using the method by Guyot et al. (2012). Beyond visual inspection, the replication process was validated by analyzing the replicated IPD and comparing the results with those reported in the corresponding publications. Subsequently, for each cancer type separately, the IPD for the 3 novel therapies were pooled, as were the data for the 3 comparators, with weighting applied based on the corresponding sample sizes such that each therapy contributed equally to the analysis. Although the framework can be applied to evaluate the impact of specific drugs, data of multiple interventions were pooled because the purpose of this work was to illustrate the proposed framework and not to perform such head-to-head comparisons. The studies used in this analysis used a common criterion, namely, the RECIST v1.1, for OS and PFS definitions. This allowed for straightforward aggregation of individual studies. It must be noted that this is not always the case, and caution must be exercised when pooling data derived from studies that use different criteria for survival outcomes.

Parametric survival modeling was performed to obtain the complete survival distributions for the pooled sets of novel therapies and comparators. Standard parametric distributions and mixtures of 2 distributions were explored, considering the following distributions: exponential, Gamma, Gompertz, log-logistic, log-normal, and Weibull (Gray et al., 2021). Relative modeling of the treatment effects, for example, through parameterization of the distributions' scale/rate parameter as hazard ratio, was not considered because that would result in increased survival inequalities by definition. More specifically, applying a single relative effect for the interventions compared with the comparators will result in larger absolute change for groups with a higher baseline and, hence, increase inequalities. To reduce the

TABLE 2 Survival inequalities in terms of OS for all cancers, reported as mean (95% confidence interval) based on the probabilistic analysis.

Cancer type	2-Group stratification			5-Group stratification			Inequality gradient		
	Interventions	Comparators	Difference[a]	Interventions	Comparators	Difference[a]	Interventions	Comparators	Difference[a]
mBC	29.6 (26.3; 33.2)	24.2 (21.4; 27.2)	5.4 (0.9; 9.9)	71.1 (61.9; 81.5)	57.1 (49.7; 65.5)	13.9 (1.4; 26.6)	16.4 (14.4; 18.8)	13.2 (11.6; 15.1)	3.2 (0.3; 6.1)
			23% (3.6%; 44.6%)			25% (2.2%; 50.8%)			24.7% (2.4%; 49.7%)
mCRC	10.1 (9.1; 11.1)	7.3 (6.4; 8.2)	2.8 (1.4; 4.2)	23.7 (21.0; 26.7)	17.0 (14.7; 19.5)	6.7 (3.0; 10.5)	5.5 (4.9; 6.2)	3.9 (3.4; 4.5)	1.6 (0.7; 2.4)
			38.9% (18.0%; 62.3%)			40.4% (16.1%; 67.9%)			40.4% (16.1%; 67.1%)
mNSCLC	39.9 (35.9; 44.9)	30.8 (23.6; 38.1)	9.1 (0.7; 17.4)	69.3 (58.3; 89.3)	84.2 (56.1; 153.3)	−14.9 (−84.5; 19.0)	17.1 (14.7; 21.3)	19.0 (13.4; 33.1)	−2.0 (−16.1; 5.1)
			31.4% (1.8%; 71.1%)			−12.2% (−55.8%; 31.1%)			−5.5% (−49.2%; 35.1%)

[a]Confidence intervals for the Difference that are strictly positive or negative, i.e., that do not cover zero, suggest the difference is significant.

TABLE 3 Survival inequalities in terms of PFS for all cancers, reported as mean (95% confidence interval) based on the probabilistic analysis.

Cancer type	2 group stratification			5 group stratification			Inequality gradient		
	Interventions	Comparators	Difference[a]	Interventions	Comparators	Difference[a]	Interventions	Comparators	Difference[a]
mBC	8.8 (8.0; 9.7)	5.5 (5.0; 6.1)	3.3 (2.3; 4.3)	21.4 (19.0; 23.9)	12.7 (11.3; 14.2)	8.7 (5.9; 11.6)	4.9 (4.4; 5.5)	3.0 (2.6; 3.3)	2.0 (1.3; 2.6)
			60.7% (39.4%; 84%)			71.7% (48.7%; 95.8%)			67.6% (42.8%; 95.2%)
mCRC	4.1 (3.7; 4.5)	0.8 (0.6; 1.1)	3.3 (2.8; 3.8)	9.1 (8.1; 10.1)	3.7 (3.1; 4.4)	5.4 (4.2; 6.6)	2.1 (1.9; 2.4)	0.8 (0.7; 0.9)	1.3 (1.1; 1.6)
			410.6% (277.4%; 563.2%)			147.6% (100.7%; 203.2%)			172.7% (121.5%; 233.3%)
mNSCLC	14.9 (13.2; 16.8)	9.2 (8.4; 10.0)	5.7 (3.8; 7.7)	37.8 (32.6; 43.6)	19.6 (17.6; 21.9)	18.2 (12.5; 24.4)	8.7 (7.5; 10.0)	4.6 (4.2; 5.1)	4.0 (2.8; 5.4)
			62.4% (39.4%; 87.7%)			93.8% (60.2%; 131.0%)			88.1% (56.7%; 122.7%)

[a]Confidence intervals for the Difference that are strictly positive or negative, i.e., that do not cover zero, suggest the difference is significant.

potential impact of structural uncertainty on the outcomes, the same type of distribution was used for the pooled novel therapies and their comparators for each cancer-outcome combination. As the selection of an inappropriate survival model can strongly bias survival estimates and lead to inaccurate results (Gray et al., 2021), an algorithm was defined to select the survival distribution used in the analyses. First, 10-year relative survival rates from the Surveillance, Epidemiology, and End Results (SEER) database were used to define an upper threshold for survival extrapolations (14.8% for mBC (National Cancer Institute Surveillance, Epidemiology, and End Results Program, 2022a), 10% for mCRC (National Cancer Institute Surveillance, Epidemiology, and End Results Program, 2022b), and 3.3% for mNSCLC (National Cancer Institute Surveillance, Epidemiology, and End Results Program, 2022c)). This study allowed for a 10% relative increase of these survival rates to account for novel therapies that may increase survival compared with the treatments used during the SEER data-capture period. Second, the survival distributions for which both the pooled novel therapies and comparators did not exceed the extrapolation threshold were ranked based on their combined Akaike information criterion (AIC) for each distribution type, where a lower AIC indicated a better fit. Finally, the survival distribution with the lowest AIC was selected after a visual inspection to ensure that it was realistic and did not substantially underestimate survival, for example. See the Supplementary Material for an illustration of the selection algorithm and the results of its application for the different cancer types and outcomes.

Analyses and availability of material

All results were generated through a probabilistic analysis to quantify the impact of parameter uncertainty on the outcomes and the uncertainty around those outcomes. Multivariate normal distributions were used to define the uncertainty in the survival model parameters. All analyses were performed in R version 4.2.2, and a simplified example of the code used in this analysis has been made available in the following GitHub repository: https://github.com/koendegeling/Survival_Inequalities. The flexsurv package, version 2.2.1 (Jackson, 2016), was used for standard parametric survival modeling.

Results

Inequalities in OS and PFS significantly increased when comparing the combined novel therapies with their comparators in RCTs, except for mNSCLC, where there was a nonsignificant decrease in OS inequality for the 5-group stratification. The full results are presented by outcome in the following 2 subsections. Tables 2, 3 show the full results of OS and PFS, respectively.

Overall survival

For mBC, Figure 2A illustrates the survival extrapolation using the selected log-logistic distribution, as well as the health

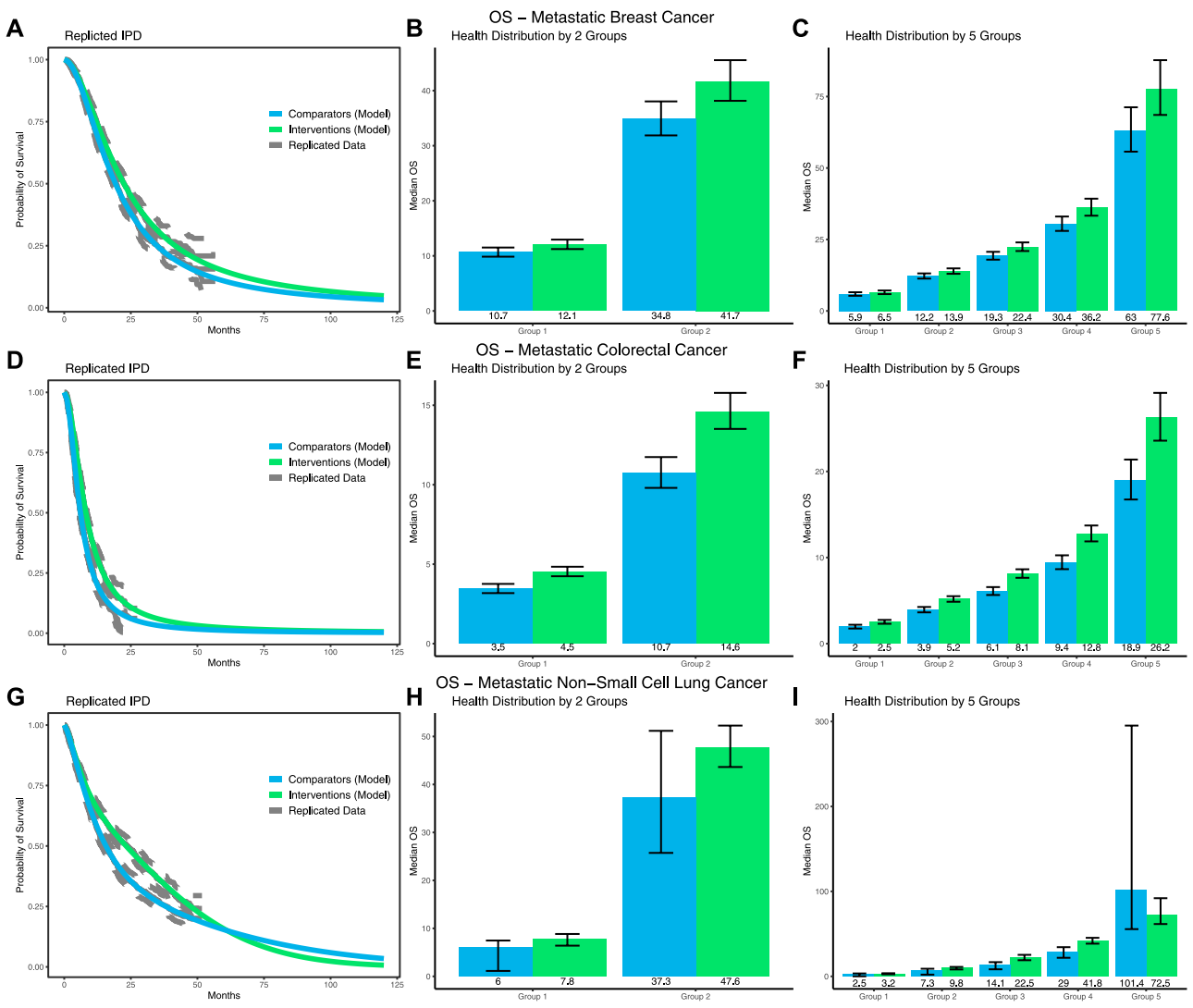

FIGURE 2
Results of the survival modeling for OS and the corresponding health distributions in terms of median OS based on the 2- and 5-group stratifications for mBC **(A–C)**, mCRC **(D–F)**, and mNSCLC **(G–I)**.

distributions based on the 2- and 5-group stratifications. Detailed results for the survival inequalities and the differences therein are presented in Table 2. The highest increase in survival inequalities was observed for the 5-group stratification. Here, the inequality in OS increased by 13.9 (95% CI: 1.4; 26.6) months, or 25% (2.2%; 50.8%), from 57.1 (49.7; 65.5) months to 71.1 (61.9; 81.5) months.

Survival was extrapolated using a log-logistic distribution for mCRC (Figure 2D). The greatest increase in inequalities was seen in the 5-group stratification, where the inequality in OS increased by 6.7 (3.0; 10.5) months, or 40.4% (16.1%; 67.9%), from 17.0 (14.7; 19.5) months to 23.7 (21.0; 26.7) months.

For mNSCLC, survival was extrapolated using a mixture of a Gamma and a Weibull distribution (Figure 2G). The results for the 2-group stratification show an increase in OS inequality by 9.1 (0.7; 17.4) months, or 31.4% (1.8%; 71.1%), from 30.8 (23.6; 38.1) months to 39.9 (35.9; 44.9) months. Notably, however, the results for the 5-group stratification showed a nonsignificant decrease in inequalities by 14.9 (−84.5; 19.0) months, or 12.2% (−55.8%; 31.1%), from 84.2 (56.1; 153.3) months to 69.3 (58.3; 89.3) months, which is the result of the crossing of the survival curve.

Progression-free survival

For all cancer types, PFS was extrapolated using a mixture of a log-logistic and log-normal distribution (Figure 3A, D, G), showing significant increases in survival inequalities. For mBC (Figure 3B, C), the greatest increase in inequalities was seen in the 5-group stratification, where the inequality in PFS increased by 8.7 (5.9; 11.6) months, or 71.7% (48.7%; 95.8%), from 12.7 (11.3; 14.2) months to 21.4 (19.0; 23.9) months.

For mCRC (Figure 3E, F), the 5-group stratification again showed the greatest increase in inequalities in absolute sense. The inequality in PFS increased by 5.4 (4.2; 6.6) months, or 147.6% (100.7%; 203.2%), from 3.7 (3.1; 4.4) months to 9.1 (8.1; 10.1) months. Note that the increase was higher in relative sense for the

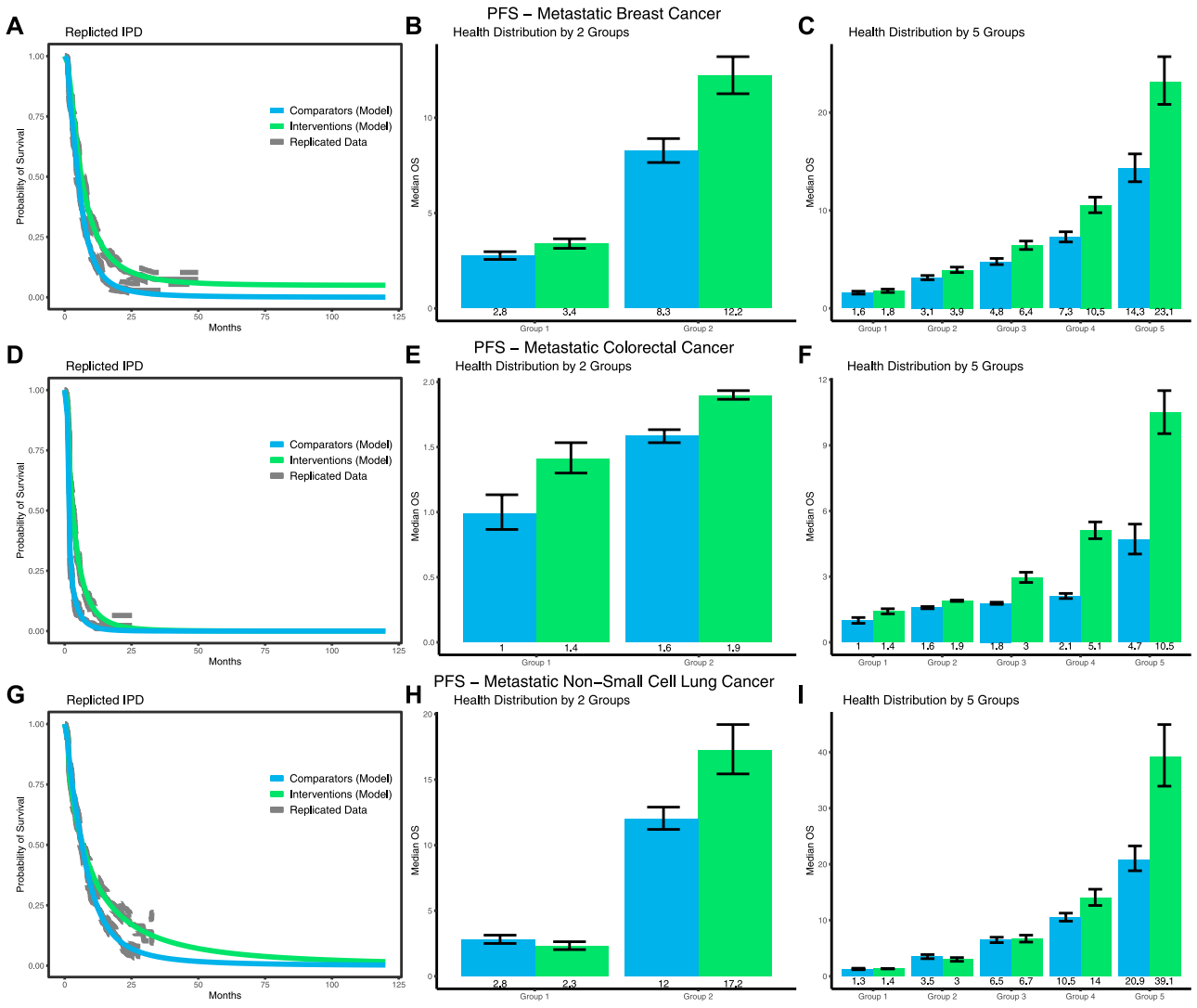

FIGURE 3
Results of the survival modeling for PFS and the corresponding health distributions in terms of median PFS based on the 2- and 5-group stratification for mBC **(A–C)**, mCRC **(D–F)**, and mNSCLC **(G–I)**.

2-group stratification, but this was caused by the low value for the comparator group as denominator.

For mNSCLC (Figure 3H, I), the greatest increase in inequalities was seen in the 5-group stratification, where the inequality in PFS increased by 18.2 (12.5; 24.4) months, or 93.8% (60.2%; 131.0%), from 19.6 (17.6; 21.9) months to 37.8 (32.6; 43.6) months.

Discussion

In this research, a framework was proposed to quantify the impact of novel metastatic cancer therapies on health inequalities in survival outcomes based on published KM curves. This comes at a pivotal point in time, where there is increasing debate about how to consider equity-related aspects in health economic analyses. For example, there has been a collective effort to show that lack of health equity consideration within a health technology assessment (HTA) could result in neglecting an important aspect of the value of interventions and potentially misallocation of healthcare resources (Cookson et al., 2017; Podolsky et al., 2022). Furthermore, the Institute for Clinical and Economic Review has recently published a whitepaper on the use of methods that support equity-informed analyses for HTA in the United States (Agboola et al., 2023). It has also been suggested that even the most popular method, namely, distributional cost-effectiveness analysis, faces significant challenges in implementation by HTA agencies due to scarcity and lack of consistency within equity-informed data (Meunier et al., 2023).

The framework was successfully applied to estimate the impact of novel therapies on OS and PFS outcomes in mBC, mCRC, and mNSCLC. Overall, the results of this analysis showed that the pooled novel therapies improved median survival for OS and PFS but widened survival inequalities in absolute terms by increasing survival the most among those patient groups who had comparatively better survival outcomes already. The findings for mNSCLC in terms of OS showed that the framework is also capable

of identifying decreases in inequalities, albeit nonsignificant for this case study. Hypotheses on what may explain this finding are beyond the scope of this research.

Although the framework was applied to pooled therapies for certain metastatic cancers, it can be generalized to other settings as well. For instance, it can be applied to specific treatments to investigate the impact of certain therapies on inequalities. It can also be applied to other types and stages of cancer, other disease areas and treatments, and other time-to-event outcomes. The philosophy behind the framework can also be used as a foundation for exploring the quantification of health inequalities based on published distributions for other types of outcomes.

In addition to its potential broad applicability, strengths of the framework include that it is a conceptually straightforward approach to visualize and explain, and it is relatively easy to apply, with the provided R code further contributing to uptake and use by other researchers. Therefore, it represents a potentially important tool that can provide useful insights when IPD are not available, facilitating an initial understanding of how an intervention may impact healthcare disparities and informing further IPD-driven research into health disparities.

The main limitation of the proposed framework is that it does not provide any direct information as to why the changes in the health distribution occur. Although various organizations have published slightly different versions, definitions generally consider inequities or disparities as unjust differences in outcomes that can be explained by SDOH, whereas inequalities are used as a synonym for inequities or to simply describe that there are differences in outcomes (Braveman, 2014; Arcaya et al., 2015). Here, we adopt the latter definition of inequalities, and therefore, one could say that the framework provides insights into the impact on health inequalities but not disparities, which would require explanation of the changes based on SDOH. Results obtained through the framework could, hence, be complemented with disease-specific evidence on links between health inequalities and SDOH or, ideally, analyses of RCT data or real-world data to understand the impact of SDOH on the outcomes. Nevertheless, the proposed approach using aggregate data provides useful initial insights into how healthcare interventions may impact the distribution of health outcomes between groups of individuals.

A natural extension of this work would be to use the results in, for example, a distributional cost-effectiveness analysis. Further research is also warranted to apply the framework to more case studies within oncology and beyond. Finally, it would be particularly interesting to apply the framework to a case study for which the corresponding IPD are available to compare the results and to investigate the extent to which the impact on inequalities can be explained by SDOH—to assess the link between health inequalities and disparities.

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

Conceptualization: KB, GS, and KD. Screening studies: KZ and ER. Data generation: KZ and RB. Data analysis: KZ, RB, and KD. Manuscript writing: KZ, ER, and KD. Review of manuscript: RB, ER, KB, GS, and KD. All authors contributed to the article and approved the submitted version.

Conflict of interest

All authors were employed by Healthcare Consultancy Group. No funding was received for performing this research and 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.

Supplementary material

The Supplementary Material for this article can be found online at: https://www.frontiersin.org/articles/10.3389/fphar.2023.1249998/full#supplementary-material

References

Acharya, S., Hsieh, S., Shinohara, E. T., DeWees, T., Frangoul, H., and Perkins, S. M. (2016). Effects of race/ethnicity and socioeconomic status on outcome in childhood acute lymphoblastic leukemia. *J. Pediatr. Hematol. Oncol.* 38, 350–354. doi:10.1097/MPH.0000000000000591

Agboola, F., Whittington, M. D., and Pearson, S. D. (2023). *Advancing health technology assessment methods that support health equity*. Boston, MA, United States: Institute for Clinical and Economic Review. Available at: https://icer.org/wp-content/uploads/2022/07/ICER_Advancing-Health-Technology-Assessment-Methods-that-Support-Health-Equity_040523.pdf (Accessed May 23, 2023).

Alnajar, A., Kareff, S. A., Razi, S. S., Rao, J. S., De Lima Lopes, G., Nguyen, D. M., et al. (2023). Disparities in survival due to social determinants of health and access to treatment in US patients with operable malignant pleural mesothelioma. *JAMA Netw. Open.* 6, e234261. doi:10.1001/jamanetworkopen.2023.4261

American Association for Cancer Research (2022). *AACR cancer disparities progress report 2022*. Available at: http://www.CancerDisparitiesProgressReport.org/(Accessed May 23, 2023).

Arcaya, M. C., Arcaya, A. L., and Subramanian, S. V. (2015). Inequalities in health: definitions, concepts, and theories. *Glob. Health Action.* 8, 27106. doi:10.3402/gha.v8.27106

Asaria, M., Griffin, S., Cookson, R., Whyte, S., and Tappenden, P. (2015). Distributional cost-effectiveness analysis of health care programmes--a methodological case study of the UK Bowel Cancer Screening Programme. *Health Econ.* 24, 742–754. doi:10.1002/hec.3058

Austin, M. T., Hamilton, E., Zebda, D., Nguyen, H., Eberth, J. M., Chang, Y., et al. (2016). Health disparities and impact on outcomes in children with primary central nervous system solid tumors. *J. Neurosurg. Pediatr.* 18, 585–593. doi:10.3171/2016.5.PEDS15704

Braveman, P. (2014). What are health disparities and health equity? We need to be clear. *Public Health Rep.* 129, 5–8. Suppl 2. doi:10.1177/00333549141291S203

Cookson, R., Mirelman, A. J., Griffin, S., Asaria, M., Dawkins, B., Norheim, O. F., et al. (2017). Using cost-effectiveness analysis to address health equity concerns. *Value Health* 20, 206–212. doi:10.1016/j.jval.2016.11.027

Cui, W., and Finkelstein, J. (2022). Using EHR data to identify social determinants of health affecting disparities in cancer survival. *Stud. Health. Technol. Inf.* 290, 967–971. doi:10.3233/SHTI220224

Curigliano, G., Mueller, V., Borges, V., Hamilton, E., Hurvitz, S., Loi, S., et al. (2022). Tucatinib versus placebo added to trastuzumab and capecitabine for patients with pretreated HER2+ metastatic breast cancer with and without brain metastases (HER2CLIMB): final overall survival analysis. *Ann. Oncol.* 33, 321–329. doi:10.1016/j.annonc.2021.12.005

Fabregas, J. C., Riley, K. E., Brant, J. M., George, T. J., Orav, E. J., and Lam, M. B. (2022). Association of social determinants of health with late diagnosis and survival of patients with pancreatic cancer. *J. Gastrointest. Oncol.* 13, 1204–1214. doi:10.21037/jgo-21-788

Gray, J., Sullivan, T., Latimer, N. R., Salter, A., Sorich, M. J., Ward, R. L., et al. (2021). Extrapolation of survival curves using standard parametric models and flexible parametric spline models: comparisons in large registry cohorts with advanced cancer. *Med. Decis. Mak.* 41, 179–193. doi:10.1177/0272989X20978958

Grothey, A., Van Cutsem, E., Sobrero, A., Siena, S., Falcone, A., Ychou, M., et al. (2013). Regorafenib monotherapy for previously treated metastatic colorectal cancer (CORRECT): an international, multicentre, randomised, placebo-controlled, phase 3 trial. *Lancet* 381, 303–312. doi:10.1016/S0140-6736(12)61900-X

Guyot, P., Ades, A. E., Ouwens, M. J. N. M., and Welton, N. J. (2012). Enhanced secondary analysis of survival data: reconstructing the data from published Kaplan-Meier survival curves. *BMC Med. Res. Methodol.* 12, 9. doi:10.1186/1471-2288-12-9

Hamers, P. A. H., Elferink, M. A. G., Stellato, R. K., Punt, C. J. A., May, A. M., Koopman, M., et al. (2021). Informing metastatic colorectal cancer patients by quantifying multiple scenarios for survival time based on real-life data. *Int. J. Cancer.* 148, 296–306. doi:10.1002/ijc.33200

Jackson, C. H. (2016). flexsurv: a platform for parametric survival modeling in R. *J. Stat. Softw.* 70, i08. doi:10.18637/jss.v070.i08

Lee, H., and Singh, G. K. (2022). Disparities in all-cancer and lung cancer survival by social, behavioral, and health status characteristics in the United States: a longitudinal follow-up of the 1997-2015 National Health Interview Survey-National Death Index Record Linkage Study. *J. Cancer Prev.* 27, 89–100. doi:10.15430/JCP.2022.27.2.89

Meunier, A., Longworth, L., Kowal, S., Ramagopalan, S., Love-Koh, J., and Griffin, S. (2023). Distributional cost-effectiveness analysis of health technologies: data requirements and challenges. *Value Health* 26, 60–63. doi:10.1016/j.jval.2022.06.011

Mok, T. S. K., Wu, Y. L., Kudaba, I., Kowalski, D. M., Cho, B. C., Turna, H. Z., et al. (2019). Pembrolizumab versus chemotherapy for previously untreated, PD-L1-expressing, locally advanced or metastatic non-small-cell lung cancer (KEYNOTE-042): a randomised, open-label, controlled, phase 3 trial. *Lancet* 393, 1819–1830. doi:10.1016/S0140-6736(18)32409-7

Namburi, N., Timsina, L., Ninad, N., Ceppa, D., and Birdas, T. (2022). The impact of social determinants of health on management of stage I non–small cell lung cancer. *Am. J. Surg.* 223, 1063–1066. doi:10.1016/j.amjsurg.2021.10.022

National Cancer Institute Surveillance, Epidemiology, and End Results Program (2022a). *Breast: SEER relative survival rates by time since diagnosis, 2000-2019*. Available at: https://seer.cancer.gov/statistics-network/explorer/application.html?site=55&data_type=4&graph_type=6&compareBy=stage&chk_stage_106=106&sex=3&race=1&age_range=1&advopt_precision=1&advopt_show_ci=on&hdn_view=0&advopt_show_apc=on&advopt_display=2#resultsRegion0 (Accessed May 23, 2023).

National Cancer Institute Surveillance, Epidemiology, and End Results Program (2022b). Colon and rectum: SEER relative survival rates by time since diagnosis, 2000-2019. Available at: https://seer.cancer.gov/statistics-network/explorer/application.html?site=20&data_type=4&graph_type=6&compareBy=stage&chk_stage_106=106&sex=3&race=1&age_range=1&advopt_precision=1&advopt_show_ci=on&hdn_view=0&advopt_show_apc=on&advopt_display=2 (Accessed May 23, 2023).

National Cancer Institute Surveillance, Epidemiology, and End Results Program (2022c). Lung and bronchus: SEER relative survival rates by time since diagnosis, 2000-2019. Available at: https://seer.cancer.gov/statistics-network/explorer/application.html?site=47&data_type=4&graph_type=6&compareBy=stage&chk_stage_106=106&sex=3&race=1&age_range=1&advopt_precision=1&advopt_show_ci=on&hdn_view=0&advopt_show_apc=on&advopt_display=2#resultsRegion0 (Accessed May 23, 2023).

Podolsky, M. I., Present, I., Neumann, P. J., and Kim, D. D. (2022). A systematic review of economic evaluations of COVID-19 interventions: considerations of non-health impacts and distributional issues. *Value Health* 25, 1298–1306. doi:10.1016/j.jval.2022.02.003

Qureshi, S., Boily, G., Boulanger, J., Pagé, É., and Strumpf, E. (2023). Inequalities in survival and care across social determinants of health in a cohort of advanced lung cancer patients in Quebec (Canada): a high-resolution population-level analysis. *Cancer Med.* 12, 12683–12704. In press. doi:10.1002/cam4.5897

Ramalingam, S. S., Vansteenkiste, J., Planchard, D., Cho, B. C., Gray, J. E., Ohe, Y., et al. (2020). Overall survival with osimertinib in untreated, EGFR-mutated advanced NSCLC. *N. Engl. J. Med.* 382, 41–50. doi:10.1056/NEJMoa1913662

Rugo, H. S., Im, S. A., Cardoso, F., Cortés, J., Curigliano, G., Musolino, A., et al. (2021). Efficacy of margetuximab vs trastuzumab in patients with pretreated ERBB2-positive advanced breast cancer: a phase 3 randomized clinical trial. *JAMA Oncol.* 7, 573–584. doi:10.1001/jamaoncol.2020.7932

Saura, C., Oliveira, M., Feng, Y. H., Dai, M. S., Chen, S. W., Hurvitz, S. A., et al. (2020). Neratinib plus capecitabine versus lapatinib plus capecitabine in HER2-positive metastatic breast cancer previously treated with ≥ 2 HER2-directed regimens: phase III NALA trial. *J. Clin. Oncol.* 38, 3138–3149. doi:10.1200/JCO.20.00147

Sengupta, R., and Honey, K. (2020). AACR Cancer Disparities Progress Report 2020: achieving the bold vision of health equity for racial and ethnic minorities and other underserved populations. *Cancer Epidemiol. Biomarkers Prev.* 29, 1843. doi:10.1158/1055-9965.EPI-20-0269

Soria, J. C., Ohe, Y., Vansteenkiste, J., Reungwetwattana, T., Chewaskulyong, B., Lee, K. H., et al. (2018). Osimertinib in untreated EGFR-mutated advanced non-small-cell lung cancer. *N. Engl. J. Med.* 378, 113–125. doi:10.1056/NEJMoa1713137

Tabernero, J., Grothey, A., Van Cutsem, E., Yaeger, R., Wasan, H., Yoshino, T., et al. (2021). Encorafenib plus cetuximab as a new standard of care for previously treated BRAF V600E–mutant metastatic colorectal cancer: updated survival results and subgroup analyses from the BEACON study. *J. Clin. Oncol.* 39, 273–284. doi:10.1200/JCO.20.02088

Tran, Y. H., Coven, S. L., Park, S., and Mendonca, E. A. (2022). Social determinants of health and pediatric cancer survival: a systematic review. *Pediatr. Blood Cancer.* 69, e29546. doi:10.1002/pbc.29546

Ward, T., Mujica-Mota, R. E., Spencer, A. E., and Medina-Lara, A. (2022). Incorporating equity concerns in cost-effectiveness analyses: a systematic literature review. *Pharmacoeconomics* 40, 45–64. doi:10.1007/s40273-021-01094-7

Wu, Y. L., Lu, S., Cheng, Y., Zhou, C., Wang, J., Mok, T., et al. (2019). Nivolumab versus docetaxel in a predominantly Chinese patient population with previously treated advanced NSCLC: CheckMate 078 randomized phase III clinical trial. *J. Thorac. Oncol.* 14, 867–875. doi:10.1016/j.jtho.2019.01.006

Xu, J., Kim, T. W., Shen, L., Sriuranpong, V., Pan, H., Xu, R., et al. (2018). Results of a randomized, double-blind, placebo-controlled, phase III trial of trifluridine/tipiracil (TAS-102) monotherapy in Asian patients with previously treated metastatic colorectal cancer: the TERRA study. *J. Clin. Oncol.* 36, 350–358. doi:10.1200/JCO.2017.74.3245

frontiers | Frontiers in Pharmacology

TYPE Perspective
PUBLISHED 12 January 2024
DOI 10.3389/fphar.2023.1289365

Check for updates

OPEN ACCESS

EDITED BY
Long Ming,
Sunway University, Malaysia

REVIEWED BY
Deepak B. Khatry,
Dassault Systemes, United States

*CORRESPONDENCE
Ravinder Claire,
ravinder.claire@nice.org.uk

RECEIVED 05 September 2023
ACCEPTED 29 December 2023
PUBLISHED 12 January 2024

CITATION
Claire R, Elvidge J, Hanif S, Goovaerts H, Rijnbeek PR, Jónsson P, Facey K and Dawoud D (2024), Advancing the use of real world evidence in health technology assessment: insights from a multi-stakeholder workshop. *Front. Pharmacol.* 14:1289365.
doi: 10.3389/fphar.2023.1289365

COPYRIGHT
© 2024 Claire, Elvidge, Hanif, Goovaerts, Rijnbeek, Jónsson, Facey and Dawoud. This is an open-access article distributed under the terms of the Creative Commons Attribution License (CC BY). The use, distribution or reproduction in other forums is permitted, provided the original author(s) and the copyright owner(s) are credited and that the original publication in this journal is cited, in accordance with accepted academic practice. No use, distribution or reproduction is permitted which does not comply with these terms.

Advancing the use of real world evidence in health technology assessment: insights from a multi-stakeholder workshop

Ravinder Claire[1]*, Jamie Elvidge[1], Shahid Hanif[2], Hannah Goovaerts[3], Peter R. Rijnbeek[4], Páll Jónsson[1], Karen Facey[5] and Dalia Dawoud[6,7]

[1]National Institute for Health and Care Excellence, Manchester, United Kingdom, [2]GetReal Institute, Utrecht, Netherlands, [3]Pfizer nv/sa, Brussels, Belgium, [4]Erasmus University Medical Center, Rotterdam, Netherlands, [5]University of Edinburgh, Member Scottish Health Technologies Group Council, Edinburgh, United Kingdom, [6]National Institute for Health and Care Excellence, London, United Kingdom, [7]Clinical Pharmacy Department, Faculty of Pharmacy, Cairo University, Cairo, Egypt

Introduction: Real-world evidence (RWE) in health technology assessment (HTA) holds significant potential for informing healthcare decision-making. A multistakeholder workshop was organised by the European Health Data and Evidence Network (EHDEN) and the GetReal Institute to explore the status, challenges, and opportunities in incorporating RWE into HTA, with a focus on learning from regulatory initiatives such as the European Medicines Agency (EMA) Data Analysis and Real World Interrogation Network (DARWIN EU®).

Methods: The workshop gathered key stakeholders from regulatory agencies, HTA organizations, academia, and industry for three panel discussions on RWE and HTA integration. Insights and recommendations were collected through panel discussions and audience polls. The workshop outcomes were reviewed by authors to identify key themes, challenges, and recommendations.

Results: The workshop discussions revealed several important findings relating to the use of RWE in HTA. Compared with regulatory processes, its adoption in HTA to date has been slow. Barriers include limited trust in RWE, data quality concerns, and uncertainty about best practices. Facilitators include multidisciplinary training, educational initiatives, and stakeholder collaboration, which could be facilitated by initiatives like EHDEN and the GetReal Institute. Demonstrating the impact of "driver projects" could promote RWE adoption in HTA.

Conclusion: To enhance the integration of RWE in HTA, it is crucial to address known barriers through comprehensive training, stakeholder collaboration, and impactful exemplar research projects. By upskilling users and beneficiaries of RWE and those that generate it, promoting collaboration, and conducting "driver projects," can strengthen the HTA evidence base for more informed healthcare decisions.

KEYWORDS

health technology assessment, regulatory, real world evidence, real world data, common data model, federated data network

1 Introduction

Health technology assessment (HTA) is a multidisciplinary process that assesses the value of health technologies to inform decision-making, aiming to enhance equity, efficiency, and quality in healthcare systems (O'Rourke et al., 2020). It is widely used throughout Europe to make decisions about the reimbursement and pricing of healthcare technologies, including new medicines. Estimates of relative effectiveness, healthcare use and costs are key inputs for assessing effectiveness, cost-effectiveness, and budget impact, which are required for HTA recommendations in several countries. Companies and HTA organisations face multiple challenges in obtaining and generating such evidence in support of their products.

Traditional HTA approaches primarily rely on randomised controlled trials (RCTs) to generate clinical evidence. However, there is growing recognition of the importance of integrating real-world evidence (RWE) derived from real-world data (RWD) sources into HTA processes. RWE may provide a more comprehensive understanding of interventions' effectiveness and safety in clinical settings, and address some of the evidence gaps faced by companies and HTA organisations. However, the uptake of RWE for HTA has been slow compared with regulatory decision making.

The European Medicines Agency (EMA) has established the Coordination Centre for the Data Analysis and Real World Interrogation Network (DARWIN EU®) (darwin-eu.org) (EMA, 2021). It aims to provide access to valid and trustworthy RWE from across Europe on diseases, populations and the use and performance of medicines. This will increasingly support regulatory decision-making, which is often followed by HTA to support reimbursement decisions (EMA, 2023).

To explore the current landscape and prospects of incorporating RWE in HTA, a multi-stakeholder workshop titled "Advancing Real-World Evidence in Health Technology Assessment" was convened by the Innovative Medicines Initiative (IMI) funded European Health Data and Evidence Network (EHDEN) project (ehden.eu) (IMI, 2018), in collaboration with the GetReal Institute. EHDEN aims to enable large-scale analysis of health data in Europe by building a large federated data network of standardised data (EHDEN, 2018). Part of the project involves supporting the transition towards outcomes-driven healthcare systems in Europe, by adopting the use of a federated data network approach for HTA purposes. The GetReal Institute is an independent, member-led non-profit organisation emerging from two IMI projects with the mission to facilitate the adoption and implementation of RWE in regulatory and HTA decision-making in Europe. The aim was to foster collaboration, share experiences, and identify key strategies to facilitate the use of RWE in HTA. This article presents an overview of the workshop discussions, highlighting key findings, recommendations, and areas for future development.

2 Materials and methods

The workshop was convened with relevant stakeholders to discuss the current state, challenges, and future directions of integrating RWE into HTA processes. Experts and stakeholders were selected based on their expertise and experience in RWE and HTA. Key individuals from academia, regulatory agencies, HTA organisations, industry, and patient organisations were invited to ensure a diverse range of perspectives.

The workshop was designed as a half-day event, comprising three panels focused on specific topics related to RWE and HTA integration. Each panel consisted of a presentation followed by a moderated discussion. The first panel discussed the progress and future of DARWIN EU®, and reflections from HTA organisations on plans for adoption of RWE. The second panel focused on reflections from industry, patient organisations, and academics. The final panel discussed the potential of EHDEN and GetReal Institute in supporting RWE integration in HTA.

Following the panel presentations, open discussions were held among the workshop participants. These discussions allowed for the exchange of ideas, identification of common challenges, and exploration of strategies to overcome barriers hindering the wider adoption of RWE in HTA. The participants shared their perspectives, experiences, and recommendations based on their respective domains of expertise. In addition, two audience polls were conducted to gather insights and perspectives from the attendees. The first poll aimed to identify the areas within HTA where RWE could help resolve decision-critical evidence gaps. The second poll aimed to determine the areas where initiatives like EHDEN and the GetReal Institute could provide support for HTA. Both polls enabled participants to select more than one option to accurately capture their views.

The data collected during the workshop, including audience poll results, presentation materials, and discussion notes, were compiled. Key themes, common challenges, and potential recommendations were identified and synthesised to provide a comprehensive understanding of the workshop outcomes.

3 Results

3.1 Panel 1: EMA, DARWIN EU® & reflections from HTA organisations

The first panel focused on the establishment of the DARWIN EU® and its Coordination Centre by EMA. The discussions highlighted the ambitious goal of providing access to valid and trustworthy RWE from across Europe, encompassing diseases, populations, and the use and performance of medicines. The panel highlighted the value of standardising health data using the Observational Medical Outcomes Partnership Common Data Model (OMOP CDM) maintained by the Observational Health Data Sciences and Informatics (OHDSI) community (www.ohdsi.org) which could help to realise the necessary need to scale up real-world data studies across Europe (Hripcsak et al., 2015).

The second part of the panel focused on reflections from HTA organisations and their plans for adopting RWE. The panel highlighted that there is growing interest in utilising RWE for HTA decision-making. The need for improved trust in RWE and availability of good quality data were identified as key factors limiting its adoption. HTA organisations expressed the need for data that reflect the target population and regional variations in healthcare. The panel recognised the potential benefits of RWE in

speeding up access to new treatments and reducing the cost of drug development programs. Some HTA organisations are investing in the development of frameworks and best practices for planning, conducting, and reporting RWE studies, though there is scope for cross-border collaboration in such efforts. The panel emphasised the importance of upskilling technical staff and committees to evaluate the quality and appropriateness of RWE.

The moderated panel discussion examined the fundamental differences between regulators and HTA organisations regarding RWE use cases. Regulators focus on safety and efficacy, while HTA organisations consider relative clinical effectiveness and cost effectiveness. However, there are potential overlapping use cases that could benefit both regulators and HTA organisations, such as characterising a given disease population and its natural history. This understanding of shared goals can shape the choice of data partners for DARWIN EU® and EHDEN. The selection of these data partners is driven by stakeholders' specific questions and the need to generate relevant evidence.

The panel acknowledged the increasing interest and adoption of the OMOP CDM in Europe, particularly stimulated by the EHDEN project and the recent DARWIN-EU® initiative. This is also resulting in the establishment of so called national nodes that drive the adoption of the data model and its use in collaborative studies at the national level (www.ohdsi-europe.org).

3.2 Panel 2: reflections from stakeholders

The second panel of the workshop featured reflections from relevant HTA stakeholders representing industry, patient organisations, a health data medical research funder, and a multi-stakeholder initiative focused on RWE generation for healthcare decisions.

The industry panelist highlighted the potential value of RWE in informing reimbursement decisions but identified challenges such as data standardisation and collaboration. The patient representative expressed support for RWE but emphasised the need for resources and training for patients to understand and engage with it. The medical research funder representative emphasised infrastructure and real-time data and the RWE initiative representative highlighted challenges in data quality and the lack of expertise in utilising RWD.

The moderated panel discussion addressed the need for training and upskilling staff, particularly in healthcare decision-making bodies. Efforts to develop educational materials and align various organisations and initiatives were discussed. The potential to extend learnings from COVID-19 projects to other conditions was explored, emphasising the importance of identifying impactful "driver projects". Aligning EU member states on RWD requirements and involving decision-making bodies in data infrastructure discussions were identified as crucial steps. Accounting for real-world context in RWE studies and involving data custodians to ensure appropriate data utilisation were also highlighted.

3.3 Panel 3: RWE and HTA: how can EHDEN and GetReal Institute help?

The final panel of the workshop focused on the role of EHDEN and the GetReal Institute in supporting the generation and use of RWE in HTA. The GetReal Institute, from its previous work with the GetReal Think Tank, identified three focus areas of interest to stakeholders; reducing barriers to using secondary data sources, bridging the gap between RCTs and RWE, and addressing evidence needs of healthcare decision-makers.

HTA use cases using EHDEN were discussed, including examples in chronic obstructive pulmonary disease (Kent et al., 2021), cancer, and COVID-19. The use cases demonstrated how EHDEN's real-world data can be used in economic models, provide insights into cancer survival, and assess treatment effectiveness for COVID-19.

Two audience polls were conducted during the moderated panel discussion. The first asked, "Where does HTA experience decision-critical evidence gaps that RWE could help to resolve?." "Generalisability of trial data" was the most selected option in this poll (Figure 1), though several other gaps received a high number of votes, such as disease population characteristics, long-term health outcomes, and identifying treatment pathways. Notably, quantifying relative effectiveness received many votes, despite HTA organisations traditionally highly prioritising randomised evidence for this purpose. These results indicate that HTA processes grapple with multiple issues that suitable RWE may help to inform.

The second poll asked, "Where should initiatives like EHDEN and GetReal Institute focus their support for HTA?". "Sharing and developing best practices for conducting studies in the HTA domain" was the most selected option (Figure 2), with "provision of educational materials" the second most popular answer.

Training and upskilling in RWD was a key theme in the moderated panel discussion. It was recognised that a multidisciplinary approach is needed. This should encompass wide-ranging learning materials including topics such as phenotyping, study design, and analytical approaches. These resources should be accessible across a variety of training levels (undergraduate through to postgraduate degrees), and to HTA staff and relevant stakeholders, such as industry. Educating healthcare workers responsible for data collection is also essential, and it should be demonstrated how the collected data informs their practice and contributes to meaningful outcomes. Both EHDEN and the GetReal Institute have educational platforms targeted to a broad audience through the EHDEN Academy (academy.ehden.eu) and the GetReal Academy (getreal-academy.org). Example courses that are directly relevant to the integration of RWE in HTA include "Real-World Evidence in Medicine Development" on the GetReal Academy, and the "Health Technology Assessment" course on the EHDEN Academy.

Discussions also revolved around the future and next steps for RWD adoption in HTA. Engaging stakeholders in ongoing discussions and projects was emphasised to drive impactful advancements in HTA. The importance of "driver projects" was emphasised, as they provide practical experience and learning opportunities. It was noted that more of these projects are needed, and prompt action is necessary due to the rapid pace of change. Involving key stakeholders in the conduct of driver projects, and structured organisation of projects were considered vital for success.

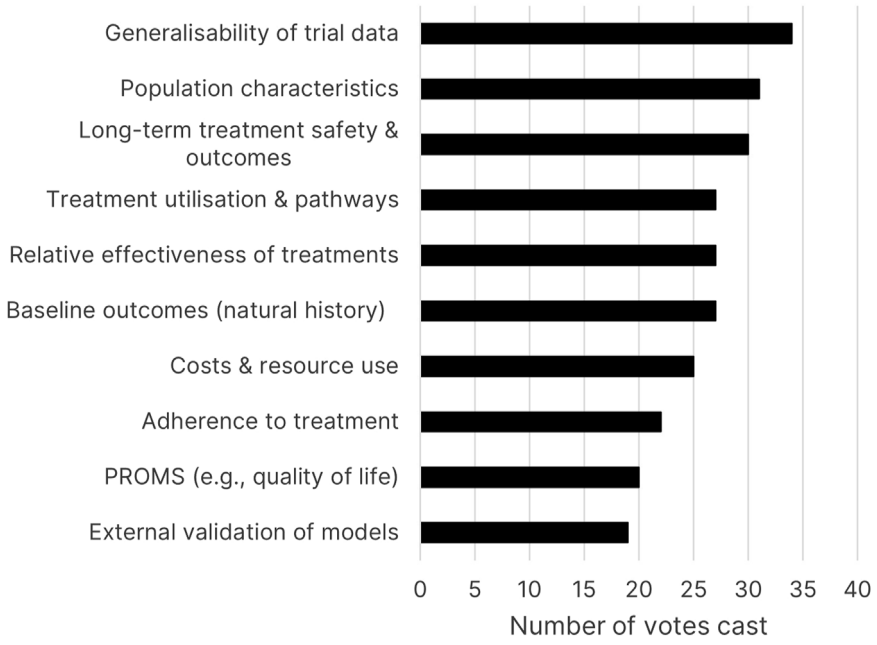

FIGURE 1
Poll results for "Where does HTA experience decision-critical evidence gaps RWE could help to resolve?" (PROMS = patient-reported outcome measures).

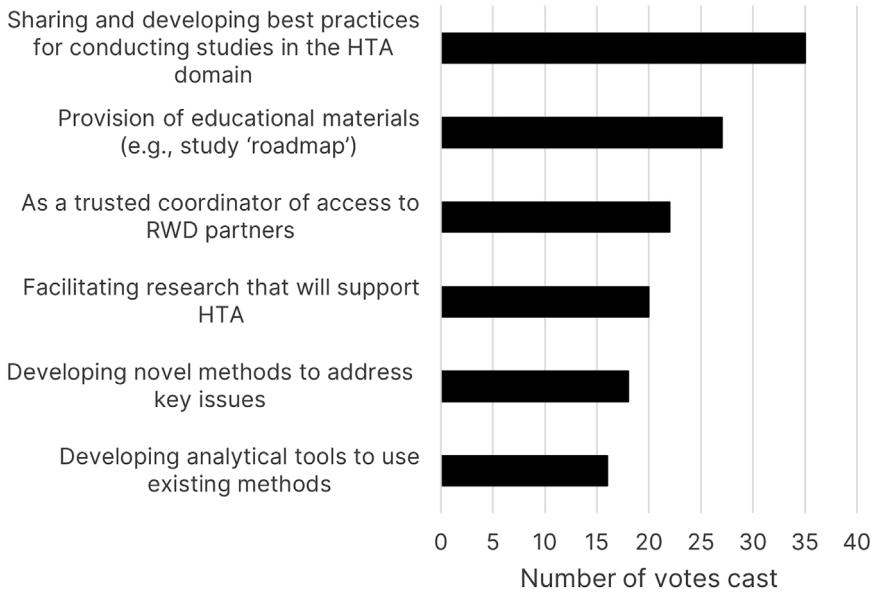

FIGURE 2
Poll results for "Where should initiatives like EHDEN and GetReal Institute focus their support for HTA?"

4 Discussion

The workshop provided valuable insights into the integration of RWE in HTA and identified key challenges and opportunities in this domain. There is a clear divergence between the acceptability of RWE for regulatory decision-making and for HTA decision-making. Traditional HTA approaches favour randomised evidence to support assessments of clinical and cost effectiveness, but HTA organisations will increasingly be presented with healthcare technologies with regulatory approval underpinned by RWE. Limited trust in RWE, concerns about data quality, and a limited understanding of what best practice is when it comes to RWE, were identified as barriers to wider adoption in HTA.

Generally, the findings are in line with published literature on barriers to adoption and use of RWE in HTA (Hogervorst et al., 2022). To address the barriers to RWE adoption, there is a need for comprehensive and multidisciplinary training and education initiatives. Starting from undergraduate levels, extending to healthcare providers responsible for data collection, and healthcare payers who make decisions about reimbursement, these efforts should aim to enhance awareness, knowledge, and expertise in assessing the quality and appropriateness of RWE.

The success of RWE integration in HTA depends on collaboration and engagement among stakeholders (Facey et al., 2020). Initiatives like EHDEN and the GetReal Institute play a crucial role in facilitating coordination, providing neutral forums, and developing resources to promote best practices and recommendations. To build trust and demonstrate the value of RWE, the identification and execution of impactful "driver projects" is essential. Initially, these projects should focus on characterising patient populations and the natural history of diseases, as these are comparatively simple analyses that can provide tangible, useful evidence for HTA quickly.

Further driver projects that focus on other identified evidence gaps, such as examining the generalisability of RCT evidence—which may require more complex studies—would be valuable. Where possible, driver projects should address evidence gaps that are common to the HTA and regulatory spaces. Such projects have been initiated within EHDEN focusing on key challenging areas for HTA including extrapolation of cancer survival data beyond the time horizon of clinical trials and assessing relative effectiveness of treatments (Claire et al., 2022). These are two key challenging methodological areas for the use of RWE in HTA, and future driver projects should similarly aim to address evidence gaps.

Based on the workshop discussions, the following recommendations are proposed to advance the integration of RWE in HTA:

- Develop and Promote Training Resources: A comprehensive strategy should be developed to create, develop, and promote training resources to upskill users and beneficiaries of RWE in HTA. These resources should cover various disciplines and target different levels of education, from undergraduates to established HTA professionals.
- Identify and Execute "Driver Projects": Key driver projects that can have a substantial impact on methodology development and build trust in the use of RWE should be identified. These projects should focus on areas of high relevance to HTA and involve collaboration among stakeholders to ensure recognition and support for their outcomes.
- Start with "Easy-Win" Projects: Initiating projects that address the characterisation of patient populations and the natural history of diseases, particularly in areas of overlap with the regulatory space, is a good starting point. These easy-win projects provide tangible outcomes and pave the way for further advancements in RWE integration in HTA.
- Collaboration and Stakeholder Engagement: Continued and deeper collaboration among stakeholders, including HTA organisations, researchers, industry representatives, and patient organisations, is crucial for the successful integration of RWE in HTA. Efforts should be made to maintain engagement, foster discussions, and drive projects that align with the vision and development of RWE adoption in HTA.

The article summarises the key findings and recommendations derived from a multi-stakeholder workshop. The insights gained from this workshop have the potential to inform future strategies and initiatives aimed at promoting the use of RWE in HTA, to support evidence-informed and patient-centred healthcare decision-making and, ultimately, better health outcomes.

Data availability statement

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

Author contributions

RC: Conceptualization, Methodology, Project administration, Writing–original draft. JE: Conceptualization, Methodology, Project administration, Writing–review and editing. SH: Conceptualization, Methodology, Writing–review and editing. HG: Conceptualization, Project administration, Writing–review and editing. PR: Writing–review and editing. PJ: Writing–review and editing. KF: Writing–review and editing. DD: Conceptualization, Methodology, Writing–review and editing.

Funding

The author(s) declare financial support was received for the research, authorship, and/or publication of this article. The European Health Data & Evidence Network has received funding from the Innovative Medicines Initiative 2 Joint Undertaking (JU) under grant agreement No. 806968. The JU receives support from the European Union's Horizon 2020 research and innovation programme and EFPIA.

Acknowledgments

The authors wish to acknowledge Juan Jose Abellan and Niklas Hedberg for their participation in the workshop.

Conflict of interest

Author HG was employed by Pfizer nv/sa.

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

Publisher's note

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

References

Claire, R., Elvidge, J., Read, C., Debray, T., Newby, D., Burn, E., et al. (2022). RWD147 demonstrating the application of real-world evidence for health technology assessment using a federated data network. *Value Health* 25, S478. doi:10.1016/j.jval.2022.09.2372

EHDEN (2018). *European health data evidence network [WWW document]*. Available at: https://www.ehden.eu/ (Accessed February 7, 2023).

EMA (2021). *Data analysis and real world interrogation network (Darwin EU) [WWW document]*. Netherlands: European Medicines Agency. Available at: https://www.ema.europa.eu/en/about-us/how-we-work/big-data/data-analysis-real-world-interrogation-network-darwin-eu (Accessed February 7, 2023).

EMA (2023). *Use of real-world evidence in regulatory decision making – EMA publishes review its studies [WWW Document]*. European Medicines Agency. Available at: https://www.ema.europa.eu/en/news/use-real-world-evidence-regulatory-decision-making-ema-publishes-review-its-studies (Accessed March 8, 2023).

Facey, K. M., Rannanheimo, P., Batchelor, L., Borchardt, M., and de Cock, J. (2020). Real-world evidence to support Payer/HTA decisions about highly innovative technologies in the EU-actions for stakeholders. *Int. J. Technol. Assess. Health Care* 36, 459–468. doi:10.1017/S026646232000063X

Hogervorst, M. A., Pontén, J., Vreman, R. A., Mantel-Teeuwisse, A. K., and Goettsch, W. G. (2022). Real world data in health technology assessment of complex health technologies. *Front. Pharmacol.* 13, 837302. doi:10.3389/fphar.2022.837302

Hripcsak, G., Duke, J. D., Shah, N. H., Reich, C. G., Huser, V., Schuemie, M. J., et al. (2015). Observational health data Sciences and Informatics (OHDSI): opportunities for observational researchers. *Stud. Health Technol. Inf.* 216, 574–578. doi:10.3233/978-1-61499-564-7-574

IMI (2018). *IMI innovative Medicines initiative | EHDEN | European health data and evidence network [WWW document]*. Belgium: Innovative Medicines Initiative. Available at: http://www.imi.europa.eu/projects-results/project-factsheets/ehden (Accessed February 7, 2023).

Kent, S., Burn, E., Dawoud, D., Jonsson, P., Østby, J. T., Hughes, N., et al. (2021). Common problems, common data model solutions: evidence generation for health technology assessment. *PharmacoEconomics* 39, 275–285. doi:10.1007/s40273-020-00981-9

O'Rourke, B., Oortwijn, W., Schuller, T., and International Joint Task Group (2020). The new definition of health technology assessment: a milestone in international collaboration. *Int. J. Technol. Assess. Health Care* 36, 187–190. doi:10.1017/S0266462320000215

frontiers | Frontiers in Pharmacology

TYPE Original Research
PUBLISHED 26 January 2024
DOI 10.3389/fphar.2024.1223858

Check for updates

OPEN ACCESS

EDITED BY
Dalia M. Dawoud,
National Institute for Health and Care Excellence, United Kingdom

REVIEWED BY
Michelle Casey,
Pfizer, United States
Enrico Capobianco,
Jackson Laboratory, United States

*CORRESPONDENCE
Letizia Polito,
letizia.polito@roche.com

†These authors have contributed equally to this work and share first authorship

RECEIVED 16 May 2023
ACCEPTED 11 January 2024
PUBLISHED 26 January 2024

CITATION
Polito L, Liang Q, Pal N, Mpofu P, Sawas A, Humblet O, Rufibach K and Heinzmann D (2024), Applying the estimand and target trial frameworks to external control analyses using observational data: a case study in the solid tumor setting.
Front. Pharmacol. 15:1223858.
doi: 10.3389/fphar.2024.1223858

COPYRIGHT
© 2024 Polito, Liang, Pal, Mpofu, Sawas, Humblet, Rufibach and Heinzmann. This is an open-access article distributed under the terms of the Creative Commons Attribution License (CC BY). The use, distribution or reproduction in other forums is permitted, provided the original author(s) and the copyright owner(s) are credited and that the original publication in this journal is cited, in accordance with accepted academic practice. No use, distribution or reproduction is permitted which does not comply with these terms.

Applying the estimand and target trial frameworks to external control analyses using observational data: a case study in the solid tumor setting

Letizia Polito[1*†], Qixing Liang[2†], Navdeep Pal[3], Philani Mpofu[2], Ahmed Sawas[2], Olivier Humblet[2], Kaspar Rufibach[1] and Dominik Heinzmann[1]

[1]Product Development Data Sciences, F Hoffmann-La Roche Ltd., Basel, Switzerland, [2]Flatiron Health, Inc., New York, NY, United States, [3]Genentech, Inc., San Francisco, CA, United States

Introduction: In causal inference, the correct formulation of the scientific question of interest is a crucial step. The purpose of this study was to apply causal inference principles to external control analysis using observational data and illustrate the process to define the estimand attributes.

Methods: This study compared long-term survival outcomes of a pooled set of three previously reported randomized phase 3 trials studying patients with metastatic non-small cell lung cancer receiving front-line chemotherapy and similar patients treated with front-line chemotherapy as part of routine clinical care. Causal inference frameworks were applied to define the estimand aligned with the research question and select the estimator to estimate the estimand of interest.

Results: The estimand attributes of the ideal trial were defined using the estimand framework. The target trial framework was used to address specific issues in defining the estimand attributes using observational data from a nationwide electronic health record-derived de-identified database. The two frameworks combined allow to clearly define the estimand and the aligned estimator while accounting for key baseline confounders, index date, and receipt of subsequent therapies. The hazard ratio estimate (point estimate with 95% confidence interval) comparing the randomized clinical trial pooled control arm with the external control was close to 1, which is indicative of similar survival between the two arms.

Discussion: The proposed combined framework provides clarity on the causal contrast of interest and the estimator to adopt, and thus facilitates design and interpretation of the analyses.

KEYWORDS

causal inference, estimand framework, target trial emulation framework, external control, oncology, real-world data

1 Introduction

Several causal inference frameworks, including the estimand framework (EF), target trial emulation framework (TTF), and PICO framework, exist to help define a precise scientific question for comparative assessments in clinical research and development (Goetghebeur et al., 2020). There are overlapping but complementary elements in these frameworks, suggesting the potential for a combined application; however, this presents challenges to investigators as there are limited practical examples and guidance for the combined application of the frameworks.

The EF has increasingly been adopted by health authorities and pharmaceutical companies since its initial publication in August 2017 (Food and Drug Administration, 2021). The EF enables researchers to specify a precise scientific question by using five attributes that define the estimand (i.e., the treatment effect of interest or the "what to estimate"). These five interrelated attributes are population, treatment, variable of interest (endpoint), intercurrent event handling, and the summary measure. An intercurrent event is an event occurring after treatment initiation that affects either the interpretation or the existence of the measurements associated with the endpoint. For example, if performing a comparative assessment on overall survival (OS) between two different treatments, candidates for intercurrent events include, among others, early discontinuation of treatment or treatment switching after disease progression. In general, the definition of the estimand comes first and is derived from the scientific objective of the trial or study. Together with considerations about missing data, the framework then informs the choice of the estimator. The addendum acknowledges that usually an iterative process will be necessary to reach an estimand that is clinically relevant for decision making and for which a reliable estimate can be computed. If it is not possible to develop an appropriate trial design or to derive an adequately reliable estimate for a particular estimand, an alternative estimand, trial design, or method of analysis may need to be considered. However, practical examples in the literature describing such an iterative process to redefine an initial target estimand, while also considering aspects of identifiability (and hence the estimator) are limited. While the focus of the ICH E9 addendum is on randomized clinical trials (RCTs), the principles are also applicable whenever estimating a treatment effect (i.e., non-randomized studies). However, estimation of a causal effect from observational data, compared to RCT data, often has additional challenges. Namely, observational data is more often incomplete, heterogeneous, and subject to different types of measurement errors and biases (e.g., selection bias, bias due to baseline confounding, and the ability to correctly define the index date for comparison) (Liu and Panagiotakos, 2023).

The TTF is another causal framework that can be used to specify the scientific question more precisely in a comparative assessment (Hernán and Robins, 2016). TTF complements the EF by addressing gaps related to the analysis of observational data and applies design principles of an RCT to the specific setting of a non-randomized comparative assessment (Hernán et al., 2008; Cain et al., 2016; Hernán and Robins, 2016; Petito et al., 2020). TTF entails defining a hypothetical randomized trial to address a precise scientific question and then further specifying how it can be emulated (i.e., approximated) by non-randomized data. The essential components of a target trial protocol are eligibility criteria, treatment strategies, treatment assignment, start/end of follow-up, outcomes, causal contrasts, and the analysis approach (estimator) (Hernán and Robins, 2016). The framework can also be utilized when a combination of clinical trial and observational data are used, for example, to contextualize a single-arm clinical trial with observational data (Thomas et al., 2021). Combining the EF and the TTF provides a structured approach to enhance the scientific rigor for causal inference for observational and/or non-randomized data. Together they bring more transparency to the causal estimand, which supports specifying the attributes of the estimand and the assumptions made to draw causal conclusions.

Another framework that aims to define the precise scientific question includes the PICO framework (Schardt et al., 2007), traditionally used in epidemiology for observational studies. The EF and TTF extend the PICO framework, with the former adding intercurrent events and ensuring that the population-level summary measure is made explicit, and the latter adding the causal contrast, assignment procedures, and the start/end of follow-up. By explicitly calling out these key elements, the treatment effect can be adequately defined.

An important goal in pharmacoepidemiology is to assess whether observational data (including electronic health record [EHR]-derived data) can emulate (and thus supplement or replace, e.g., for regulatory decision-making) the control arm of a RCT, while acknowledging that there are differences between clinical trial and routine clinical settings, at baseline and post-baseline, that may have an impact on the outcome independently from the treatment received. In this study we jointly apply the EF and TTF to perform a comparative effectiveness assessment in patients with non-small cell lung cancer (NSCLC) using data from a set of pooled control arms of three RCTs as well as EHR-derived de-identified observational data (West et al., 2019; Jotte et al., 2020; Nishio et al., 2021). The objective of our case study was to determine whether there is a difference in OS between patients with metastatic NSCLC receiving front-line chemotherapy in pivotal trials *versus* patients with metastatic NSCLC who received front-line chemotherapy as part of routine care, had patients not received a subsequent therapy. This case study aims to illustrate the application of the EF to observational data, and the benefits of complementing the EF with the TTF to account for specific challenges in observational data that are not directly addressed by the EF (and *vice versa*, as the handling of intercurrent events is not explicitly addressed in the TTF). The iterative process (as indicated in the EF) to arrive at the final scientific question is illustrated in the Methods section. In sum, the present study provides insights into where the two frameworks are complementary and provides a practical example of their joint application.

2 Materials and methods

2.1 Applying the frameworks to the research question

Before discussing details of the joint application of the EF and TTF to define the final scientific question, we want to provide insights and stepwise practical guidance on the iterative process outlined in the EF to arrive at the final question:

TABLE 1 EF/TTF attributes based on the scientific research question.

Scientific research question			
Would there be a difference in OS between patients with metastatic NSCLC receiving front-line chemotherapy (control arms) in IMpower trials (130, 131 and 132) vs. patients with metastatic NSCLC who received front-line chemotherapy as part of routine care, had patients not received a subsequent therapy?			
EF/TTF Attributes	Target trial	Emulation of the target trial	Assumptions
Target population/ Eligibility criteria	Metastatic squamous and non-squamous NSCLC patients, 18 years of age or older, with ECOG PS 0,1 and with adequate hematological and end-organ function. The population is defined through the common I/E criteria of IMpower130, 131 and 132 (limited to those criteria applicable retrospectively to observational data). To align the I/E criteria of the 3 trials, and to reflect the targeted population treated with 1L chemotherapy, patients with a sensitizing mutation in the *EGFR* gene or an *ALK* fusion oncogene were excluded	Same as the target trial for the RCT arm, with some assumptions for the OC arm	Observational data does not perfectly emulate the trial I/E criteria. We attempt to define the study cohort that best approximates the target population by including additional rules
			• Time window for the eligibility assessment (ECOG PS, lab values, biomarker)
			• How to handle missing values (ECOG PS, lab values, biomarker)
			○ Excluding patients with missing value may introduce selection bias
			• Rules to account for difference between trial structured visits and routine clinical care
			○ E.g., Patients with structured activity within 90 days of advanced diagnosis
Treatment/Treatment strategies	The investigational arm (pooled trial control arms) and the OC arm received the following chemotherapies	Same as the target trial with some assumptions for both arms	Assumption on treatment: • For this study we assume equivalence of nab-paclitaxel and paclitaxel. However, the two molecules are known to have different safety profiles. The decision to include paclitaxel was to limit treatment assignment bias since nab-paclitaxel is not the standard of care in the real world while it was adopted in IMpower trials
	Patients with non-squamous NSCLC		
	-Pemetrexed + cisplatin/carboplatin		
	-nab-paclitaxel/paclitaxel + carboplatin*		
	Patients with squamous NSCLC		
	-nab-paclitaxel/paclitaxel + carboplatin*		
	The investigational arm received care according to the trial protocol, whereas the comparator arm received care according to routine clinical practice		
Endpoint/Outcomes	OS	Same as the target trial	None. The validity of the rwOS from Flatiron Health has been demonstrated (Zhang et al., 2021) against clinical trial OS as the gold standard to capture death occurrence. For this reason, in this study we refer to OS and not to rwOS for routine clinical practice
Intercurrent events (IE) and strategy/Causal contrast	IE: Receipt of any subsequent cancer therapy	Same as the target trial	None
	Strategy: hypothetical		
	Causal contrast: Per-protocol effect of adhering to treatment after initiation. Receipt of any subsequent cancer therapy is a deviation from the study protocol.		
Population-level summary/analysis plan	HR with 95% CI	Same as the target trial	None
Assignment procedures	Participants were randomly assigned to one of the two treatment settings	Randomization is emulated by weighting observations for the inverse probability of treatment setting assignment following some assumptions	Clinical assumptions
			Treatment setting assignment was assumed to be conditional on the following baseline covariates
			• Age, gender, race, metastatic tumor type (*de novo* Stage IV/recurrent disease), time from initial diagnosis to index date,

(Continued on following page)

TABLE 1 (*Continued*) EF/TTF attributes based on the scientific research question.

Scientific research question			
Would there be a difference in OS between patients with metastatic NSCLC receiving front-line chemotherapy (control arms) in IMpower trials (130, 131 and 132) vs. patients with metastatic NSCLC who received front-line chemotherapy as part of routine care, had patients not received a subsequent therapy?			
EF/TTF Attributes	Target trial	Emulation of the target trial	Assumptions
			smoking history, histology, and treatment type
			Statistical assumptions
			Statistical assumptions include consistency, conditional exchangeability, positivity and correct model specification. These are explained in the text
Start/end follow-up	Start of follow-up occurs at the time when the treatment is assigned (i.e., when eligibility is met) End of follow-up is reported in Supplementary Table 1	Same as target trial. To emulate the start of follow up for the OC arm, some assumptions are needed. To emulate the end of follow up we truncated the follow-up time at Month 21 because there were few patients remaining in the RCT arm after Month 21	For the OC arm, the actual start of follow-up occurs at the time when the treatment is initiated (dose 1 cycle 1)
			The risk of comparing different time zero is to introduce immortal time bias. This cannot be quantified. The primary estimate is unbiased if the following assumptions are met.
			Assumptions in the OC
			• There are no reasons for a patient to not initiate treatment other than death once assigned to treatment
			• Death is unlikely to have occurred in between assignment and start of treatment because we assume
			○ The time between assignment and start of therapy is short
			○ mNSCLC is a disease with no rapid course in first line
			No assumption for RCT. We verified that
			• All patients assigned to treatment started treatment
			• Median time between assignment and start of therapy was 2 days

Notes: 1L, first-line therapy; ALK, anaplastic lymphoma kinase; CI, confidence interval; EGFR, epidermal growth factor receptor; HR, hazard ratio; I/E, inclusion and exclusion; mNSCLC, metastatic non-small cell lung cancer; NSCLC, non-small cell lung cancer; OC, observational comparator; OS, overall survival; RCT, randomized clinical trials; rwOS, real-world overall survival.

Step 1: determine the comparison of interest.

Step 2: develop the scientific question.

Step 3: discuss the implications of estimating the estimand aligned with the scientific question, thinking in terms of estimand attributes, including potential intercurrent events and the consequences of different strategies used to handle them.

Step 4: refine the scientific question if needed and iterate Steps 3–4 until the question is clear enough to leave no ambiguity about the estimand.

Applying these steps, we were interested in comparing the treatment effect of the same front-line treatment given in a clinical trial *versus* in the clinical practice when subsequent treatments would be similar. We started with the scientific question: "Is there a difference in OS between patients with metastatic NSCLC receiving front-line chemotherapy in pivotal trials *versus* patients with metastatic NSCLC who received front-line chemotherapy as part of routine care?" EHR-derived observational data from routine clinical practice suggests a larger heterogeneity in subsequent second-line cancer treatments as compared to the clinical trial setting (Signorovitch et al., 2022). This difference in the range of potential subsequent therapies may introduce complexities in estimating causal treatment effects for longer-term outcomes such as OS and ultimately complicate interpretations. Therefore, the initial research question has been iterated to: "Is there a difference in OS between patients with metastatic NSCLC receiving front-line chemotherapy in pivotal trials *versus* patients with metastatic NSCLC who received front-line chemotherapy as part of routine care, had patients not received a subsequent therapy?" Hence, instead of considering the entire

treatment strategy (front-line and subsequent therapy) which is complicated by heterogeneity in subsequent therapies among clinical trial and clinical practice settings, the iteration resulted in the scientific question of treatment effect of the front-line regimens.

Now we focus on jointly applying the EF and TTF to the final scientific question. Table 1 displays the EF/TTF attributes that define the estimand aligned with the scientific research question. We define the hypothetical target trial structured according to the EF and the study that attempts to emulate it, leveraging elements from the EF and TTF. The average treatment effect on the treated (ATT) is the estimand of primary interest. This is the treatment effect difference of using front-line chemotherapy in a clinical trial *versus* in clinical practice, where the target population is defined by the population of the three clinical trials.

2.2 Data source

2.2.1 Clinical trial data

Individual patient-level data (IPD) were used from Roche-sponsored phase III, open-label randomized clinical trials IMpower130 (ClinicalTrials.gov identifier: NCT02367781), 131 (ClinicalTrials.gov identifier: NCT02367794), and 132 (ClinicalTrials.gov identifier: NCT02657434). Methods and primary findings have been previously reported (West et al., 2019; Jotte et al., 2020; Nishio et al., 2021). These three trials included patients who were chemotherapy-naive and had stage IV NSCLC. OS was the primary endpoint for the three trials. To address the objective of the present study, only the IPD from the control arms were used. The control arms received platinum-based chemotherapy as follows:

- IMpower130 included patients with non-squamous NSCLC treated with carboplatin plus nab-paclitaxel
- IMpower131 included patients with squamous NSCLC treated with carboplatin plus nab-paclitaxel
- IMpower132 included patients with non-squamous NSCLC treated with carboplatin or cisplatin plus pemetrexed

As these three clinical trial control arms had similar settings in terms of disease, therapy, and inclusion/exclusion criteria and had similar survival outcomes such as median survival time (Supplementary Figure S1), they were pooled together to increase the sample size and are collectively referred to as the RCT arm in this study.

2.2.2 Observational data

The observational comparator (OC) arm of this study was developed using the nationwide (US-based) Flatiron Health EHR-derived de-identified database. This longitudinal database is comprised of patient-level structured (e.g., laboratory values and prescribed treatments) and unstructured data (e.g., biomarker reports) curated from technology-enabled chart abstraction from physicians' notes and other documents (Birnbaum et al., 2020; Ma et al., 2020). During the study period, the de-identified data originated from approximately 280 cancer clinics (approximately 800 sites of care, primarily community-based cancer centers). The studies involving human participants were reviewed and approved by the IRB of WCG IRB and included a waiver of informed consent.

2.3 Cohort selection/study sample

The OC cohort was selected to align, as closely as possible, to the eligibility (inclusion/exclusion) criteria of the three clinical trials, which reflected the eligibility criteria of the target trial (Table 1 and Supplementary Table S2). This deliberate selection allowed us to define a pooled sample of one common target population. To be eligible for entry into the de-identified database, the patient's EHR must include >1 visit to a community oncology clinic and have confirmation of an advanced NSCLC diagnosis and histological subtype (squamous vs. non-squamous histology) through a review of unstructured data (i.e., clinical notes, radiology reports, or pathology reports). A front-line therapy start date for advanced or metastatic NSCLC on or after 16 April 2015, and on or before 31 May 2017, to match the clinical trials' start and end dates of enrollment was also required. Patients with an Eastern Cooperative Oncology Group performance status (ECOG PS) of 0, 1, or unknown were included. Patients had to have received at least one administration of the regimens of interest (i.e., carboplatin plus paclitaxel/nab-paclitaxel, carboplatin, or cisplatin plus pemetrexed). Patients who had potentially incomplete historical treatment data (i.e., >90-day gap between advanced diagnosis and structured activity in the EHR), therapy within 6 months before the start of front-line therapy for advanced-stage disease, receipt of a clinical study drug, or multiple primary tumors were excluded. Patients with missing information or known to have a sensitizing mutation in the epidermal growth factor receptor (*EGFR*) gene or anaplastic lymphoma kinase (*ALK*) fusion oncogene were excluded. All patients were followed until 18 July 2019. Detailed inclusion/exclusion criteria were included in Supplementary Table S2.

2.4 Statistical analyses

We applied the following estimation approach to target the ATT estimand with attributes as specified in Table 1. First, the inverse probability of treatment weighting (IPTW) method was used to balance baseline patient characteristics between the RCT arm and the OC arm. A multiple logistic regression model was used to estimate propensity scores (PS) that are defined as probabilities of being assigned to the RCT arm conditional on all confounders that were selected based on clinical experts' knowledge and availability of the relevant variables. Because we target the ATT as described above, patients from clinical trials were given a weight of one. In contrast, patients' weights from the OC cohort were defined as the ratio of the estimated PS to one minus the estimated PS (i.e., odds of being treated in the clinical setting). We refer to these weights as IPTW-ATT weights. Before and after IPTW-ATT weighting, differences in baseline characteristics were assessed through standardized mean and proportion differences (SMD) (Table 2; Figure 1). Patient characteristics were considered statistically different if SMD ≥0.10 (Austin and Stuart, 2015). In addition, we examined the propensity score distribution to ensure a

TABLE 2 Baseline characteristics.

Variable	Categories	RCT arm, N = 849	OC arm, N = 3340	SMD
Age group (years), n (%)	<65	435 (51.2)	1222 (36.6)	0.42
	≥65 and <75	322 (37.9)	1268 (38.0)	
	≥75	92 (10.8)	850 (25.4)	
Gender, n (%)	Female	248 (29.2)	1457 (43.6)	0.30
Race, n (%)	Asian	105 (12.4)	46 (1.4)	0.75
	White	699 (82.3)	2373 (71.0)	
	Other	45 (5.3)	921 (27.6)	
ECOG PS, n (%)	0	314 (37.0)	714 (21.4)	0.05a
	1	532 (62.7)	1179 (35.3)	
	Unknown	2 (0.2)	1447 (43.3)	
Tumor diagnosis type, n (%)	*De novo* Stage IV	706 (83.2)	2118 (63.4)	0.46
	Recurrent disease	143 (16.8)	1221 (36.6)	
Smoking history, n (%)	No	69 (8.1)	257 (7.7)	0.02
	Yes	780 (91.9)	3070 (91.9)	
	Unknown	0 (0.0)	13 (0.4)	
Histology, n (%)	Non-squamous	509 (60.0)	2278 (68.2)	0.17
	Squamous	340 (40.0)	1062 (31.8)	
Time from initial diagnosis to index date (months), median [IQR]		1.41 [0.92, 2.89]	1.25 [0.79, 2.27]	0.15
Treatment, n (%)	Carboplatin + Pacli/Nab-pacli	568 (66.9)	1877 (56.2)	0.22
	Platinum + Pemetrexed	281 (33.1)	1463 (43.8)	

Notes: ECOG PS, eastern cooperative group performance status; OC, observational comparator; RCT, randomized clinical trial; SMD, standardized mean and proportion differences.
[a]The "unknown" category was not considered for SMD, calculation.
[b]ECOG PS, variable was not included in the propensity score model because of the high proportion of missing values.

reasonable overlap between the two cohorts. The weighted population was used in the subsequent analyses.

Secondly, the inverse probability of censoring weighting (IPCW) method was used to handle informative censoring introduced by censoring patients upon the occurrence of the intercurrent event of interest, i.e., receipt of any subsequent cancer therapy, as per the hypothetical strategy of handling intercurrent events (Table 1). We artificially censored patients at the time of receipt of first second-line treatment and used the IPCW method to estimate weights for the follow-up information for the remaining patients using both baseline and time-varying variables, which are likely to impact treatment switching based on clinical experts' knowledge to adjust for any potential confounding created by the artificial censoring. Specifically, we fit a Cox model for each arm that was used to estimate the probability of not being censored by time (t) given baseline and time-varying covariates (listed in Table 2) for the specific group. The IPCW weights are calculated as the inverse of the conditional probability of not being censored. We truncated the follow-up time at Month 21 because there were few patients remaining in the RCT arm after Month 21 and thus, the positivity assumption was unlikely to hold. This approach was adopted to emulate the end of follow-up of the target trial (Table 1). Then, to reduce variance of the weighted estimator, we calculated the stabilized IPCW weight (Austin and Stuart, 2015), which is the probability of not being censored conditional on selected baseline covariates, divided by the probability of not being censored, conditional on both baseline and time-varying covariates. The mean, standard deviation, minimum, and maximum estimated weights were used to inspect the robustness of the estimator. Estimated weights with the mean far from one—or very extreme values—are indicative of non-positivity or misspecification of the weight model (Hernán and Robins, 2006).

The treatment effects were estimated using weighted survival analysis methods. Hazard ratio (HR) and 95% CI were adopted for the population-level summary (Table 1). Specifically, we estimated the HR, using an IPTW-ATT-IPCW weighted Cox proportional hazard model and the 95% CI for the HR using the bootstrap approach (Schaubel and Wei, 2011). We also used the IPTW-ATT-IPCW weighted Kaplan-Meier method to compute OS function estimates and weighted log-rank test to compare across groups. Hence, the double weighting estimation approach targets the ATT estimand with attributes of the EF and TTF as specified in Table 1.

Missing values for covariates with a missing rate less than 30% were imputed using median (for age and time from initial diagnosis to index date) or mode (smoking history). Covariates with more than 30% of values missing (i.e., ECOG PS) were not imputed and

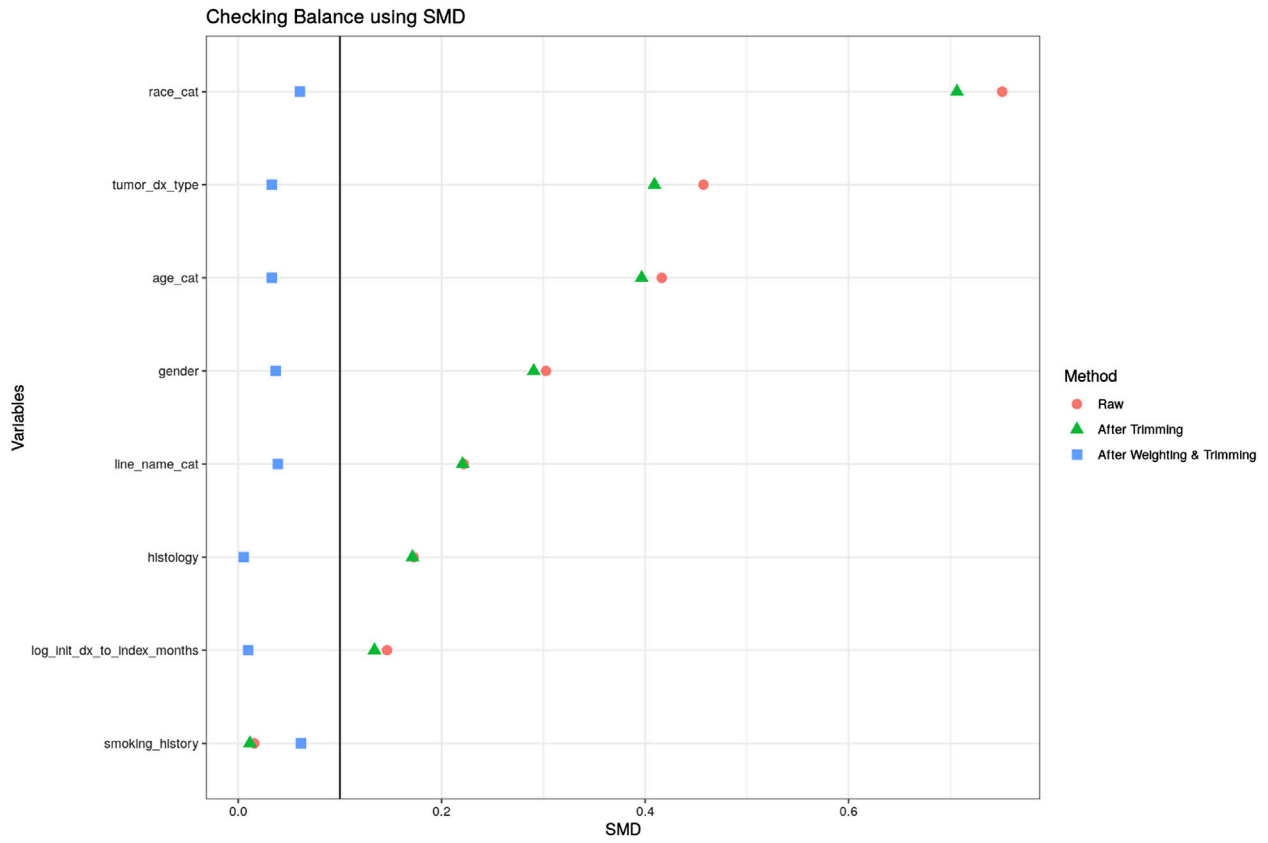

FIGURE 1
Covariate balance after IPTW-ATT.

TABLE 3 Characteristics of intercurrent events.

	RCT	OC
Number of patients	849	3340
Median (95% CI) follow-up time, months	26.5 (19.9–28.8)	35.6 (29.4–43)
Switch to subsequent therapy (any), n (%)	449 (52.9%)	1881 (56.3%)
Median (IQR) time to switch (among patients who switched), months	6.24 (4.27–9.69)	5.45 (3.12–9.43)
Number of patients who switched prior to 6 months/Number of patients who ever switched, n (%)	207/449 (46.1)	1049/1881 (55.8)

Notes: CI, confidence interval; IQR, interquartile range; OC, observational comparator; RCT, randomized clinical trial.

excluded from the IPW models. We performed a sensitivity analysis by analyzing the whole follow-up period for RCT and OC arms instead of truncating them at Month 21. Also, to evaluate to what extent our estimation methods remove the potential bias on OS due to baseline confounders and intercurrent events, we performed the traditional IPTW-only method that adjusts for baseline characteristics but not intercurrent events in terms of Kaplan-Meier (K-M) estimate and HR, and compared it to our proposed method. To follow the structure of the EF, we consider this IPTW-ATT-only estimation as a supplementary analysis because it estimates an estimand different from our target estimand.

No formal hypothesis testing was conducted, and thus, no statistical significance was explicitly assessed.

R (3.6.1) was used for the analyses.

3 Results

3.1 Cohort characteristics

A total of 849 patients were in the RCT arm and 3,340 patients were in the OC arm (refer to Supplementary Table S3 for the OC cohort attrition table). Demographic and clinical characteristics of the study sample at baseline are presented in Table 2 (and in Supplementary Table S4 stratified by RCT). Statistically significant differences between the RCT and OC arms were observed in age, gender, race, ECOG PS,

10.3389/fphar.2024.1223858

TABLE 4 Baseline and clinical characteristics among patients who switched treatment and who did not switch treatment.

Variable	Category, n (%)	RCT		OC	
		Patients who switched treatment	Patients who did not switch treatment	Patients who switched treatment	Patients who did not switch treatment
		N = 449 (52.9%)	N = 400 (47.1%)	N = 1881 (56.3%)	N = 1459 (43.7%)
Age	<65	227 (50.6)	207 (51.9)	708 (37.6)	514 (35.2)
	65–75	179 (39.9)	143 (35.8)	717 (38.1)	551 (37.8)
	≥75	43 (9.6)	49 (12.3)	456 (24.2)	394 (27.0)
Histology	Non-squamous	251 (55.9)	258 (64.5)	1287 (68.4)	991 (67.9)
	Squamous	198 (44.1)	142 (35.5)	594 (31.6)	468 (32.1)
Treatment	Carboplatin + Pacli/Nab-pacli	287 (63.9)	281 (58.5)	1034 (55.0)	843 (57.8)
	Platinum + Pemetrexed	162 (36.1)	119 (41.5)	847 (45.0)	616 (42.2)
Progression during the follow-up[a]	Yes	390 (86.9)	230 (57.5)	1360 (72.3)	397 (27.2)
	No	59 (13.1)	170 (42.5)	521 (27.7)	1062 (72.8)

Notes: OC, observational comparator; RCT, randomized clinical trial.
[a]Follow-up is up to switch or, in absence of switch until last activity before study end date (end date for the specific data source).

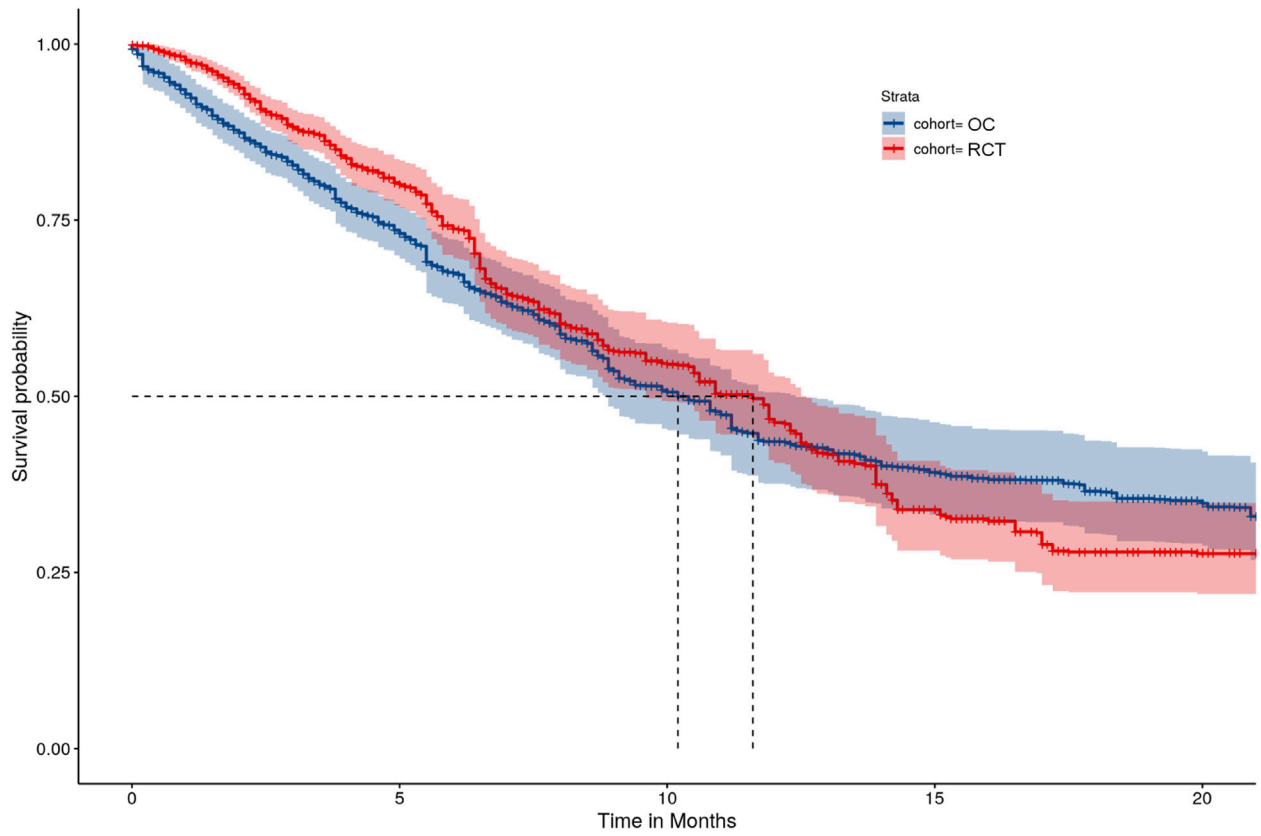

FIGURE 2
IPTW-ATT-IPCW weighted Kaplan-Meier curves.

tumor diagnosis type (*de novo* Stage IV/recurrent disease), histology, time from initial diagnosis to index date, and treatment type. Patients in the OC arm were older, with a higher percentage of females, races other than White and Asian, recurrent disease and non-squamous histology, shorter time from initial diagnosis to index date, and less frequently treated with carboplatin plus paclitaxel/nab-paclitaxel.

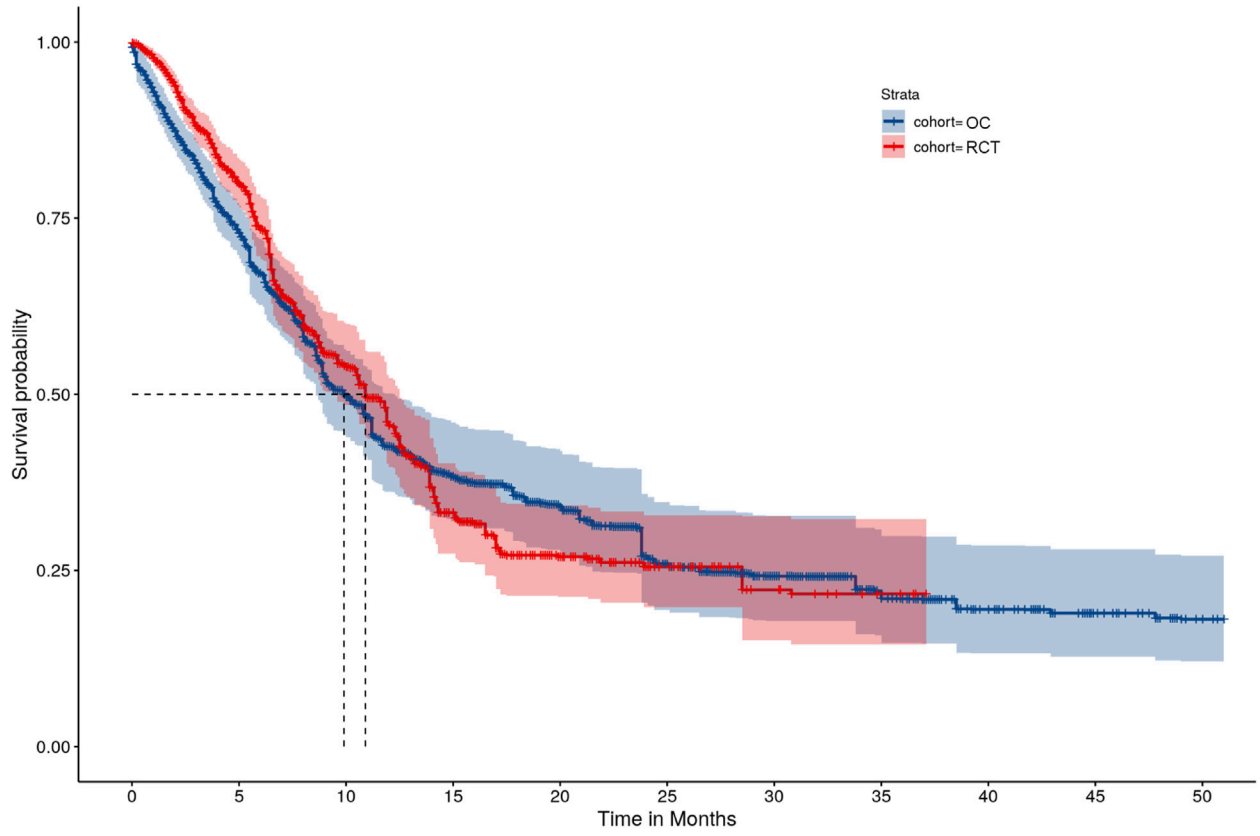

FIGURE 3
IPTW-ATT-IPCW weighted Kaplan-Meier survival function estimates without truncating the follow-up time.

The percentage of patients who switched to subsequent antineoplastic treatment, i.e., the intercurrent event of interest, was higher in the OC arm compared to the RCT arm (56.3% vs. 52.9%; Table 3) during the whole follow-up period. Among patients who switched, the median time to treatment switch was shorter in the OC arm compared to the RCT arm (5.45 vs. 6.24 months; 55.8% vs. 46.1% switched in the first 6 months). Differences in pre-specified confounders for treatment switching including age, histology, treatment type, and progression were observed. Specifically, we saw a higher percentage of switching among patients with progression events during the follow-up period in both RCT and OC arms (Table 4).

3.2 Main analyses

A logistic regression model was fitted to account for imbalances between the RCT and OC arms on baseline characteristics and estimate the PS. Then IPTW-ATT weights were calculated using the PS estimated from the logistic model and we excluded a small percentage of patients (0.4%) with extreme weights (weight >10) in the OC arm to avoid undesirable variability in estimates due to extremely large weights (Potter and Zheng, 2015). Supplementary Figure S2 shows the distribution of the PSs in the OC and RCT arms, which served as the basis to compute the IPTW-ATT weights. SMDs for patient variables were all below 0.1 after IPTW-ATT (Figure 1), suggesting balance achieved on the selected baseline characteristics through IPTW-ATT weighting (Austin, 2009) when trying to emulate randomization (more detail in Table 1).

Patients were artificially censored at the time of treatment switching (i.e., the intercurrent event of interest), then the censoring mechanism was modeled via a Cox regression model, and the probability of not being censored conditional on patient/clinical characteristics that were pre-specified was estimated (Table 4). The stabilized IPCW weights were calculated as the ratio of the inverse of the probability of not being censored conditional on race only and the probability of not being censored conditional on age, race, histology, and progression. Here, different from the traditional stabilized weights, race was added to both the numerator and denominator to further increase the stability of the IPCW weight (Cole and Hernán, 2008). To make a stable estimation and reduce variability, extreme weights were trimmed at the 99th percentile for the OC arm and the 98th percentile for the RCT arm. The distribution of the weights after trimming is displayed in Supplementary Table S5. The mean stabilized weights had means close to one, a necessary condition for correct model specification (Hernán and Robins, 2006).

After accounting for treatment setting assignment at baseline and treatment switching using the IPTW-ATT-IPCW method, the HR estimated from the weighted Cox model was equal to 0.94 (95% CI: [0.77, 1.13]), which suggests comparable OS between the RCT and OC arms. Weighted K-M estimates of survival functions overall were

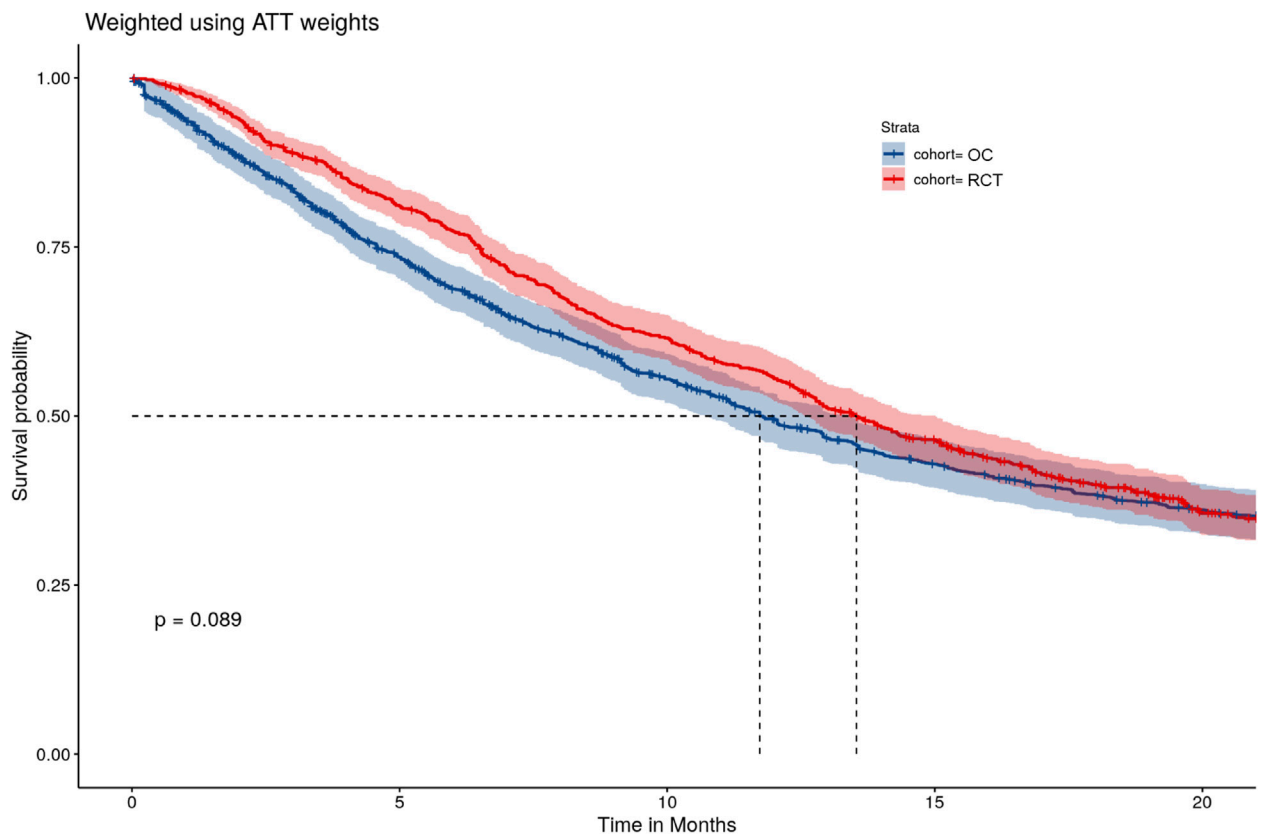

FIGURE 4
IPTW-ATT weighted Kaplan-Meier curves.

comparable (Figure 2), however, there was crossing hazard between the two arms. The two curves align well at months 7–14, while RCT performed better at months 0–6 and worse at months 15–23. The difference in median survival time between the two arms was small (9.9 months with 95% CI: [8.6, 12.3] for the OC cohort *versus* 10.9 months with 95% CI: [9.6, 12.5] for the RCT cohort). These results suggested that after accounting for imbalances of baseline characteristics and removing the confounding effects of treatment switching, patients in the OC arm had similar OS as those in the RCT arm.

3.3 Sensitivity analyses

A sensitivity analysis was performed to analyze the entire follow-up period (i.e., no truncation at 21 months) for the RCT and OC arms. The HR was 0.93 (95% CI: [0.77, 1.13]), which was similar to the primary analysis results. However, there were wider confidence intervals for K-M curves after month 21 for both the RCT and OC arms due to the low number of events (Figure 3).

3.4 Supplemental analyses

In a supplemental analysis, we performed an IPTW-ATT-only analysis that adjusted for baseline characteristics only by IPTW-ATT weighting but without IPCW. This is a commonly used method in analyses of external control arms, resulting in a different estimand compared to the primary analysis. Although the HR was similar to the primary analysis (0.92, 95% CI: [0.81, 1.05]), there was a larger discrepancy in K-M estimates between the RCT and OC arms, especially during Months 6 and 14, compared to the primary analysis (Figure 4).

4 Discussion

In this study, we applied the EF and TTF to define a precise scientific question in comparative-effectiveness research. As a case study to illustrate how to apply the EF and TTF when designing an external control study using observational data, we conducted a retrospective cohort study to compare OS among patients with metastatic NSCLC exposed to front-line chemotherapy in RCTs *versus* routine clinical practice settings, while accounting for differences in subsequent treatments between these settings. To achieve this objective, we pooled clinical trial patients from the control arms of three RCTs (IMPOWER 130, 131, and 132) and derived an OC cohort from de-identified EHR data obtained from routine clinical practice. OS was compared between the two arms, assuming a hypothetical scenario wherein patients in neither setting received subsequent therapy after the first-line chemotherapy. We found no relevant difference in OS between the two arms. Hence, when accounting for baseline confounding as well

as differences in patterns of subsequent treatments in clinical trial and routine clinical practice care patients, the long-term outcome of first-line treatment for patients with metastatic NSCLC is similar despite the lack of full trial entry criteria implementation.

Our approach attempts to clarify the causal contrast of interest by combining elements of the EF and TTF. The EF and TTF serve complementary purposes in answering the scientific question. As formulated by Hernán and Robins, the TTF ensures that an appropriate comparative study is designed to help estimate the causal effect from the observed data (Hernán and Robins, 2016). While the causal contrast can be specified within the TTF, the EF adds clarity to the causal contrast through the explicit consideration of intercurrent events (i.e., events occurring post-baseline that can affect the assessment of treatment effects). Combined, the EF and TTF improves transparency in the: 1) target of estimation (causal contrast), 2) assumptions and data needed to identify the causal contrast, and 3) limitations of available data.

To our knowledge, there are a limited number of studies that combine the EF and TTF. Recently, Hampson et al. combined the EF and TTF using routine clinical care data to generate an external control arm (Hampson et al., 2023). The approach described in our study adds to the limited number of use cases by accounting for a scenario where patterns of subsequent treatments are different between the sources of clinical trials and routine clinical care. We anticipate that many researchers will likely encounter this scenario in applications involving real-world external controls. Our study, unlike other studies, also illustrates the iterative nature of specifying an estimand. In practice, such iteration allows a comprehensive and transparent dialogue among stakeholders to reach a consensus on the scientific question and its tractability given the available data (i.e., discern the identifiability of the estimand).

Strengths of this study include the combination of the EF and TTF, its large sample size, extensive follow-up, and its high proportion of patients with an event of interest. In addition, to mitigate possible sources of bias due to heterogeneity from comparing the RCT and OC arms, we emulated randomization with IPTW. Furthermore, the real-world data source we selected reports key variables with high accuracy and clinical relevance. For example, the composite real-world mortality endpoint was previously validated using the National Death Index, and the real-world disease progression endpoint, although following a clinician-anchored approach supported by radiology report data, was previously found to be comparable to trial RECIST-based disease progression (Griffith et al., 2019; Zhang et al., 2021; Mhatre et al., 2023). Lastly, model diagnostics indicated that the weights from the IPTW-ATT and IPCW induced balance in the measured baseline and post-baseline confounders.

There are notable limitations with this study. First, because data were pooled from disparate sources, full information was not available for all possible confounders. For example, there was limited capture of comorbidities, sites of metastasis, and smoking status within the OC arm compared to the RCT arm. The assumption of no unmeasured confounders underlies both IPCW (i.e., baseline as well as time-varying covariates jointly predicting treatment switch and outcome (Howe et al., 2011)) as well as IPTW (i.e., baseline covariates jointly predicting treatment setting and outcome). About 43% (Table 2) of the patients in routine clinical care included in our study had missing ECOG PS at the start of front-line therapy, some of whom may have had an ECOG PS value above 1. For context, among adults with NSCLC who received first-line chemotherapy in the real-world setting, 13.6% had an ECOG PS greater than 1 (Supplementary Table S3). A second limitation was that the definition of time-zero differed across the RCT and OC arms. Time-zero was the date of randomization in the clinical trials compared to the date of treatment initiation in the routine clinical practice cohort. The impact is believed to be small given that typically, treatment was initiated within a few days post-randomization. A third limitation is that patients in the IMpower trials were global while patients in the OC arm were from the United States only. Although we account for potential patient confounders in our models, there could be residual confounding effects due to regional differences. A fourth limitation was that we pooled data from the control arms of the RCTs and hence assumed negligible heterogeneity in outcomes among the three clinical-trial cohorts. However, we believe trial heterogeneity posed little bias risk to our study because the three trials were conducted by the same sponsor and had similar visit schedules, data quality monitoring, and survival estimates (Supplementary Figure S1). As a final limitation, this work does not present guidelines regarding size and power because formal hypothesis testing was not conducted during the study. A proper power analysis would need to specify and model the impact of the (time-varying) confounders on the effect size. There are limited examples applying time-varying covariate weighting in external control analyses, and guidelines on how to compute sample size are needed. Future work should aim to establish size and power guidelines to ensure quality and meaningful inferences from these types of analyses.

5 Conclusion

In conclusion, this study showed that combining the EF and TTF approaches can improve the rigor in the design and analysis of comparative effectiveness studies, including retrospective observational studies. The EF approach alone does not suffice in specifying a study design, and the TTF alone can leave ambiguity in the inferential target. The combination of the two frameworks should be considered more often by researchers.

Data availability statement

The datasets presented in this article are not readily available because the real-world data that support the findings of this study have been originated by Flatiron Health, Inc. Requests for data sharing by license or by permission for the specific purpose of replicating results in this manuscript can be submitted to publicationsdataaccess@flatiron.com. For eligible clinical trials, qualified researchers may request access to individual patient-level clinical data for each separate study through a data request platform. At the time of writing, this request platform is Vivli. https://vivli.org/ourmember/roche/. For up to date details on Roche' Global Policy on the Sharing of Clinical Information and how to request access to related clinical study documents, see here: https://go.roche.com/data_sharing. Anonymized records for individual patients across more than one data

source external to Roche cannot, and should not, be linked due to a potential increase in risk of patient re-identification.

Ethics statement

The studies involving humans were approved by the institutional review board of WCG IRB. The studies were conducted in accordance with the local legislation and institutional requirements. The ethics committee/institutional review board waived the requirement of written informed consent for participation from the participants or the participants' legal guardians/next of kin because human participants were not directly involved in the study.

Author contributions

LP, QL, PM, AS, OH, KR, and DH contributed to study design (literature, background search), NP, QL, and LP contributed to data analysis (including creation of figures/tables). All authors contributed to the article and approved the submitted version.

Funding

This study was sponsored by Flatiron Health, Inc., which is an independent member of the Roche Group.

Acknowledgments

The authors would like to thank Cody Patton, Hannah Gilham, and Darren Johnson of Flatiron Health, Inc. for publication management and editing support, as well as Somnath Sarkar and Meghna Samant of Flatiron Health, Inc. for their scientific input and support of this research. A version of this manuscript appears on arXiv as a preprint, as allowed by Frontiers in Pharmacology (Polito et al., 2022).

Conflict of interest

QL, PM, AS and OH were employed by Flatiron Health, Inc., an independent member of the Roche Group, and stock ownership in Roche. OH was employed by and owning stock in Regeneron Pharmaceuticals, Inc. LP, KR, and DH were employed by F. Hoffmann-La Roche and stock ownership in Roche. NP was employed by Genentech, Inc., and stock ownership at Roche/Genentech.

The authors declare that this study received funding from Flatiron Health, Inc. The funder was involved in the study design, collection, analysis, interpretation of data, the writing of this article, and the decision to submit it for publication.

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.

Supplementary material

The Supplementary Material for this article can be found online at: https://www.frontiersin.org/articles/10.3389/fphar.2024.1223858/full#supplementary-material

References

Austin, P. C. (2009). Balance diagnostics for comparing the distribution of baseline covariates between treatment groups in propensity-score matched samples. *Stat. Med.* 28, 3083–3107. doi:10.1002/sim.3697

Austin, P. C., and Stuart, E. A. (2015). Moving towards best practice when using inverse probability of treatment weighting (IPTW) using the propensity score to estimate causal treatment effects in observational studies. *Stat. Med.* 34, 3661–3679. doi:10.1002/sim.6607

Birnbaum, B., Nussbaum, N., Seidl-Rathkopf, K., Agrawal, M., Estevez, M., Estola, E., et al. (2020). Model-assisted cohort selection with bias analysis for generating large-scale cohorts from the EHR for oncology research. Available at: https://arxiv.org/abs/2001.09765.

Cain, L. E., Saag, M. S., Petersen, M., May, M. T., Ingle, S. M., Logan, R., et al. (2016). Using observational data to emulate a randomized trial of dynamic treatment-switching strategies: an application to antiretroviral therapy. *Int. J. Epidemiol.* 45, 2038–2049. doi:10.1093/ije/dyv295

Cole, S. R., and Hernán, M. A. (2008). Constructing inverse probability weights for marginal structural models. *Am. J. Epidemiol.* 168, 656–664. doi:10.1093/aje/kwn164

Food and Drug Administration, (2021). *E9(R1) statistical principles for clinical trials: addendum: estimands and sensitivity analysis in clinical trials; guidance for industry*. USA: U.S. Department of Health and Human Services. Available at: https://www.fda.gov/media/148473/download (Accessed May 10, 2023).

Goetghebeur, E., le Cessie, S., De Stavola, B., Moodie, E. E., and Waernbaum, I., (2020). Formulating causal questions and principled statistical answers. *Stat. Med.* 39, 4922–4948. doi:10.1002/sim.8741

Griffith, S. D., Tucker, M., Bowser, B., Calkins, G., Chang, C. J., Guardino, E., et al. (2019). Generating real-world tumor burden endpoints from electronic health record data: comparison of RECIST, radiology-anchored, and clinician-anchored approaches for abstracting real-world progression in non-small cell lung cancer. *Adv. Ther.* 36, 2122–2136. doi:10.1007/s12325-019-00970-1

Hampson, L. V., Chu, J., Zia, A., Zhang, J., Hsu, W., Parzynski, C. S., et al. (2023). Combining the target trial and estimand frameworks to define the causal estimand: an application using real-world data to contextualize a single-arm trial. Available at: https://arxiv.org/abs/2202.11968.

Hernán, M. A., Alonso, A., Logan, R., Grodstein, F., Michels, K. B., Willett, W. C., et al. (2008). Observational studies analyzed like randomized experiments: an application to postmenopausal hormone therapy and coronary heart disease. *Epidemiology* 19, 766–779. doi:10.1097/EDE.0b013e3181875e61

Hernán, M. A., and Robins, J. M. (2006). Estimating causal effects from epidemiological data. *J. Epidemiol. Community Health* 60, 578–586. doi:10.1136/jech.2004.029496

Hernán, M. A., and Robins, J. M. (2016). Using big data to emulate a target trial when a randomized trial is not available. *Am. J. Epidemiol.* 183, 758–764. doi:10.1093/aje/kwv254

Howe, C. J., Cole, S. R., Chmiel, J. S., and Muñoz, A. (2011). Limitation of inverse probability-of-censoring weights in estimating survival in the presence of strong selection bias. *Am. J. Epidemiol.* 173, 569–577. doi:10.1093/aje/kwq385

Jotte, R., Cappuzzo, F., Vynnychenko, I., Stroyakovskiy, D., Rodríguez-Abreu, D., Hussein, M., et al. (2020). Atezolizumab in combination with carboplatin and nab-paclitaxel in

advanced squamous NSCLC (IMpower131): results from a randomized phase III trial. *J. Thorac. Oncol.* 15, 1351–1360. doi:10.1016/j.jtho.2020.03.028

Liu, F., and Panagiotakos, D. (2023). Correction: real-world data: a brief review of the methods, applications, challenges and opportunities. *BMC Med. Res. Methodol.* 23, 109–111. doi:10.1186/s12874-023-01937-1

Ma, X., Long, L., Moon, S., Adamson, B., and Baxi, S. (2020). *Comparison of population characteristics in real-world clinical oncology databases in the US: flatiron Health, SEER, and NPCR. medRxiv*.

Mhatre, S. K., Machado, R. J. M., Ton, T. G. N., Trinh, H., Mazieres, J., Rittmeyer, A., et al. (2023). Real-world progression-free survival as an endpoint in lung cancer: replicating atezolizumab and docetaxel arms of the OAK trial using real-world data. *Clin. Pharmacol. Ther.* 114, 1313–1322. doi:10.1002/cpt.3045

Nishio, M., Barlesi, F., West, H., Ball, S., Bordoni, R., Cobo, M., et al. (2021). Atezolizumab plus chemotherapy for first-line treatment of nonsquamous NSCLC: results from the randomized phase 3 IMpower132 trial. *J. Thorac. Oncol.* 16, 653–664. doi:10.1016/j.jtho.2020.11.025

Petito, L. C., García-Albéniz, X., Logan, R. W., Howlader, N., Mariotto, A. B., Dahabreh, I. J., et al. (2020). Estimates of overall survival in patients with cancer receiving different treatment regimens: emulating hypothetical target trials in the surveillance, epidemiology, and end results (SEER)-Medicare linked database. *JAMA Netw. Open* 3, e200452. doi:10.1001/jamanetworkopen.2020.0452

Polito, L., Liang, Q., Pal, N., Mpofu, P., Sawas, A., Humblet, O., et al. (2022). Applying the Estimand and Target Trial frameworks to external control analyses using observational data: a case study in the solid tumor setting. Available at: https://arxiv.org/abs/2208.06707.

Potter, F., and Zheng, Y. (2015). Methods and issues in trimming extreme weights in sample surveys. Available at: http://www.asasrms.org/Proceedings/y2015/files/234115.pdf (Accessed May 10, 2023).

Schardt, C., Adams, M. B., Owens, T., Keitz, S., and Fontelo, P. (2007). Utilization of the PICO framework to improve searching PubMed for clinical questions. *BMC Med. Inf. Decis. Mak.* 7, 16. doi:10.1186/1472-6947-7-16

Schaubel, D. E., and Wei, G. (2011). Double inverse-weighted estimation of cumulative treatment effects under nonproportional hazards and dependent censoring. *Biometrics* 67, 29–38. doi:10.1111/j.1541-0420.2010.01449.x

Signorovitch, J., Moshyk, A., Zhao, J., Le, T. K., Burns, L., Gooden, K., et al. (2022). Overall survival in the real-world and clinical trials: a case study validating external controls in advanced melanoma. *Future Oncol.* 18, 1321–1331. doi:10.2217/fon-2021-1054

Thomas, D. S., Lee, A. Y., Müller, P. L., Schwartz, R., Olvera-Barrios, A., Warwick, A. N., et al. (2021). Contextualizing single-arm trials with real-world data: an emulated target trial comparing therapies for neovascular age-related macular degeneration. *Clin. Transl. Sci.* 14, 1166–1175. doi:10.1111/cts.12974

West, H., McCleod, M., Hussein, M., Morabito, A., Rittmeyer, A., Conter, H. J., et al. (2019). Atezolizumab in combination with carboplatin plus nab-paclitaxel chemotherapy compared with chemotherapy alone as first-line treatment for metastatic non-squamous non-small-cell lung cancer (IMpower130): a multicentre, randomised, open-label, phase 3 trial. *Lancet Oncol.* 20, 924–937. doi:10.1016/S1470-2045(19)30167-6

Zhang, Q., Gossai, A., Monroe, S., Nussbaum, N. C., and Parrinello, C. M. (2021). Validation analysis of a composite real-world mortality endpoint for patients with cancer in the United States. *Health Serv. Res.* 56, 1281–1287. doi:10.1111/1475-6773.13669

Frontiers in Pharmacology

Explores the interactions between chemicals and living beings

The most cited journal in its field, which advances access to pharmacological discoveries to prevent and treat human disease.

Discover the latest Research Topics

See more →

Frontiers

Avenue du Tribunal-Fédéral 34
1005 Lausanne, Switzerland
frontiersin.org

Contact us

+41 (0)21 510 17 00
frontiersin.org/about/contact