BMJ Health &
Care Informatics

# Bibliometric analysis of the 3-year trends (2018–2021) in literature on artificial intelligence in ophthalmology and vision sciences

Hayley Monson ,[1] Jeffrey Demaine ,[2] Adrianna Perryman,[3] Tina Felfeli [4,5]

**To cite:** Monson H, Demaine J, Perryman A, *et al*. Bibliometric analysis of the 3-year trends (2018–2021) in literature on artificial intelligence in ophthalmology and vision sciences. *BMJ Health Care Inform* 2024;**31**:e100780. doi:10.1136/bmjhci-2023-100780

► Additional supplemental material is published online only. To view, please visit the journal online (https://doi.org/10.1136/bmjhci-2023-100780).

Received 03 April 2023
Accepted 31 January 2024

[1]Mathematics, McMaster University, Hamilton, Ontario, Canada
[2]Research Impact Services, McMaster University, Hamilton, Ontario, Canada
[3]Global Health, York University, Toronto, Ontario, Canada
[4]Department of Ophthalmology and Vision Sciences, University of Toronto, Toronto, Ontario, Canada
[5]Institute of Health Policy, Management and Evulation, University of Toronto, Toronto, Ontario, Canada

**Correspondence to**
Dr Tina Felfeli;
tina.felfeli@mail.utoronto.ca

## ABSTRACT

**Objectives** The objective of this analysis is to present a current view of the field of ophthalmology and vision research and artificial intelligence (AI) from topical and geographical perspectives. This will clarify the direction of the field in the future and aid clinicians in adapting to new technological developments.

**Methods** A comprehensive search of four different databases was conducted. Statistical and bibliometric analysis were done to characterise the literature. Softwares used included the R Studio bibliometrix package, and VOSviewer.

**Results** A total of 3939 articles were included in the final bibliometric analysis. Diabetic retinopathy (391, 6% of the top 100 keywords) was the most frequently occurring indexed keyword by a large margin. The highest impact literature was produced by the least populated countries and in those countries who collaborate internationally. This was confirmed via a hypothesis test where no correlation was found between gross number of published articles and average number of citations (p value=0.866, r=0.038), while graphing ratio of international collaboration against average citations produced a positive correlation (r=0.283). Majority of publications were found to be concentrated in journals specialising in vision and computer science, with this category of journals having the highest number of publications per journal (18.00 publications/journal), though they represented a small proportion of the total journals (<1%).

**Conclusion** This study provides a unique characterisation of the literature at the intersection of AI and ophthalmology and presents correlations between article impact and geography, in addition to summarising popular research topics.

### WHAT IS ALREADY KNOWN ON THIS TOPIC

⇒ Bibliometric analysis as a method of characterising research in a field has become increasingly popular in recent years. Some bibliometric analyses on the body of ophthalmological literature have been published in specialised areas, as well as a small number in the intersection of artificial intelligence (AI) and ophthalmology.

### WHAT THIS STUDY ADDS

⇒ This study will provide a more recent and comprehensive profile of the intersection of AI and ophthalmology than previous studies, as well as examining a broader range of subspecialties and data sources.

### HOW THIS STUDY MIGHT AFFECT RESEARCH, PRACTICE OR POLICY

⇒ A better understanding of the existing literature on AI will provide insight into the growing influence of AI on ophthalmology, and will allow medical researchers and academics to anticipate emerging areas of research and allocate funds more effectively.

## INTRODUCTION

Coined over 60 years ago by McCarthy and Minsky, the term artificial intelligence (AI) refers to the ability of a computer system to complete complex tasks normally requiring human abilities.[1] The popularity of this idea has grown in medicine in recent years as there is great potential for the increase in the efficiency of medical systems via AI, particularly in the areas of visual processing for diagnosis and determination of treatment pathways. To date, AI has been applied to ophthalmology with great efficacy in diagnosis of common diseases such as diabetic retinopathy, retinopathy of prematurity, glaucoma and macular degeneration.[2] A review by Grzybowski *et al* suggested that recent diagnostic software for diabetic retinopathy demonstrated a sensitivity of 87.0% and a specificity of 96.8%.[3]

Bibliometric analysis as a method of characterising research in a field has become increasingly popular in recent years.[4] Previously published bibliometric analyses in ophthalmology and intersecting fields include an analysis on uveal melanoma literature, and keratoconus.[5 6] In particular, the growing use of AI in ophthalmology has been profiled by AlRyalat *et al* who performed a comparative bibliometric analysis between the fields of glaucoma research and AI.[7] Boudry *et al* have also demonstrated the growth of AI in the field of ophthalmology over several decades between 1966 and 2019.[8]

Here, we aim to provide a bibliometric profile of the intersection of ophthalmology and AI. Our study complements previous studies in this area by examining a more recent timeframe (2018 to August 2021) across a broader range of data sources and all subspecialties in ophthalmology. A better understanding of the existing literature on AI will provide insight into the growing influence and importance of AI on the field of ophthalmology. This will allow medical researchers and academics to anticipate emerging areas of research and allocate funds more effectively, to seek out research partners and institutions with common interests, and will allow the medical community to adapt to new technologies and integrate them into the future model of patient care.

## METHODS

This is a bibliometric analysis of articles relating to AI technology and ophthalmology and vision research. A detailed review of the bibliometric analysis study methods is reported elsewhere.[9] The protocol for this study was also prospectively registered on Open Science Framework registry (https://doi.org/10.17605/OSF.IO/BZ9YJ).

### Search strategy

A comprehensive search was conducted in Web of Science, Scopus, Dimensions and Cochrane from 1 January 2018 up to 4 August 2021. These specific databases were chosen as they encompass a wide selection of journals and articles pertaining to the selected topics and are compatible with a wide variety of bibliometric analytic softwares.[10 11] A 3-year timeline for the citation analysis was chosen with regard to the feasibility of analyses as well as its focused overview of the latest and most relevant technology in AI and ophthalmology. Search strategy keywords were carefully selected from relevant literature and online medical and computer science glossaries to ensure only relevant documents were analysed. No language or study design restrictions were placed on the search strategy. The details of the search query are provided in online supplemental file 1.

### Screening

All citations were uploaded to the DistillerSR software and deduplicated.[12] Following de-duplication, all articles were screened by title and abstract by a single reviewer for relevance. More information on the methods of extraction and data-cleaning processes are included in online supplemental file 2. Only articles directly pertaining to the field of ophthalmology and AI were included, and given that each article had to meet certain search criteria to be included in the preliminary dataset, articles passing the screening either clearly fell within the scope of ophthalmology and AI or did not.

### Analytic methods

Several analytic methods were applied to this dataset to elucidate the present focus of the field and its future direction. Preliminary analyses were applied to the dataset using RStudio to obtain the number of articles and mean number of citations per year. Then charts displaying the most popular journals and countries and their gross publications were produced. Journals were categorised by topic and then an analysis was conducted using Excel. The journals contained in the dataset were categorised as belonging to medicine (M), vision (V), computer science (CS), engineering (E), artificial intelligence (AI) and general science (G). Journals belonging to both medicine and computer science were labelled as intersectional (I). A metric measuring average publications per journal, and by extension the significance of that journal in the field, was calculated by summing all the articles and then dividing by the number of journals in that category. This value corresponds to the average number of articles per journal in that category.

The international distribution of the publications was analysed. The raw number of publications per country was extracted along with the number of mean citations in the literature for each country. The countries were ranked by the number of publications, the number of citations to those publications and the average number of citations per publication based on the principal investigator. A statistical analysis was performed on the dataset to investigate if a statistically significant correlation existed between gross number of publications by a country and their average number of citations.

The data including the countries, their total number of articles published, and their average citations was exported, and a citation network was created using the VOSviewer software. A statistical analysis comparing countries by their published output and its average citations was performed. This was done via a Spearman rank correlation test. The null hypothesis ($H_0$) was that there is no correlation between the number of publications produced by a country and the average number of citations received by those publications (ie, that the value of r is 0). Further, single country publication (a ratio representative of the proportion of total publications with intra-national collaborations) and multiple country publication (MCP, the proportion of total publications with international collaborators) ratios were used to investigate the linkage between international collaboration and rate of citation. Average citations by country were graphed against MCP to see if correlation between the two variables could be established.

Author keywords were extracted, and a co-occurrence map was created with all words with a minimum of five connections to others. A link between words is established if two keywords are listed in conjunction by more than one author. The number of occurrences of each keyword was represented by the size of the nodes.

## RESULTS

From the initial search, 5917 articles were obtained from Dimensions, 5771 from Scopus, 3717 from Web of

Table 1 Number of journals and articles in each category

| Category | Journals (n) | Articles (n) |
|---|---|---|
| Medicine (M) | 371 | 949 |
| Vision (V) | 128 | 1454 |
| Computer science (CS) | 141 | 446 |
| Engineering (E) | 49 | 182 |
| Artificial intelligence (AI) | 47 | 124 |
| General (G)<br>nature, science, etc | 120 | 306 |
| Intersection of CS and medicine (I) | 95 | 667 |

Science and 136 from Cochrane. Following deduplication, and screening, 3939 articles were included in the analysis, with 433 articles collected from 2018, 697 articles from 2019, 1416 from 2020 and 1393 from 2021.

The number of journals and articles in each discrete category is summarised in table 1. The highest number of articles were categorised as medicine, with computer science being second and vision being a close third. No journals were categorised as specialising in vision and AI, while only two journals were categorised as specialising in vision and computer science. Vision and computer science had the highest average number of publications (18.00 publications/journal), although it accounted for less than 1% of the total journals. The second highest average number of publications was in the vision (V) category, with 11.36 publications/journal. General medical journals (M), while accounted for the highest number of journals, had only 2.73 publications/journal whereas medical and computer science journals had an average of 8.19 publications/journal. The top three journals were *Translational Vision Sciences and Technology*, categorised as vision with 133 articles; *Scientific Reports* categorised as general science with 129 articles; and *IEEE Access* categorised as engineering with 120 articles. Below, we present the top five articles from *IEEE Access*, the engineering journal with the greatest number of publications, to exemplify the growing popularity of the field of ophthalmology and AI outside of medicine.

Based on corresponding authors' affiliations, China (946, 25%) and the USA (719, 19%) produced the most number of publications overall (table 2). The rest of the publications came from a wide range of countries

Table 2 Countries ranked in order of most publications, accompanied by citation data

| Publication rank | Citation rank | Average article citations | Corresponding author's country | Publications | Total citations | Average article citations |
|---|---|---|---|---|---|---|
| 1 | 2 | 11 | China | 946 | 7769 | 8.21 |
| 2 | 1 | 6 | USA | 719 | 8108 | 11.28 |
| 3 | 4 | 21 | India | 367 | 1894 | 5.16 |
| 4 | 6 | 16 | Korea | 178 | 1190 | 6.69 |
| 5 | 3 | 3 | UK | 150 | 2254 | 15.03 |
| 6 | 8 | 13 | Japan | 134 | 998 | 7.45 |
| 7 | 9 | 10 | Germany | 106 | 871 | 8.22 |
| 8 | 11 | 14 | Spain | 106 | 747 | 7.05 |
| 9 | 5 | 2 | Singapore | 95 | 1460 | 15.37 |
| 10 | 7 | 5 | Australia | 94 | 1116 | 11.87 |
| 11 | 14 | 23 | Turkey | 85 | 372 | 4.38 |
| 12 | 13 | 20 | Italy | 82 | 466 | 5.68 |
| 13 | 10 | 4 | Canada | 52 | 772 | 14.85 |
| 14 | 15 | 17 | Brazil | 51 | 329 | 6.45 |
| 15 | 17 | 19 | France | 51 | 311 | 6.10 |
| 16 | 18 | 18 | Iran | 47 | 291 | 6.19 |
| 17 | 12 | 1 | Austria | 42 | 742 | 17.67 |
| 18 | 19 | 12 | Pakistan | 34 | 259 | 7.62 |
| 19 | 21 | 15 | Saudi Arabia | 34 | 235 | 6.91 |
| 20 | 16 | 8 | Netherlands | 33 | 312 | 9.46 |
| 21 | 23 | 22 | Egypt | 26 | 127 | 4.89 |
| 22 | 20 | 7 | Switzerland | 26 | 252 | 9.69 |
| 25 | 22 | 9 | Portugal | 24 | 226 | 9.42 |

in Europe and Asia, with no country (aside from India) accounting for more than 5% of the total number of publications (figure 1). Austria had the highest average article citations, collaborated with authors from nine different countries, had 42 articles by corresponding authors, and 138 total publications. China collaborated with 17 distinct countries, had 946 articles by corresponding authors and had 2911 total publications (figure 2). When comparing countries by their published output and average citations, the findings did not reveal a significant correlation (p value=0.866, r=0.038). This suggests that there is no statistically significant correlation between gross amount of literature published by a country and average number of articles citations for that country, which is a surrogate metric for literature quality.

Austria had a higher MCP/total fraction, at 0.4762, as compared with China, which had an MCP/total fraction of 0.243. Plotting countries by their average citations per publication against their proportion of international collaborations yielded a weakly positive correlation coefficient of $R^2$=0.283 (figure 3). This suggests that there is association between number of international collaborators and global popularity of literature.

The top five most frequent indexed keywords included 'deep learning' (677, 11%), 'diabetic retinopathy' (391, 6%), 'machine learning' (364,6%), 'artificial intelligence' (332, 5%) and 'optical coherence tomography' (311, 5%, figure 4). Diabetic retinopathy was the most frequently occurring ophthalmological disease by a margin of 291 occurrences (5% of the top 100 occurrences), with 'age-related macular degeneration' being the next most frequently occurring ophthalmic disease at only 100 occurrences.

## DISCUSSION

We conducted a bibliometric analysis of the intersection of ophthalmology and AI between January 2018 and August 2021. Many aspects of the dataset were analysed in order to gain both quantitative and qualitative insights. In particular, investigation into countries of publication and their correlation (or lack thereof) with literature quality was performed, and it was found that smaller countries tended to produce more highly cited literature. There was a direct correlation between country population and gross quantity of published literature. Furthermore, countries with more international collaboration tended to have higher average article citations. With respect to research topics, the most common application of the AI technology to ophthalmology tended to be in diagnostic imaging.

Our findings suggested that the field of ophthalmology and AI has been growing at an exponential rate as predicted by Lotka's law until 2020 when the scientific production dropped sharply.[13] The authors hypothesise that there are two main reasons for this finding. First, it is likely that SARS-CoV-2 affected scientific production in the field of ophthalmology and AI as the broad scientific community shifted to focus on developing a body of research on the novel virus. Second, articles were only collected up to August 2021, and had the articles been collected up to December it is predicted that the growth rate of the field would have increased rather than decreased, though likely not with the same increase in rate as in previous years.

It was noted in our analysis that China and the USA collectively account for over 40% of the literature in the dataset. This is not surprising in consideration of the population size and large number of research institutions in both countries. Within the dataset there is an over-representation in the advanced economies of Southeast Asia, where Japan, Korea and Singapore accounted for more research in this field than the UK and Germany combined.

Popular AI ranking indices have consistently placed the USA and China at the top of research, development and implementation of new AI technologies over the past 5 years, with Japan and Korea ranking in the top 10.[14 15] According to the Stanford AI index, in 2021, East Asia accounted for 26.7% of all published academic articles pertaining to AI globally, while the USA accounted for 14.0%.[14 15] Further, global AI publications have seen a steep growth curve recently, with total international journal publications having increased 2.5 times since 2015. This rapid growth is seen in conjunction with an exponential increase in AI patent filings globally, with a compound annual growth rate of 76.9% between 2015 and 2021.[16] As more research is published, more innovation is spurred, while new technology promotes new research, in a positive and fast accelerating feedback loop. In 2021, China held the greatest number of AI patent filings, while the USA had the most granted patents as a percentage of the world total filed and granted patents.[16]

We have used the number of citations as a measurement of literature impact. Previous studies have suggested that the correlation between citation numbers and value of scientific knowledge and influence is not perfect, and citations might also be influenced by factors such as author prominence and randomness.[17] Although, there are important factors that should be considered when using number of citations as an absolute measure of literature quality,[17] the large size of our data set may give an accurate overall picture of global impact.[18] Our findings showed no statistically significant correlation between the gross number of publications for a country and mean number of citations. This result indicates that while China and the USA may produce nearly half of the articles in this field, they do not also attract the most citations. Our findings suggested that research from countries such as Austria, had the most citations per publication and high proportional international collaboration than China. It is well-established for scientometric characteristics that collaboration between institutions, in particular internationally, tends to produce research that is cited more frequently than less-collaborative work.[19] As such China and the USA, although produce most publications they

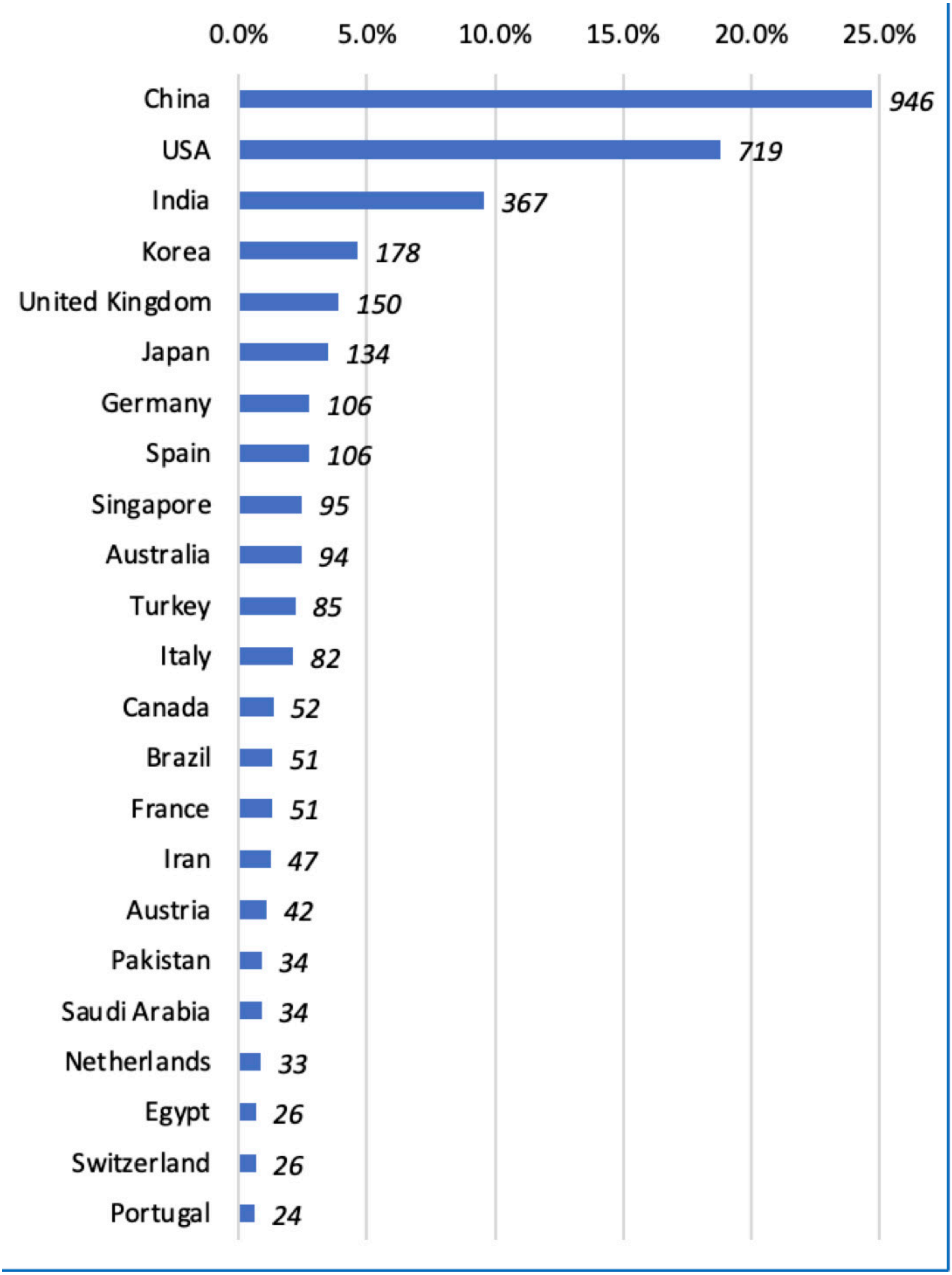

**Figure 1** Breakdown of percentage of total number of publications identified based on the country of the corresponding author.

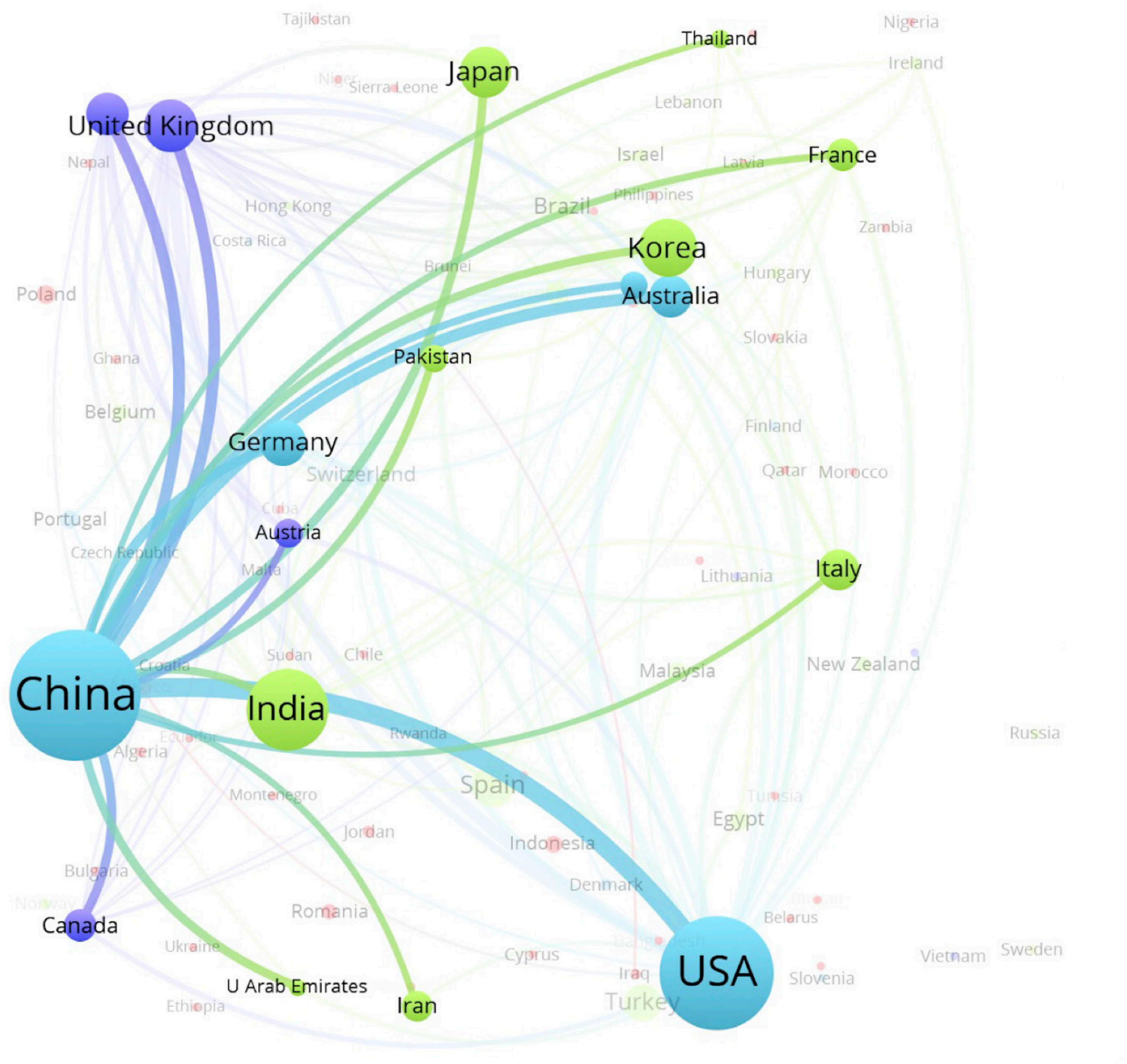

**Figure 2** Countries were clustered via unique colours representing the average number of citations for that country. Purple countries had the highest average citations (>12), light blue countries had between 8 and 12 average citations, light green countries had between 4 and 8, and red countries had the fewest, between 0 and 4. The sizes of the country names indicate their gross number of publications, the larger the label being correlated with the total number of publications for that country. Links between countries indicate which tend to collaborate, and the thickness of the linkage corresponds to the strength of the connection. Countries which collaborate on many papers will have a thicker connecting line. Links between countries are only displayed if there has been a minimum of five collaborative publications.

tend to collaborate less with institutions in other countries. The reasons behind this effect are multi-faceted and beyond the scope of this paper. Besides the cultural and geographic factors that would limit their international connections, both China and the USA have many universities within their own borders with whom to collaborate. In contrast, the high impact of smaller countries such as Singapore and Austria are surrounded by many other countries to collaborate with and have some of the highest citations-per-publication alongside a high proportion of MCPs.

We noted that the most collaborative countries, as well as those with the highest average citation impact, tend to be smaller countries in Europe with the exception of Singapore. As an Asian city-state with a British colonial heritage, Singapore's cultural-linguistic connections both to Europe and to South-East Asia enable it to have the second-highest

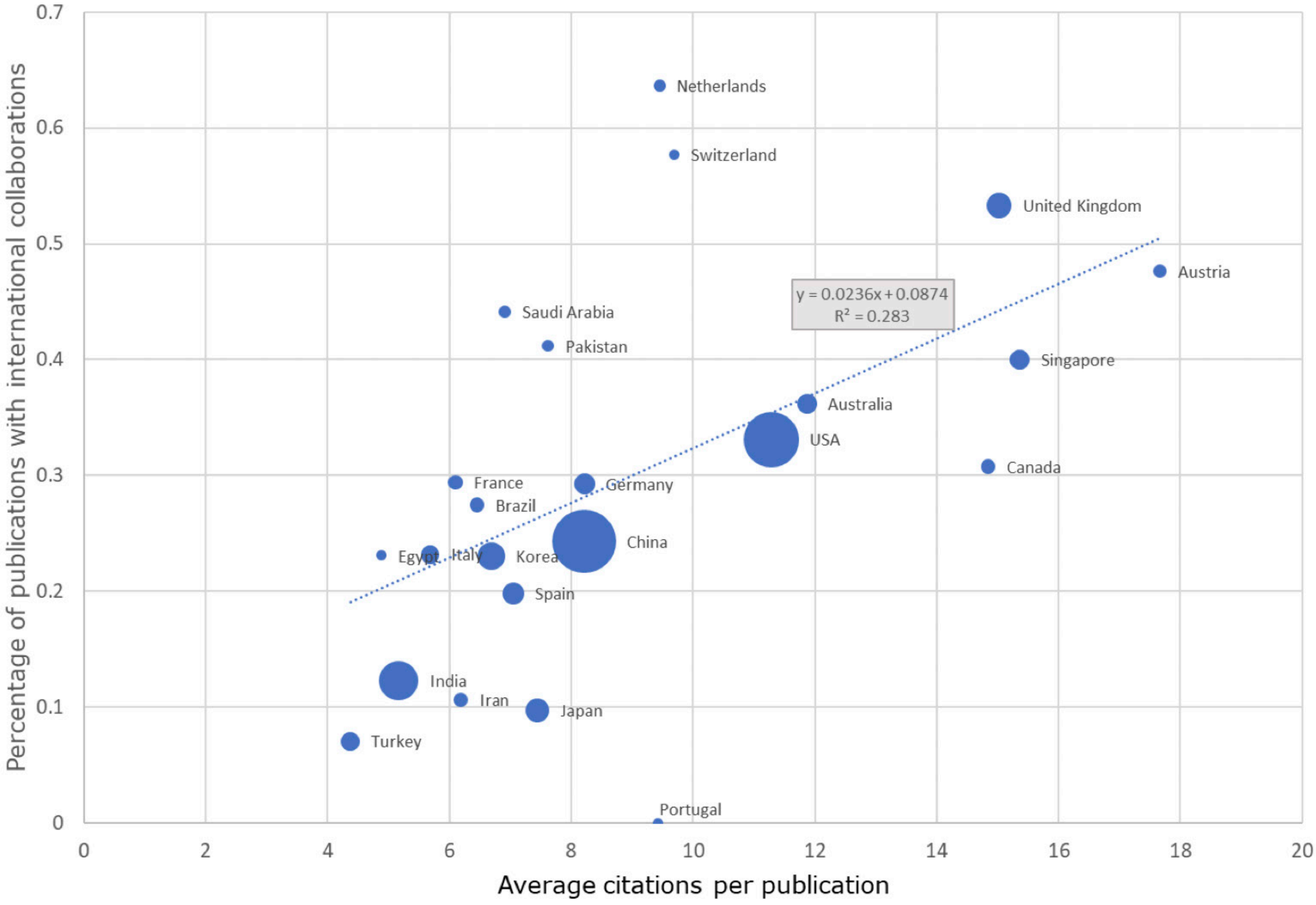

**Figure 3** A plot depicting countries by their average citations per publication against their proportion of international collaborations.

citations-per-paper of all the countries in this survey, showing how collaborations are more important than size. We also found that while China is the most productive country, it lags behind the only other country of comparable output (the USA) which tends to have more international collaborations. This is corroborated by two popular AI index reports, which find that while China leads the USA in gross publications, the USA 'leads on the most significant research into cutting-edge developments'.[14–16]

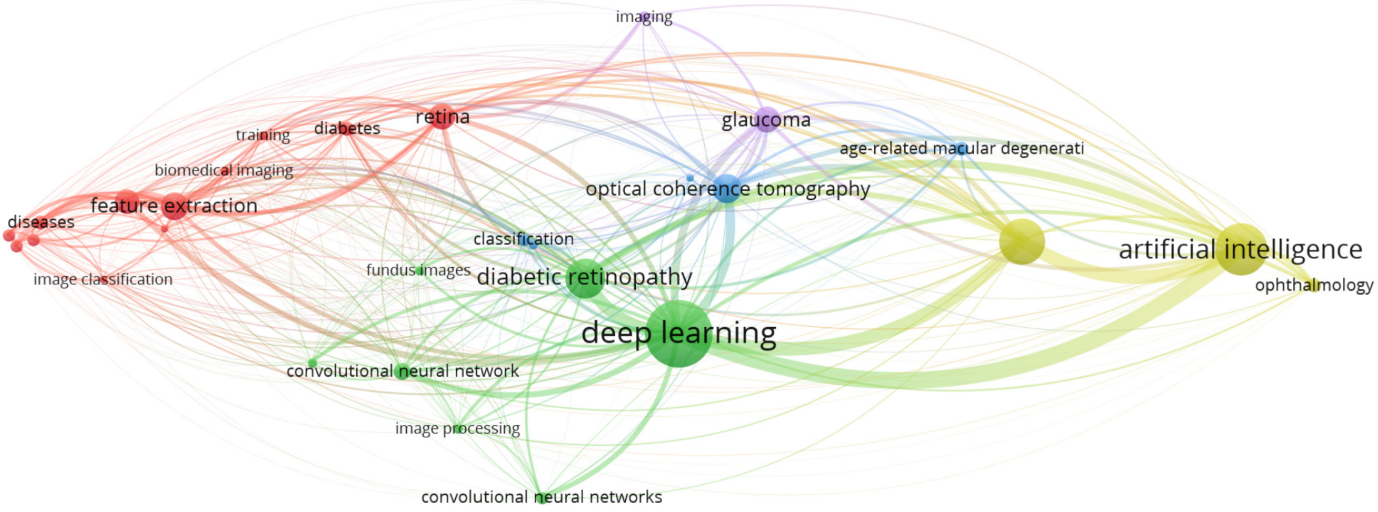

**Figure 4** A co-occurrence network showing the top 20 keywords among all listed author keywords. Larger nodes correspond to a higher number of occurrences of that keyword, thicker connections indicate a higher frequency of two keywords being listed together.

From the co-occurrence network created diabetic retinopathy is most connected with the terms 'deep learning', 'machine learning' and 'artificial intelligence'. Further, other popular terms relate to types of diagnostic imaging, such as 'optical coherence tomography' and 'image segmentation'. This implies that the focus of the field is on applications of AI to diagnosis, and creation of algorithms for automating diagnosis and triage of ophthalmic diseases. Many medical fields follow a progression of care model, where diagnosis is the first step, followed by prognostication, development and administration of treatment protocols, and surgical management if necessary. As such, new technology may begin to develop first in the areas of need, in the case of the field of ophthalmology this is diagnosis and triage. Additionally, there is more cost and resource associated with research in robotics than computer research.[20]

## CONCLUSION

This paper presents an in-depth bibliometric analysis of literature in the field of ophthalmology and AI. Articles were collected from a wide variety of sources over a 3-year time period in order to gain a detailed perspective on the current state of the technology and its future trajectory. We have characterised the field via both qualitative and quantitative methods. We have investigated trends in topics in the field, and which varieties of research are currently gaining the most traction and may have practical application in the near future. We have determined that the USA and China together produce the highest volume of research, though they have among the lowest rates of international collaboration, while smaller countries with high rates of international collaboration such as Singapore and Austria produce the most cited research. Increasing international collaborations may be an effective way for geographic areas which are behind in this field to strengthen their body of research in AI and ophthalmology. Encouraging researchers to provide open source access to research, particularly to newly developed code for AI algorithms, can aid in increasing participation and collaboration from previously dormant countries. These findings will aid the ophthalmology medical and research community in adapting their practices to the changing landscape of vision care.

**Twitter** Tina Felfeli @TinaFelfeli

**Contributors** Conception and design: TF. Acquisition of data: HM, TF, AP. Data analysis: HM, JD. Interpretation of data: HM, JD, TF. First draft of the article: HM, JD, TF. Critical revision: HM, JD, TF. Final approval of the version to be published: HM, JD, AP, TF. Act as guarantor of the work: TF.

**Funding** Funding for the publication of this study was provided by Fighting Blindness Canada, Clinician Scientist Emerging Leader Award given to Dr. Tina Felfeli.

**Competing interests** None declared.

**Patient consent for publication** Not applicable.

**Ethics approval** Ethics approval from our Institutional Review Board was not required as this is a review of published studies and does not involve human subjects.

**Provenance and peer review** Not commissioned; externally peer reviewed.

**Data availability statement** Data are available upon reasonable request. The data is available upon request sent to the corresponding author.

**Supplemental material** This content has been supplied by the author(s). It has not been vetted by BMJ Publishing Group Limited (BMJ) and may not have been peer-reviewed. Any opinions or recommendations discussed are solely those of the author(s) and are not endorsed by BMJ. BMJ disclaims all liability and responsibility arising from any reliance placed on the content. Where the content includes any translated material, BMJ does not warrant the accuracy and reliability of the translations (including but not limited to local regulations, clinical guidelines, terminology, drug names and drug dosages), and is not responsible for any error and/or omissions arising from translation and adaptation or otherwise.

**ORCID iDs**

Hayley Monson http://orcid.org/0000-0001-5460-7893
Jeffrey Demaine http://orcid.org/0000-0003-4586-1317
Tina Felfeli http://orcid.org/0000-0002-0927-3086

## REFERENCES

1 Anyoha R. The history of artificial intelligence - science in the news. Harvard graduate school of arts and sciences; 2017. 1.
2 Lee A, Taylor P, Kalpathy-Cramer J, *et al*. Machine learning has arrived! *Ophthalmology* 2017;124:1726–8.
3 Grzybowski A, Brona P, Lim G, *et al*. Artificial intelligence for diabetic retinopathy screening: a review. *Eye (Lond)* 2020;34:451–60.
4 González-Alcaide G. Bibliometric studies outside the information science and library science field: uncontainable or uncontrollable. *Scientometrics* 2021;126:6837–70.
5 Li S, Guo Y, Hou X, *et al*. Mapping research trends of uveal melanoma: a bibliometric analysis. *Int Ophthalmol* 2022;42:1121–31.
6 Efron N, Morgan PB, Jones LW, *et al*. Bibliometric analysis of the keratoconus literature. *Clin Exp Optom* 2022;105:372–7.
7 AlRyalat SA, Al-Ryalat N, Ryalat S. Machine learning in glaucoma: a Bibliometric analysis comparing computer science and medical fields' research. *Expert Review of Ophthalmology* 2021;16:511–5.
8 Boudry C, Al Hajj H, Arnould L, *et al*. Analysis of international publication trends in artificial intelligence in ophthalmology. *Graefes Arch Clin Exp Ophthalmol* 2022;260:1779–88.
9 Monson H, Demaine J, Banfield L, *et al*. Three-year trends in literature on artificial intelligence in ophthalmology and vision sciences: a protocol for bibliometric analysis. *BMJ Health Care Inform* 2022;29:e100594. 10.1136/bmjhci-2022-100594 Available: https://informatics.bmj.com/lookup/doi/10.1136/bmjhci-2022-100594
10 Aria M, Cuccurullo C. Bibliometrix: an R-tool for comprehensive science mapping analysis. *J Inf* 2017;11:959–75.
11 R Core Team. R: A language and environment for statistical computing. R foundation for statistical computing, Vienna, Austria. 2021. Available: https://www.R-project.org/
12 DistillerSR. Version 2.35. Distillersr Inc. 2023. Available: https://www.distillersr.com/
13 Lotka AJ. The frequency distribution of scientific productivity. *J Washingt Acad Sci* 1926;16:317–23. Available: https://www.jst. J Washingt Acad Sci. 1926;16(12):317–23
14 Zhang D, Mishra S, Brynjolfsson E, *et al*. The AI index 2021 annual report. AI Index Steering Committee, Human-Centered AI Institute, Stanford University: Stanford, CA; 2021.
15 Mostrous A, White J, Cesareo S. The global artificial intelligence. Tortoise Media; 2023.
16 Zhang D, Maslei N, Brynjolfsson E, *et al*. The AI index 2022 annual report. AI Index Steering Committee, Stanford Institute for Human-Centered AI, Stanford University: Stanford, CA; 2022.

17 Clancy M. Do academic citations measure the impact of new ideas? New things under the sun. 2022. Available: https://www.newthingsunderthesun.com/pub/ko1l8fgf

18 Phelan TJ. A compendium of issues for citation analysis. *Scientometrics* 1999;45:117–36.

19 Larivière V, Haustein S, Börner K. Long-distance interdisciplinarity leads to higher scientific impact. *PLoS One* 2015;10:e0122565.

20 Barbash GI, Glied SA. New technology and health care costs — the case of robot-assisted surgery. *N Engl J Med* 2010;363:701–4.

BMJ Health & Care Informatics

# Seamless EMR data access: Integrated governance, digital health and the OMOP-CDM

Christine Mary Hallinan ,[1] Roger Ward,[1] Graeme K Hart,[2] Clair Sullivan,[3] Nicole Pratt,[4] Ashley P Ng ,[5,6] Daniel Capurro,[2,7] Anton Van Der Vegt,[8] Siaw-Teng Liaw ,[9] Oliver Daly,[2] Blanca Gallego Luxan,[10] David Bunker,[8] Douglas Boyle[1]

**To cite:** Hallinan CM, Ward R, Hart GK, *et al.* Seamless EMR data access: Integrated governance, digital health and the OMOP-CDM. *BMJ Health Care Inform* 2024;**31**:e100953. doi:10.1136/bmjhci-2023-100953

Received 29 October 2023
Accepted 14 January 2024

For numbered affiliations see end of article.

**Correspondence to**
Dr Christine Mary Hallinan; hallinan@unimelb.edu.au

Douglas Boyle; dboyle@unimelb.edu.au

## ABSTRACT

**Objectives** In this overview, we describe theObservational Medical Outcomes Partnership Common Data Model (OMOP-CDM), the established governance processes employed in EMR data repositories, and demonstrate how OMOP transformed data provides a lever for more efficient and secure access to electronic medical record (EMR) data by health service providers and researchers.

**Methods** Through pseudonymisation and common data quality assessments, the OMOP-CDM provides a robust framework for converting complex EMR data into a standardised format. This allows for the creation of shared end-to-end analysis packages without the need for direct data exchange, thereby enhancing data security and privacy. By securely sharing de-identified and aggregated data and conducting analyses across multiple OMOP-converted databases, patient-level data is securely firewalled within its respective local site.

**Results** By simplifying data management processes and governance, and through the promotion of interoperability, the OMOP-CDM supports a wide range of clinical, epidemiological, and translational research projects, as well as health service operational reporting.

**Discussion** Adoption of the OMOP-CDM internationally and locally enables conversion of vast amounts of complex, and heterogeneous EMR data into a standardised structured data model, simplifies governance processes, and facilitates rapid repeatable cross-institution analysis through shared end-to-end analysis packages, without the sharing of data.

**Conclusion** The adoption of the OMOP-CDM has the potential to transform health data analytics by providing a common platform for analysing EMR data across diverse healthcare settings.

## INTRODUCTION

### The Observational Medical Outcomes Partnership Common Data Model

Adoption of the Observational Medical Outcomes Partnership Common Data Model (OMOP-CDM) internationally and in Australia has enabled the conversion of vast amounts of complex, and heterogeneous electronic medical record (EMR) data into a standardised structured data model. The conversion of data has the potential to provide hospitals, health departments, auditors, regulators and universities valuable insights tailored to each institution's needs, both for operational and research purposes. This is achievable as long as the secure utilisation of an institution's EMR clinical and administrative data for purposes beyond its initial collection, known as 'secondary use', is effectively managed and employed.

Such data can be transformative, especially if used to monitor, evaluate and audit healthcare to improve clinical practice, reduce inefficiencies, contribute to the evidence base and develop a 'learning healthcare system' for improved patient care.[1–4] However, this potential is often not realised due to the inherent complexity of EMR databases—that comprise thousands of data elements across thousands of proprietary tables—where vast amount of data needs to be transformed, cleaned and restructured to make it 'fit' for 'secondary use'.[5] For highly powered collaborative research, where large volumes of EMR data are combined, use is further constrained by the heterogeneity of each institution's EMR schema[6]; concern over data sharing and privacy breaches and lack of clarity over governance and consent.[7]

The Observational Health Data Sciences and Informatics (OHDSI) consortium[8] is addressing these challenges through the transformation of each EMR database into the open-source OMOP-CDM, where EMR data elements are translated into the OMOP-CDM using standardised terminologies such as SNOMED-CT,[9] LOINC[10] or RxNORM.[11] Importantly, these transformed data are also able to be securely stored within their dedicated environment, complete with the necessary validation, analysis and reporting tools.[12] Given the OMOP-CDM is 'open source', the original source code is freely available to the public. This allows anyone to view, use, modify and distribute the software's source code which fosters collaboration and community-driven development. This 'open-source' approach promotes transparency, innovation and widespread accessibility.

### The utility and adoption of the OMOP-CDM

An increasing number of Australian and international organisations are transforming their EMR data into the OMOP-CDM as these converted databases provide health services and researchers a valuable data source to monitoring health service utilisation, contribute to the evidence base through research and develop clinical decision support systems to improve quality of care. Furthermore, it enables researchers to 'scale-up' and 'de-risk' collaborative research, by securely sharing deidentified and aggregated data and executing analyses across multiple OMOP-converted databases, ensuring that patient-level data remains securely firewalled within its respective local site.[12]

The adoption of OMOP-CDM has been on the rise globally, with the conversion of approximately 12% of EMRs worldwide by 2022, which encompasses data from 453 databases, that accounts for more than 928 million unique patient records across 41 countries.[12] This substantial adoption demonstrates the recognition of OMOP-CDM's utility in leveraging EMR data for various purposes.

An Australian OHDSI Chapter has been established to support the use of OMOP and develop collaborations between database stakeholders. OMOP members include clinicians and researchers from the University of Melbourne, the University of South Australia, the University of Queensland and the University of New South Wales and Western Australia.[13] The Australian databases that have undergone OMOP-CDM conversion include those that contain data from large tertiary hospitals in major cities, specialised hospitals that hold data for children's and cancer care services, joint replacement registries, Australian Electronic Practice-Based Research Network (AU-ePBRN),[14] local health district databases, the Primary Care Audit, Teaching and Research Open Network (PATRON) database[15], pharmaceutical registries, and the Australian Department of Veterans Affairs.[12] However, it is important to acknowledge that this progress is not without its limitations. Currently, there exists a gap in data integration, notably the absence of a seamless linkage between hospital and primary care data OMOP data sources. Despite the comprehensive approach to data integration across various healthcare contexts, the lack of connectivity between these crucial components of the healthcare system represents a constraint. This limitation highlights an area for potential improvement in Australia's data infrastructure. Addressing this gap and establishing effective linkage between hospital and primary care data could lead to even more comprehensive and impactful research outcomes.

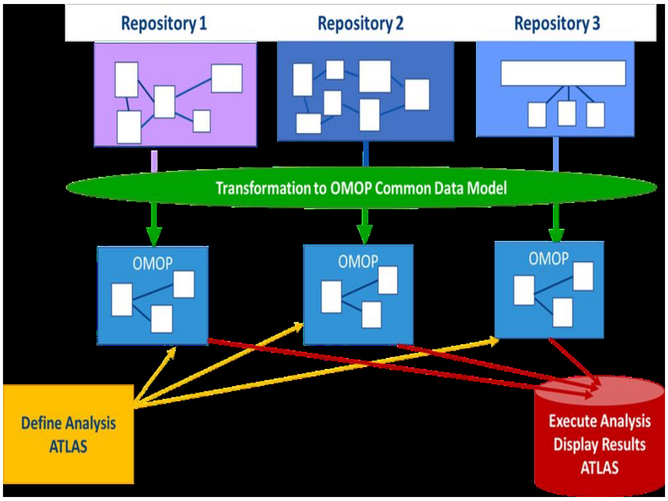

Figure 1 Observational Medical Outcomes Partnership Common Data Model (OMOP-CDM). Adapted from Standardised Data: The OMOP Common Data Model.[12]

### Aim

In this overview, we describe the OMOP-CDM, the established governance processes employed in EMR data repositories, and demonstrate how OMOP transformed data provides a lever for more efficient and secure access to EMR data, by health service providers, evaluators, auditors and researchers. Governance, privacy, consent and ethics vary by country or jurisdiction. For this review, we have applied an Australian context, however, the general nature of the guidance here is applicable internationally.

## THE OBSERVATIONAL MEDICAL OUTCOMES PARTNERSHIP COMMON DATA MODEL

### The OMOP-CDM: structure and process

The OMOP-CDM can be implemented using many of the existing database management systems. The OMOP-CDM extraction, transformation and loading process converts complex clinical and administrative EMR data into a simplified standard format consisting of 16 data tables and other derived tables.[8] Through this process, it is important to note the source EMR data are not changed or lost, OMOP conversion just provides a new representation of existing EMR data. For the deployment and installation of the OMOP-CDM into existing information system infrastructure (figure 1), we recommend the OMOP-CDM instance, that model's institution-specific data, is maintained under the existing repository data access and governance mechanisms established by each data custodian.

### OMOP, data quality and the principles of Findable, Accessible, Interoperable and Reusable, Collective benefit, Authority, Researcher and Ethics and Five Safes

The use of OMOP-CDM aligns well with the need for systematic data evaluation and adherence to data quality standards and the principles of FAIR (Findable, Accessible, Interoperable and Reusable), CARE (Collective benefit, Authority, Researcher and Ethics) and the Five Safes.

Before use, OMOP-CDM data undergoes a rigorous data quality assessment process, which includes checks for completeness, concordance, plausibility and currency when compared with the source EMR data.[16] These quality checks are predefined and configured to run on datasets conforming to OMOP standards, and they can be executed using tools such as Achilles, which is accessible via the OHDSI Data Quality Dashboard.[17] In addition, the OMOP-CDM enables researchers to work within a secure and firewalled environment while conducting advanced analytics and prediction techniques. This aligns with the principles of making data 'FAIR, ensuring that data are available for a wide range of research applications.[18 19] Data accessed through an OMOP-CDM also adheres to the CARE Principles for Indigenous Data. CARE operates within the governance framework established by the custodians of each local data repository. The CARE principles complement FAIR principles by aligning data sharing with the rights and interests of Indigenous Peoples. By adhering to CARE, Indigenous Peoples worldwide gain greater control over their data and the knowledge derived from it, ensuring alignment with their worldviews and the knowledge economy. This framework emphasises the Indigenous Peoples' right to derive value from data while promoting responsible and ethical data usage, for collective and equitable benefit of researchers, evaluators and the broader community.[18 19]

OMOP data adheres to the 'Five Safes' guiding principles by providing a structured and secure framework for managing and sharing healthcare data while ensuring privacy and security are maintained.[20] These frameworks were selected for their compatibility with the principles of ethical research, data quality and data governance. Their widespread adoption and acceptance within the research community make them robust and suitable choices for guiding data management practices in the context of the OMOP-CDM. The responsibility for applying the SAFES framework typically falls on various stakeholders involved in data access and usage, including government agencies, research institutions and data custodians (table 1).

## OMOP, DATA GOVERNANCE, ETHICAL REVIEW AND CONSENT

### OMOP-CDM and governance

By virtue of its design and objectives, the OMOP-CDM enhances the governance of secondary health data, by ensuring data utilisation in both research and healthcare decision-making is ethical, transparent and effective.

With the transformation of EMR data into a standardised structure, the OMOP-CDM ensures there is a uniform representation of these data regardless of the data's original source. This uniformity streamlines data governance and, importantly, eases the complexities associated with conducting single site studies that contain native EMR data (raw and/or curated), and multisite studies that involve integrating data from various disparate sources.[21 22] In addition, the common data model emphasises data quality, allowing for consistent checks and ensuring that research data meets the highest standards.[23 24] The standardised model also ensures that security and privacy protocols are uniformly applied, safeguarding secondary health data from data breaches to maintain patient privacy. Given the structured approach of the OMOP-CDM, an institution can easily implement access controls, thereby ensuring that only authorised parties can access or interact with the data. As a result, the OMOP-CDM acts as a cornerstone for the conduction of rigorous and ethically sound research as it builds trust among stakeholders, mitigates information disparities and encourages the production of high-quality medical evidence for rigorous and ethical research[25]

### Operational use and quality assurance activities in a hospital or healthcare setting

For operational use quality assurance activities where the 'primary purpose is to monitor or improve the quality of service delivered by an individual or an organisation'[26] data governance and principles for ethical use apply. However, within healthcare institutions, ethics approval is not mandated for the establishment of the OMOP database or data use, provided:

- The data being collected and analysed, is coincidental to standard operating procedures with standard equipment and/or protocols.
- The data are being collected and analysed expressly for the purpose of, maintaining standards or identifying areas for improvement in the environment from which the data were obtained.
- The data being collected and analysed, is not linked to individuals.
- None of the triggers for consideration of ethical review are present.[26]

### Research in a university setting

For research use, the data custodian is usually the agency or organisation that commissioned the research and paid for the data collected by the owner (ie, hospital/general practice). Existing local governance principles already developed by custodians can be applied to OMOP standardised data including: (1) data only being made accessible to named researchers on relevant ethics applications approved by the relevant institution, (2) appropriate secure data management strategies for transfer and management of data using password-protected computers or servers with multifactor authentication, 3) data restrictions that align with project scope and objectives and (4) storage of data outputs extracted from the OMOP-CDM as approved by the HREC. For

**Table 1** Guiding principles of FAIR, CARE and the Five Safes

| | | |
|---|---|---|
| The FAIR guiding principles | | |
| F | Findability | Metadata and data should be easily found by both humans and computers through the assigment of globally unique and permanent identifier to enable the automatic discovery of datasets and services via machine learning.[18] |
| A | Accessability | Metadata and data should easily retrieved by authorised and authenticated users via a standard communication protocol.[18] |
| I | Interoperabilty | Data from one data source can be integrated with data from other sources so that it can be aggregated into a single, unified view and refers to the intergration and exchange of applications, analysis, storage and workflow processing across different data sources.[18] |
| R | Reusability | Metadata and data characteristics are specified in detail to enable replication and/or linkage in different settings. Reusability includes the release of data usage licenses, provenance details and disclosure around community standards relevant to the domain.[18] |
| The CARE principles | | |
| C | Collective benefit | Collective benefit including where the well-being of Indigenous Peoples' rights is of primary concern.[19] |
| A | Authority | Indigenous Peoples' rights and interests about their peoples, communities, cultures and territories with regard to data are recognised and clearly articulated.[19] |
| R | Researcher | Researchers have a responsibility to develop and nurture respectful relationship with Indigenous Peoples' from whom the data originate.[19] |
| E | Ethics | Minimise harm and maximise benfit for Indigenous Peoples', for justice and future use.[19] |
| The Five Safes framework | | |
| People | | Safe People<br>Is the researcher appropriately trained and authorised to access and use the data?[20] |
| Projects | | Safe Projects<br>Is data used for an appropriate purpose that is valid and of public benefit?[20] |
| Settings | | Safe Settings<br>Does IT access and physical environment prevent unauthorised use?[20] |
| Data | | Safe Data<br>Has appropriate and sufficient protection been applied to the data to avoid risk of disclosure?[20] |
| Outputs | | Safe Outputs<br>Are the statistical results non-disclosive?[20] |

OMOP converted data that contains linked data, for example, AU-ePBRN where primary care data are linked with hospital admissions data,[27] governance and liability procedures would need to be explicitly developed to ensure the governance interests of all institutions are considered.

## Consent

In any research, regardless of whether it is conducted by an individual researcher, clinician or collaborative research team, it is imperative to determine the nature of the consent obtained from a patient for the secondary use of their data. This assessment should consider the risks and the potential for psychological, social, economic and legal harm that may arise from data collection, utilisation or any potential breaches.

In Australia, a 'waiver of consent' as per National Health and Medical Research Council (NHMRC) guidelines can be applied to secondary use of health data[26] (box 1). Some ethics committees may request an 'opt-out' model, necessitating the consideration of options for patients who wish to decline to participate.[26, 27] Through the deidentification methods employed by OMOP-CDM, the risks related to data breaches, such as the reidentification of individuals, are significantly reduced. This is achieved by exclusively using aggregated results from OMOP-CDM and by refraining from reporting small cell sizes. Reidentification is further minimised by ensuring only aggregated outputs from OMOP-CDM are used and that small cell sizes are not reported.

## Risk mitigation

OMOP mitigates many of the risks of using EMR data for secondary purposes including: (1) replacement of all personal identifiers with a generic number that does not allow reidentification back to the original personal identifier[12]; (2) an option for data custodians to perform analyses on behalf of an individual researcher and auditor (ie, no data release); (3) the use of a user interface tool such as ATLAS, where researcher or auditor access to data in all tables can

### Box 1 Consent

If an ethics committee deems a research project or a healthcare evaluation to be of minimal risk to the individual, an exception to obtaining the legislated requirement for patient consent can be managed using a 'waiver of consent.' A 'waiver of consent' can be applied based on a duty of 'easy rescue,' where the potential benefits of data access are considered significant, and the harm associated with the risk of a loss of privacy are considered minimal.[30] It is also hypothesised a 'waiver of consent' avoids the consequence of consent bias where individuals who provide informed consent to participate in a study differ in important ways from those who do not consent or choose not to participate.[30]

Numerous research and evaluation initiatives have employed a 'waiver of consent' approach, allowing for the secondary utilisation of electronic medical record (EMR) data.[31–34]

Arguments against a 'waiver of consent' considers the societal costs and potential patient harm, against the benefits of patient data utilisation. Costs include privacy breaches per se and the use of data for nefarious purposes, both of which contribute to a heightened risk of eroding trust.[35] This includes potential for a loss of informational privacy where an individual's personal or sensitive information is exposed, shared or accessed by others without their consent, or in a manner that violates their expectations of privacy.[35] Additional rational against the application of a 'waiver of consent' stems from the primary rationale for an individual's involvement in research lies in the process of duty of care to obtain, 'informed consent'. This justification is grounded in the idea that depending solely on research and evaluation might not adequately protect the values and interests of those participating. Further to this, informed consent is regarded as a means of building trust, not only in the research and evaluation process itself but in the researcher/clinician understanding of health data use.[35]

Notwithstanding ethics committee considerations for patient consent, there should also be considerable social engagement across a breadth of stakeholders on research that uses health data, even if it is deidentified. This engagement provides options for the provision of 'social permission' and 'social licence' for consent, where the determination of consent is cocreated by patients and therefore morally legitimised—beyond the limits of law and outside of what is acceptable by an ethics committee—to preserve societal trust.[36]

Patient and community acceptability of the use of data within their EMR for research and healthcare evaluation indicates, for social licence to be assumed, a breadth of patient and public values, needs and interests should be incorporated into governance frameworks.[37]

### Box 2 Risk mitigation

An access control policy is crucial for ensuring the privacy, management and security of data, especially when it is related to research. This ensures use of data is managed appropriately and underpinned by respect of the rights and expectations of the individuals it represents.

Access control measures include the application of strong passwords that are complex and contain alphanumeric characters as well as symbols; multifactor authentication where data users apply two or more evidence pieces (or factors) to verify their identity; safe connectivity where the standard practice for data access is via devices connected to secure and private networks rather than devices that are connected to public networks; the prompt reporting of data breaches to mitigate the impact of any cybersecurity attack and prevent further vulnerabilities; the verification of ethics approvals before granting access to ensure that research, healthcare evaluation and audit is conducted in an ethically sound manner; and the permissions for data access limited to those researchers and health service evaluators who are authorised and working within the confines of an institution's environment.

A review of data that is due to be transmitted to researchers and health service evaluators provides another important safety check, as does maintenance of version control, where the most recent database is always held as back up. Additional risk management controls include the delivery of explicit instructions to researchers and evaluators on the appropriate use of the dataset; the incorporation of additional hardware authentication such as the YubiKey, Titan, Thetis and Kensington Verimark hardware keys; restricted access to identifiers in the underlying Structured Query Language database and continuous evaluation of anonymisation adequacy instructions on appropriate use for dataset.

be configured to protect privacy[8]; (4) collaborative analyses are always conducted within each institution's firewalled network[8]; (5) use of standardised terminology only removes potential identifiers in the source terminology and (6) there is an option to obscure dates from view, such that temporal association can be calculated from a relative date (box 2).[28]

## BENEFITS, LIMITATIONS AND CONSIDERATIONS OF OMOP-CDMS

OMOP-converted databases offer a secure and standardised approach to EMR data analysis within an open-source framework, which produces aggregated results which are free of patient identifiers. This eliminates the need for direct access to native EMR data or external sources that have data sharing restrictions, it also sets it apart from the less structured EMR 'data lakes' that contain vast amounts of native and disparate data. These 'data lakes' that lack standardised schemas, make data management and analysis more challenging. In contrast, OMOP-CDM benefits from OHDSI's open-source tools and standardised analytics, by enhancing transparency, reducing coding errors and supporting validation processes.

As an extension of OMOP data conversion, the OHDSI consortium has developed the OHDSI Quality Dashboard, a tool to ensure the quality of data converted into the OMOP-CDM to improve transparency, reduce coding errors and enable validation.[29] The OHDSI Quality Dashboard is designed to assess and monitor the quality of data that has been converted into the OMOP-CDM. It provides a set of data quality checks and validation tools that help identify issues or anomalies in the converted data.[16] In doing so, it identifies and addresses data quality issues that may arise during the conversion process by checking data for completeness, consistence, accuracy and adherence to standard terminologies. This ensures that the data in the OMOP-CDM are reliable and suitable for research, analysis, evaluation and audit.

Specialist fields such as oncology and pregnancy have unique data requirements.[8] For instance, cancer-related data elements can vary among healthcare sites due to clinical practice variations. The OMOP-CDM may not always fully standardise these elements during the mapping process. To address specialised data needs, the OHDSI community actively develops and shares OMOP-CDM extensions, particularly for specific cancer types.

Specialist fields such as oncology and pregnancy have specific data needs[8] and data elements may vary among healthcare sites due to differences in clinical practices. The OMOP-CDM may not always fully standardise these elements during the mapping process. To cater to these specialised data needs, the OHDSI community actively creates, refines and disseminates OMOP-CDM extensions that are highly specific to cancer types and treatments, as well as pregnancy episodes and outcomes.

Given the EMR captures similar data across various healthcare sites, such as specific pathology indicators, they may contain different data elements due to differences in pathology classifications (ie, pathology definitions and units). To address these variations and maintain the OMOP-CDM's relevance and flexibility, the OHDSI community also actively develops extension to address variation in pathology phenotypes to preserve the OMOP-CDM's overall compatibility and interoperability across diverse healthcare sites and research projects.

Institutional governance and privacy frameworks have evolved independently alongside the adoption of secondary EMR use,[7] therefore, achieving a consensus on governance practices across institutions is an ongoing endeavour. This underscores the importance of ongoing collaboration and standardisation in the healthcare data field to ensure that valuable health data can be leveraged effectively and ethically for research and healthcare improvement. Given all the opportunities the OMOP-CDM offers for integrated data governance, these opportunities are limited by lack of standalone funding required for the comprehensive mapping of data from local EMRs to the common format. Despite these challenges, the commitment of the global community to the OMOP-CDM signifies a promising future for standardised health data, which will pave the way to transform healthcare research, evaluate operational processes and facilitate quality improvement within healthcare organisations.

## CONCLUSION

Adoption of the OMOP-CDM internationally and locally is well worth the investment, as it enables conversion of large amounts of complex, and heterogeneous EMR data into a standardised structured data model, simplifies governance processes and facilitates rapid repeatable cross-institution analysis through shared end-to-end analysis packages, without the sharing of native data. Combined with pseudonymisation and common data quality assessments, the OMOP-CDM provides a powerful model to support ethical real-world 'big' data research. The continued adoption of OMOP-CDM, ongoing development efforts, and the emphasis on sound governance practices all contribute to the realisation of OMOP's utility in unlocking valuable EMR data. These factors collectively support a wide range of applications, from health service operational reporting to diverse clinical, epidemiological and translational research projects.

While the adoption of OMOP and the collaborative efforts in data integration in Australia is commendable, there is room for further development in bridging the gap between hospital and primary care data. This ongoing endeavour has the potential to significantly enhance Australia's capacity for data-driven research and improve healthcare outcomes for its population.

**Author affiliations**

[1]Health and Biomedical Informatics Centre, Research Information Technology Unit (HaBIC R2), Department of General Practice and Primary Care, The University of Melbourne Faculty of Medicine Dentistry and Health Sciences, Melbourne, Victoria, Australia
[2]School of Computing and Information Systems, Faculty of Engineering and Information Technology, Centre for the Digital Transformation of Health, The University of Melbourne Faculty of Medicine Dentistry and Health Sciences, Melbourne, Victoria, Australia
[3]Queensland Digital Health Centre (QDHeC), Centre for Health Services Research, The University of Queensland Faculty of Medicine, Woolloongabba, Queensland, Australia
[4]Quality Use of Medicines and Pharmacy Research Centre, Clinical and Health Sciences, University of South Australia, Adelaide, South Australia, Australia
[5]Clinical Haematology Department, The Royal Melbourne Hospital, Peter MacCallum Cancer Centre, Melbourne, Victoria, Australia
[6]Sir Peter MacCallum Department of Oncology, The University of Melbourne Faculty of Medicine Dentistry and Health Sciences, Melbourne, Victoria, Australia
[7]Department of General Medicine, The Royal Melbourne Hospital, Parkville, Victoria, Australia
[8]Queensland Digital Health Centre (QDHeC), Centre for Health Services Research, The University of Queensland Faculty of Medicine, Herston, Queensland, Australia
[9]School of Population Health, UNSW, Sydney, New South Wales, Australia
[10]Centre for Big Data Research in Health (CBDRH), UNSW, Sydney, New South Wales, Australia

**Twitter** Christine Mary Hallinan @Cmhallinan, Ashley P Ng @drng and Siaw-Teng Liaw @ElAstee

**Collaborators** Not applicable.

**Contributors** The research presented in this paper was a collaborative effort involving multiple authors who contributed to various aspects of the review. The review was conceived by RW, GKH, NP and DBu. CMH structured the review, synthesised the evidence, drafted the manuscript, reviewed the manuscript and completed the final version of the manuscript. RW, GKH, CS, NP, AN, DC, AVDV, S-TL, OD, BGL, DBu and DBo provided evidence and edits for the review, and reviewed the full manuscript when completed.

**Funding** This study was funded by Australian Government, Australian Research Data Commons (ARDC) Public Sector Bridges Program 'Electronic Medical Records as a National Data Asset' (Doi: 10.47486/PS014).

**Competing interests** None declared.

**Patient consent for publication** Not applicable.

**Provenance and peer review** Not commissioned; externally peer reviewed.

**Data availability statement** Data sharing not applicable as no datasets generated.

**ORCID iDs**
Christine Mary Hallinan http://orcid.org/0000-0002-0471-4444
Ashley P Ng http://orcid.org/0000-0001-9690-0879
Siaw-Teng Liaw http://orcid.org/0000-0001-5989-3614

## REFERENCES

1 Safran C, Bloomrosen M, Hammond WE, *et al*. Toward a national framework for the secondary use of health data: an American medical informatics association white paper. *J Am Med Inform Assoc* 2007;14:1–9.
2 Danciu I, Cowan JD, Basford M, *et al*. Secondary use of clinical data: the vanderbilt approach. *J Biomed Inform* 2014;52:28–35.
3 Budrionis A, Bellika JG. The learning healthcare system: where are we now? A systematic review. *J Biomed Inform* 2016;64:87–92.
4 Park K, Cho M, Song M, *et al*. Exploring the potential of OMOP common data model for process mining in healthcare. *PLoS One* 2023;18:e0279641.
5 Weiskopf NG, Weng C. Methods and dimensions of electronic health record data quality assessment: enabling reuse for clinical research. *J Am Med Inform Assoc* 2013;20:144–51.
6 Jean-Baptiste L, Mouazer A, Sedki K, *et al*. Translating the observational medical outcomes partnership - common data model (OMOP-CDM) electronic health records to an OWL ontology. *Stud Health Technol Inform* 2022;290:76–80.
7 Evans RS. Electronic health records: then, now, and in the future. *Yearb Med Inform* 2016;Suppl 1(Suppl 1):S48–61.
8 OHDSI. Observational health data sciences and Informatics data standardization 2021. 2021 Available: https://www.ohdsi.org/data-standardization/the-common-data-model/
9 SNOMED CT. SNOMED International leading Healthcare terminology, worldwide 2022. 2022. Available: https://www.snomed.org/
10 LOINC. LOINC the International standard for identifying health measurements, observations, and documents 2022. 2022. Available: https://loinc.org/
11 National Library of Medicine. Unified medical language system Rxnorm. 2022. Available: https://www.nlm.nih.gov/research/umls/rxnorm/index.html
12 OHDSI. Our journey: where the OHDSI community has been and where we are going observational health data sciences and Informatics; 2022.
13 Electronic medical records National data asset (Internet). 2022. Available: https://doi.org/10.26188/6295c4a5d7c5c
14 Electronic Practice Based Research Network. Centre for Primary Health Care and Equity, UNSW Sydney, Available: https://cphce.unsw.edu.au/research/electronic-practice-based-research-network
15 Boyle D, Sanci L, Emery J, *et al*. PATRON Primary Care Research Data Repository, . 2019Available: https://medicine.unimelb.edu.au/school-structure/general-practice-and-primary-care/research/data-for-decisions
16 Kahn MG, Callahan TJ, Barnard J, *et al*. A harmonized data quality assessment terminology and framework for the secondary use of electronic health record data. *EGEMS (Wash DC)* 2016;4:1244.
17 OHDSI. Observational health data sciences and Informatics ACHILLES for data characterization 2022. n.d. Available: https://www.ohdsi.org/analytic-tools/achilles-for-data-characterization/
18 Wilkinson MD, Dumontier M, Jan Aalbersberg I, *et al*. Addendum: the fair guiding principles for scientific data management and stewardship. *Sci Data* 2019;6:6.
19 The Global Indigenous Data Alliance. CARE principles for indigenous data governance. 2022. Available: https://www.gida-global.org/care
20 ABS. Five safes framework Australian Bureau of Statistics Canberra: Australia. n.d. Available: https://www.abs.gov.au/about/data-services/data-confidentiality-guide/five-safes-framework
21 Biedermann P, Ong R, Davydov A, *et al*. Standardizing registry data to the OMOP common data model: experience from three pulmonary hypertension databases. *BMC Med Res Methodol* 2021;21:238.
22 OHDSI. The book of OHDSI: observational health data sciences and Informatics. 2023. Available: https://ohdsi.github.io/TheBookOfOhdsi/
23 Blacketer C, Defalco FJ, Ryan PB, *et al*. Increasing trust in real-world evidence through evaluation of observational data quality. *J Am Med Inform Assoc* 2021;28:2251–7.
24 Blacketer C, Voss EA, DeFalco F, *et al*. Using the data quality dashboard to improve the EHDEN network. *Applied Sciences* 1192;11:11920.
25 Kim J-W, Kim C, Kim K-H, *et al*. Scalable infrastructure supporting reproducible nationwide healthcare data analysis toward FAIR stewardship. *Sci Data* 2023;10:674.
26 NHMRC. National Stnational statement on ethical conduct in human research (2007) - updated 2018; 2018. National health and medical research Council
27 UNSW. The electronic practice based research network Sydney: centre for primary health care and equity medicine. 2021. Available: https://cphce.unsw.edu.au/research/electronic-practice-based-research-network
28 Hripcsak G, Mirhaji P, Low AF, *et al*. Preserving temporal relations in clinical data while maintaining privacy. *J Am Med Inform Assoc* 2016;23:1040–5.
29 Schuemie MJ, Ryan PB, Pratt N, *et al*. Principles of Large-scale Evidence Generation and Evaluation across a Network of Databases (LEGEND). *J Am Med Inform Assoc* 2020;27:1331–7.
30 Porsdam Mann S, Savulescu J, Sahakian BJ. Facilitating the ethical use of health data for the benefit of society: electronic health records, consent and the duty of easy rescue. *Phil Trans R Soc A* 2016;374:20160130.
31 Tu K, Sarkadi Kristiansson R, Gronsbell J, *et al*. Changes in primary care visits arising from the COVID-19 pandemic: an international comparative study by the International consortium of primary care big data researchers (INTRePID). *BMJ Open* 2022;12:e059130.
32 Lu Y, Van Zandt M, Liu Y, *et al*. Analysis of dual combination therapies used in treatment of hypertension in a multinational cohort. *JAMA Netw Open* 2022;5:e223877.
33 Ahmadi N, Peng Y, Wolfien M, *et al*. OMOP CDM can facilitate data-driven studies for cancer prediction: a systematic review. *Int J Mol Sci* 2022;23:19.
34 Lane JCE, Weaver J, Kostka K, *et al*. Risk of hydroxychloroquine alone and in combination with azithromycin in the treatment of rheumatoid arthritis: a multinational, retrospective study. *Lancet Rheumatol* 2020;2:e698–711.
35 Ploug T. In Defence of informed consent for health record research - why arguments from “easy rescue”, “no harm” and “consent bias” fail. *BMC Med Ethics* 2020;21:75.
36 Muller SHA, Kalkman S, van Thiel G, *et al*. The social licence for data-intensive health research: towards co-creation, public value and trust. *BMC Med Ethics* 2021;22:110.
37 Kalkman S, van Delden J, Banerjee A, *et al*. Patients’ and public views and attitudes towards the sharing of health data for research: a narrative review of the empirical evidence. *J Med Ethics* 2022;48:3–13.

BMJ Health & Care Informatics

# Explainable machine learning for breast cancer diagnosis from mammography and ultrasound images: a systematic review

Daraje kaba Gurmessa ,[1,2] Worku Jimma[1]

**To cite:** Gurmessa Dkaba, Jimma W. Explainable machine learning for breast cancer diagnosis from mammography and ultrasound images: a systematic review. *BMJ Health Care Inform* 2024;**31**:e100954. doi:10.1136/bmjhci-2023-100954

► Additional supplemental material is published online only. To view, please visit the journal online (https://doi.org/10.1136/bmjhci-2023-100954).

Received 06 November 2023
Accepted 21 January 2024

[1]Department of Information Science, Jimma Institute of Technology, Jimma University, Jimma, Oromia, Ethiopia
[2]Computer Science, Mattu University, Mattu, Oromīya, Ethiopia

**Correspondence to**
Daraje kaba Gurmessa;
darajekaba2020@gmail.com

## ABSTRACT

**Background** Breast cancer is the most common disease in women. Recently, explainable artificial intelligence (XAI) approaches have been dedicated to investigate breast cancer. An overwhelming study has been done on XAI for breast cancer. Therefore, this study aims to review an XAI for breast cancer diagnosis from mammography and ultrasound (US) images. We investigated how XAI methods for breast cancer diagnosis have been evaluated, the existing ethical challenges, research gaps, the XAI used and the relation between the accuracy and explainability of algorithms.

**Methods** In this work, Preferred Reporting Items for Systematic Reviews and Meta-Analyses checklist and diagram were used. Peer-reviewed articles and conference proceedings from PubMed, IEEE Explore, ScienceDirect, Scopus and Google Scholar databases were searched. There is no stated date limit to filter the papers. The papers were searched on 19 September 2023, using various combinations of the search terms 'breast cancer', 'explainable', 'interpretable', 'machine learning', 'artificial intelligence' and 'XAI'. Rayyan online platform detected duplicates, inclusion and exclusion of papers.

**Results** This study identified 14 primary studies employing XAI for breast cancer diagnosis from mammography and US images. Out of the selected 14 studies, only 1 research evaluated humans' confidence in using the XAI system—additionally, 92.86% of identified papers identified dataset and dataset-related issues as research gaps and future direction. The result showed that further research and evaluation are needed to determine the most effective XAI method for breast cancer.

**Conclusion** XAI is not conceded to increase users' and doctors' trust in the system. For the real-world application, effective and systematic evaluation of its trustworthiness in this scenario is lacking.

**PROSPERO registration number** CRD42023458665.

## INTRODUCTION

Breast cancer is the first and most common type of cancer in women.[1 2] Anatomically, the breast consists of healthy blood vessels, connective tissue, ductal lobules and lymph nodes.[3] Breast cancer is a problem with abnormal growth of the breast cells. By 2040, the burden of breast cancer is predicted to increase to over three million new cases and one million deaths every year because of population growth and ageing alone.[2]

Breast cancer is highly treatable if identified at an early stage, and hence, early detection is crucial to save lives. Among the methods of breast cancer detection, the most popular are ultrasound (US),[4] mammography[5] and MRI. However, traditional computer-aided design systems generally depend on manually created features and experience of the physiologist, therefore weakening the overall performance of breast cancer identification. Therefore, artificial intelligence (AI) methods like machine learning and deep learning-based techniques have emerged for breast cancer diagnosis with high accuracy. Additionally, improved breast cancer classification by combining graph convolutional network and convolutional neural network[6] and abnormal breast identification by a nine-layer convolutional neural network with parametric rectified linear unit and rank-based stochastic pooling are used to support patients and doctors' decisions.[7] However, the algorithms lack ethical AI, right of explanation and trustworthy AI. These concepts are considered critical issues by high-level political and technical bodies (eg, G20, EU expert groups, Association of Computing Machinery in the USA).[8 9]

Additionally, AI algorithms like machine learning and deep learning are vulnerable to bad stuff (bad decisions, bad medical diagnosis and bad prediction) is the most common drawback of AI algorithms today. They are also black box for predictive interpretation.

To overcome this issue, the science of explainable AI (XAI) has grown exponentially with its successful application in breast cancer diagnosis. However, it still requires a comprehensive review of existing studies to help researchers and practitioners gain insight and understanding of the field. Therefore, his systematic review is conducted.

**Table 1** Search term combination

<table>
<tr><th></th><th></th><th></th><th>AND (&&)</th><th></th><th></th><th>OR (||) or AND (&&)</th><th></th></tr>
<tr><td>OR (||)</td><td>Term 1</td><td></td><td>Term 2</td><td>Term 3</td><td>Term 4</td><td>Term 5</td><td>Term 6</td></tr>
<tr><td></td><td>Explainable</td><td>AND (&&)</td><td>Machine Learning</td><td>Breast</td><td>Cancer</td><td>Mammography</td><td>Ultrasound</td></tr>
<tr><td></td><td></td><td></td><td>Artificial Intelligence</td><td></td><td></td><td></td><td></td></tr>
<tr><td></td><td>Interpretable</td><td></td><td>Deep learning</td><td></td><td></td><td></td><td></td></tr>
<tr><td></td><td></td><td></td><td>AI</td><td></td><td></td><td></td><td></td></tr>
<tr><td></td><td>XAI</td><td></td><td></td><td></td><td></td><td></td><td></td></tr>
</table>

XAI is the extent to which people can easily understand the model. It has received much attention over the past few years. The purpose of a model explanation is to clarify why the model makes a certain prediction, to increase confidence in the model's predictions[10] and to describe exactly how a machine learning model achieves its properties.[11] Therefore, using machine learning explanations can increase the transparency, interpretability, fairness, robustness, privacy, trust and reliability of machine learning models. Recently, various methods have been proposed and used to improve the interpretation of machine learning models.

There are different taxonomies for machine learning explainability. An interactive explanation allows consumers to drill down or ask for different types of explanations until they are satisfied, while a static explanation refers to one that does not change in response to feedback from the consumer.[12] A local explanation is for a single prediction, whereas a global explanation describes the behaviour of the entire model. A directly interpretable model is one that by its intrinsic transparent nature is understandable by most consumers, whereas a post hoc explanation involves an auxiliary method to explain a model after it has been trained.[13] Self-explaining may not necessarily be a directly interpretable model. By itself, it generates local explanations. A surrogate model is usually a directly interpretable model that approximates a more complex model, while visualisation of a model may focus on parts of it and is not itself a full-fledged model.

No single method is always the best for interpreting machine learning.[12] For this reason, it is necessary to have the skills and equipment to fill the gap from research to practice. To do so, XAI toolkits like AIX360,[12] Alibi,[14] Skater,[15] H2O,[16 17] InterpretML,[18 19] EthicalML-XAI,[19 20] DALEX,[21 22] tf-explain,[23] Investigate.[24] Most interpretations and explanations are post hoc (local interpretable model-agnostic explanations (LIME) and SHapley Additive exPlanations (SHAP). LIME and SHAP are broadly used explanation types for machine learning models from physical examination datasets. But these made explanations with limited meaning as they lacked fidelity and transparency. However, deep learning and ensemble gradients are preferable in performance for image processing and computer vision. This research is processing mammography and US images. Therefore, deep learning is recommended for breast cancer image processing.

Ensemble gradients are used to interpret deep neural networks,[11] GradientSHAP is a sample interpretation algorithm that approximates SHAP values.[25] Occlusion methods are most useful in situations such as image processing. Biological nurturing(BN) is ideal for clinical decision-making and, in general, for all assessments and studies involving multiple interventions and orientations. The oriented, modified integrated gradient (OMIG) interpretability method is inspired by the integrated gradients method. Since there is no one-size-fits-all approach to learning machine explanation, it needs a comprehensive evaluation of published papers and tools to bridge the gap in research to practice.

The research that does not consider objective metrics for evaluating XAI may lack significance and experience controversy, especially if negative reviews are not used.[8] To avoid the issues, a study[8] suggests four metrics based on performance differences, D, between the explanation's logic and the agent's actual performance, the number of rules, *R*, outputted by the explanation, the number of features, F, used to generate the explanation, and the stability, S, of the explanation. It is believed that user studies that focus on D, R, F and S metrics in their evaluations are inherently more valid.

The main contributions of this systematic review are:

1. Investigating XAI methods popularly applied for breast cancer diagnosis.
2. Identifying the algorithm's explainability and their performance relation in breast cancer diagnosis.
3. Summarise the evaluation metrics used for breast cancer diagnosis using XAI methods.
4. Summarise existing ethical challenges that XAI overcomes in breast cancer diagnoses.
5. Analysing the research gaps and future direction for XAI for breast cancer detection.

## METHODOLOGY

The methodology employed in this systematic review is devoid of any medical (either prospective or retrospective) data of patients. This study applies the Preferred Reporting Items for Systematic Reviews and Meta-Analyses (PRISMA 2020) guiding principles for conducting systematic reviews.[26] PRISMA 2020 was adopted because

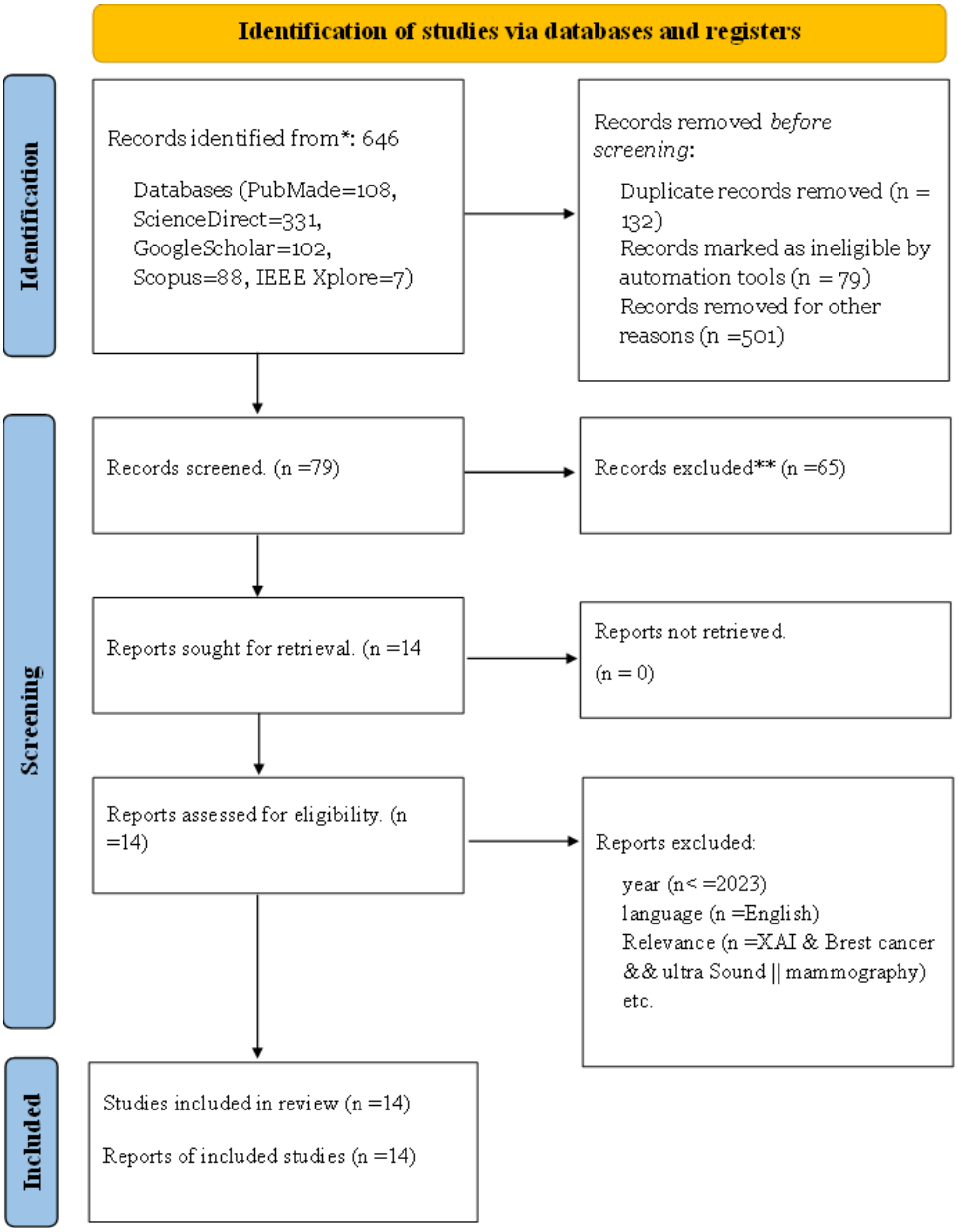
Identification of studies via databases and registers
Identification
Records identified from*: 646
Databases (PubMade=108, ScienceDirect=331, GoogleScholar=102, Scopus=88, IEEE Xplore=7)
Records removed *before screening*:
Duplicate records removed (n = 132)
Records marked as ineligible by automation tools (n = 79)
Records removed for other reasons (n =501)
Screening
Records screened. (n =79)
Records excluded** (n =65)
Reports sought for retrieval. (n =14
Reports not retrieved.
(n = 0)
Reports assessed for eligibility. (n =14)
Reports excluded:
year (n< =2023)
language (n =English)
Relevance (n =XAI & Brest cancer && ultra Sound || mammography)
etc.
Included
Studies included in review (n =14)
Reports of included studies (n =14)

**Figure 1** Preferred Reporting Items for Systematic Reviews and Meta-Analyses flow chart of explainable artificial intelligence (XAI) for breast cancer diagnosis.

of the clear guidelines it offers to ease robust systematic reviews. Therefore, this review article follows the recommendations of the guidelines. There is no stated date limit to filter the papers. The papers were searched on 19 September 2023. Peer-reviewed manuscripts and conference proceedings from PubMed, IEEE Explore, ScienceDirect, Scopus and Google Scholar databases published were searched. Rayyan for systematic review was used for duplicate removing, inclusion and exclusion term visualisations. The systematic review protocol was registered through PROSPERO with ID CRD42023458665.[27] Preplanned subgroup analyses were detailed.

## Search strategy

Five databases (PubMed, IEEE Explore, ScienceDirect, Scopus and Google Scholar) were searched systemically on 19 September 2023. There is no stated date limit to filter the papers. The terms and logical operations are combined and arranged as per tables 1 and 2.

## Inclusion and exclusion criteria

After applying the search equation, the criteria for inclusion and exclusion are as follows:

- Literature or systematic review articles were excluded.
- All articles focusing specifically on using XAI and strategies for breast cancer diagnosis using US, mammography or both (practical or theoretical) were included.
- Articles dealing with relevant technologies but, used procedures other than breast cancer diagnosis using US, mammography, or both were excluded, even if these systems were mentioned elsewhere in the article.

- Articles published in languages other than English were excluded.
- Articles by year of publication were not excluded, given the novelty of using XAI for breast cancer diagnosis using US mammography or both.

### Study selection

The selection process of the articles was conducted based on the inclusion and exclusion criteria defined (figure 1). A bibliography of 646 papers was extracted from databases (PubMed=118, ScienceDirect=331, Scopus=88, Google Scholar=102 and IEEE Xplore=7). All the extracted papers were imported into the Rayyan online platform for systematic review. In total, 132 articles were found to be duplicates and were deleted. Moreover, 501 papers were excluded (systematic review, scoping review, breast cancer diagnosis without explainable AI and explainable AI without breast cancer diagnosis). In total, 79 papers with XAI for breast cancer terms were retained. Their full documents were downloaded and reviewed. From these, 65 papers with XAI for breast cancer without mammography or US terms were excluded again. Finally, 14 studies with XAI for breast cancer and mammography or US or both terms were included and used for this systematic review.

### Risk of bias (quality) assessment method

Quality and risk of bias are assessed using Risk of Bias Visualization assessment tool in a systematic review assessment tool.[28] The tool creates traffic light plots of the domain-level judgments for each result and weighted bar plots of the distribution of risk-of-bias judgments within each bias domain.[28]

## RESULTS AND DISCUSSION

### Results

A total of 646 papers were extracted using search queries and terms defined in tables 1 and 2 from the selected databases. From a total of 646 papers, 134 were duplicates and removed. As depicted in figure 1, based on inclusion and exclusion stated in section Inclusion and exclusion criteria above, 79 papers (14%) with XAI for breast cancer were included (figure 1). Figure 2 depicts the included and excluded ratios. All screenshots added to these results are taken from Rayyan for a systematic review online platform.

US and mammography are the most recommended methods for breast cancer diagnosis. From 79 included papers based on XAI for breast cancer, 14 papers with XAI for breast cancer and mammography or US or both terms were either included or excluded based on criteria set in section Inclusion and exclusion criteria above. So, table 3 presents that 64.29% (9 papers from included 14) of papers were on US images, whereas 35.71% (5 papers from included 14) of papers were on mammography images.

Figure 2 shows that 97% were excluded and 3% were included based on inclusion criteria. Table 3 shows that 100% of the included papers visualised are XAI for breast cancer from mammography, US or both. It shows that 50% of them used heatmaps for visualisation.

The main objective of XAI is to encounter ethical challenges and to increase doctors' and patients' thrust on XAI. Different XAI are used for breast cancer. However, only one paper compared doctors' trust in the system.

In most of the papers, 50% (7 from 14 papers) used heatmaps for visualisation of areas of interest[29–35] and.[36] Additionally, Zhang *et al*[37] used BI-RADS-Net, Zhang *et al*[38] and Shen *et al*[35] used a saliency map, Ortega-Martorell *et al*[39] used uniform manifold approximation and projection (UMAP), Mital and Nguyen[40] used a tornado diagram, Rezazadeh *et al*[41] used histogram and Rezazade Mehrizii *et al*[34] used class activation map (CAM)-based heatmaps.

Shen *et al*'s study[35] used the largest number of datasets when compared with included studies. The study proves that the artificial intelligence system reduces false-positive findings in the interpretation of breast US examinations.[35] Breast cancer is most common in women, based on evidence on the ground in all of the studies most of the data are from women. This implies the ground truth. However, most of the datasets are taken from women and do not keep the existence of breast cancers in the ratio from man to women.

**Table 2** Search equations

| No | Database | Query | Number of papers |
|---|---|---|---|
| 1 | ScienceDirect | ((('Explainable' \|\|'Interpretable') && ('AI' \|\|'Artificial Intelligence' \|\|'machine learning' \|\|'deep Learning') && 'Breast Cancer')) | 331 |
| 2 | PubMed | ((((Breast Cancer) AND (((Explainable) OR (Explainable))) AND ((AI) OR (Artificial Intelligence) OR (machine learning) OR (deep Learning)) | 118 |
| 3 | IEEE Explore | Explainable Machine Learning for Breast Cancer Diagnosis from Mammography and Ultrasound Images | 7 |
| 4 | Scopus | ((('Explainable' \|\|'Interpretable') && ('AI' \|\|'Artificial Intelligence' \|\|'machine learning' \|\|'deep Learning') &&'Breast Cancer')) | 88 |
| 5 | Google Scholar | Explainable Machine Learning for Breast Cancer Diagnosis from Mammography and Ultrasound Images | 102 |

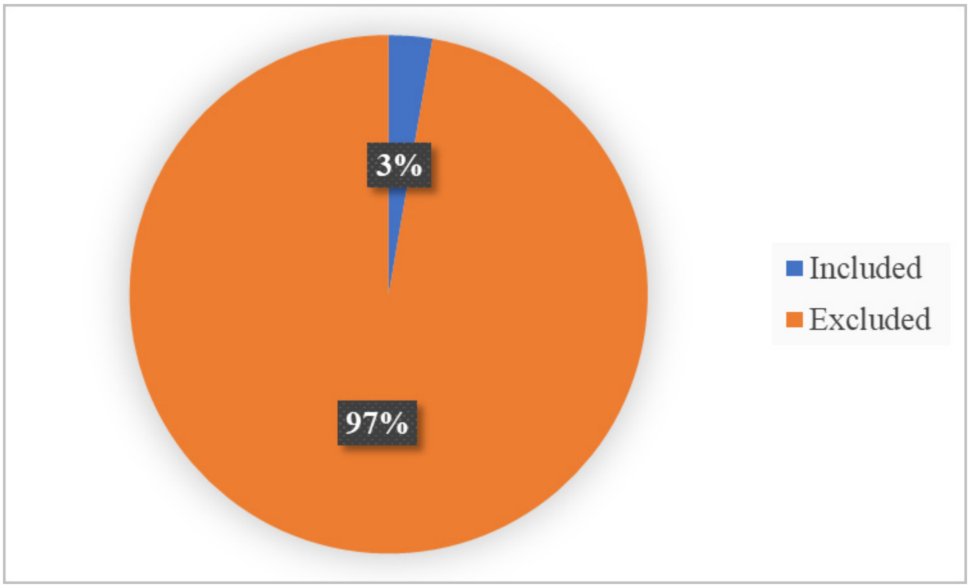

**Figure 2** Included and excluded ratio graph for explainable artificial intelligence, breast cancer and mammography or ultrasound.

A total of 5 648 066 datasets are used by all included papers. From all the included papers, US-based datasets were used by 99% of studies. Mammography-based datasets used by only 1% of the total studies. For example, the maximum datasets used by Shen *et al*[35] used 5 442 907 US images, and the study by Mital and Nguyen[40] used 100 000 mammography images. This shows that there are many works left to work on improving the number of datasets on mammography images when compared with US images. We recommend that data should be collected from suspected patients with breast cancer but all the included studies said nothing about it.

Explainable/interpretable algorithms used are deep learning explanation algorithms: Of 14 papers, Explainer alone or with Grad-CAM,[29] interpretable deep learning,[30] Grad-CAM,[31] Fisher information network (FIN),[39] AI and Polygenic Risk Scores (PRS) algorithms,[40] DenseNet,[35] Explainability-partial,[34] Explainability-full,[34] VGG-16,[37] fine-tuned MobileNet-V2 convolutional neural network,[33] OMIG explainability[32] and BI-RADS-Net-V2[38] are used in 11 papers (78.57 %), SHAP[41 42] is used in 2 papers (14.3%) and LIME[36] is used in 1 paper (7.14%).

### Risk of bias

The study population was known in all articles. We have obtained complete outcome variables in all articles. In all articles involved, selective reporting and publication bias were not obtained (figure 3). 'Traffic light' plots of the domain-level judgments for each result are sh0wn in figure 3.

## DISCUSSION

Explainer is the situation that is explainable by itself rather than explaining black box.[29] They proved that physicians perform better when assisted by Explainer than when diagnosing alone. The study compares the use of Explainer with the post hoc technique. Based on this, they prove that Explainer can locate more reasonable and feature-related regions than the classic post hoc technique. Robustness is a characteristic expected from XAI. The study by Song *et al*[29] also tested the robustness of the proposed framework. Explainability[29] is not only related to AI performance but also to responsibility and risk in medical diagnosis. For phantom object detection,[30] accuracy and mean intersection over union were used to test the model over a total of 6369 out of 6400 objects. Finally, Oh *et al*'s study[30] concludes interpretable deep learning model using large-scale data from multiple centres shows high performance.

In the study by Qian *et al*,[31] BI-RADS scores for breast cancer were compared with experienced radiologists, areas under the receiver operating curve (ROC) and CI for multimodal images. Explanation using principal component analysis, visualisation using UMAP, FIN visualisations of the training cases and projecting the test cases onto the trained embedding.[39] the study propose a novel visualisation using FIN containing accurate information about data points' similarities that can provide intelligence about neighbouring data points.

The finding by Mital and Nguyen[40] explained AI's ability to identify high-risk women more accurately than PRS, and family history reduces the possibility of delayed breast cancer diagnosis and fewer false-positive diagnoses from not screening low-risk women.

In Sun *et al*'s study,[42] model-agnostic methods versus model-specific methods, post hoc (black box+SHAP) technique and three algorithms, namely, logistic regression, extreme gradient boosting and random forest performance, were evaluated by sensitivity, specificity and AUC.[42] This evaluation was used to evaluate the black box model only. Moreover, SHAP was used for visualising feature importance using a heatmap but it was not tested.

Table 3 Overview of reviewed articles on explainable artificial intelligence data

| Reference | Number of images | Type of data | Population | Image type | Features used | XAI used | Visualisations |
|---|---|---|---|---|---|---|---|
| 29 | 19 341 | Image | 19 341 | Ultrasound (US) | Physician-annotated TI-RADS features | Explainer alone or with Grad-CAM | Heatmaps |
| 30 | 2208 (training 1808, testing 400) | Images | 1755 | Mammography | Phantom object detection | Interpretable deep learning | Heatmaps |
| 31 | 10 815 (1633 lesions) | Images | 775 | US | Convolutional features | Grad-CAM | Heatmaps |
| 39 | 2000 | Images | 1246 women | Mammography | CNN features | FIN | UMAP |
| 40 | 100 000 | Images | Images | US | ICER | AI and PRS algorithms | Tornado diagram |
| 42 | 11 294: 45–49 years (5709) and 50–54 years (5585) | Images | 11 294 | Mammography | Random forest | SHAP | SHAP values |
| 36 | 153 | Images | 153 patients (59 with metastasis and 94 without metastasis) | US | CNN features | LIME | CAM-based heatmaps |
| 35 | 5 442 907 | Images | 143 203 | US | Coarse and fine ROIs | DenseNet | Saliency maps heatmaps |
| 34 | 2760 | Image | 2760 | Mammography | Morphological and numerical inputs | Explainability-partial Explainability-full | Heatmap and numerical attributes |
| 37 | 1192 (BUSIS 562 and BUSI 630) | Image | 1192 | US | BI-RADS descriptors | VGG-16 | BI-RADS-Net |
| 33 | 624 | Image | 624 | US | BI-RADS descriptors | Fine-tuned MobileNet-V2 convolutional neural network | Heatmaps |
| 41 | 780 | Image | 600 female patients (age 25–75 years) | US | GLCM texture features | SHAP | Histogram |
| 32 | 52 800 simulated, 4800 real and 48 augmentations | Image | 4800 | Mammography | Phantom features | OMIG explainability | Heatmaps |
| 38 | 1192 (727 benign (negative) and 465 malignant (positive)) | Image | 1192 | US | Morphological features | BI-RADS-Net-V2 | Saliency map |

*BUSIS, Breast Ultrasound Image Segmentation

AI, artificial intelligence; BI-RADS, breast imaging reporting and data system; BUSI, Breast Ultrasound Image; BUSIS, Breast Ultrasound Image Segmentation; CAM, class activation map; CNN, Convolutional Neural Network; FIN, Fisher information network; GLCM, Gray-level cooccurrence matrix; ICER, incremental cost effectiveness ratio; OMIG, oriented, modified integrated gradient; ROIs, regions of interest; TI-RADS, Thyroid Imaging Reporting & Data System; UMAP, uniform manifold approximation and projection; XAI, explainable artificial intelligence.

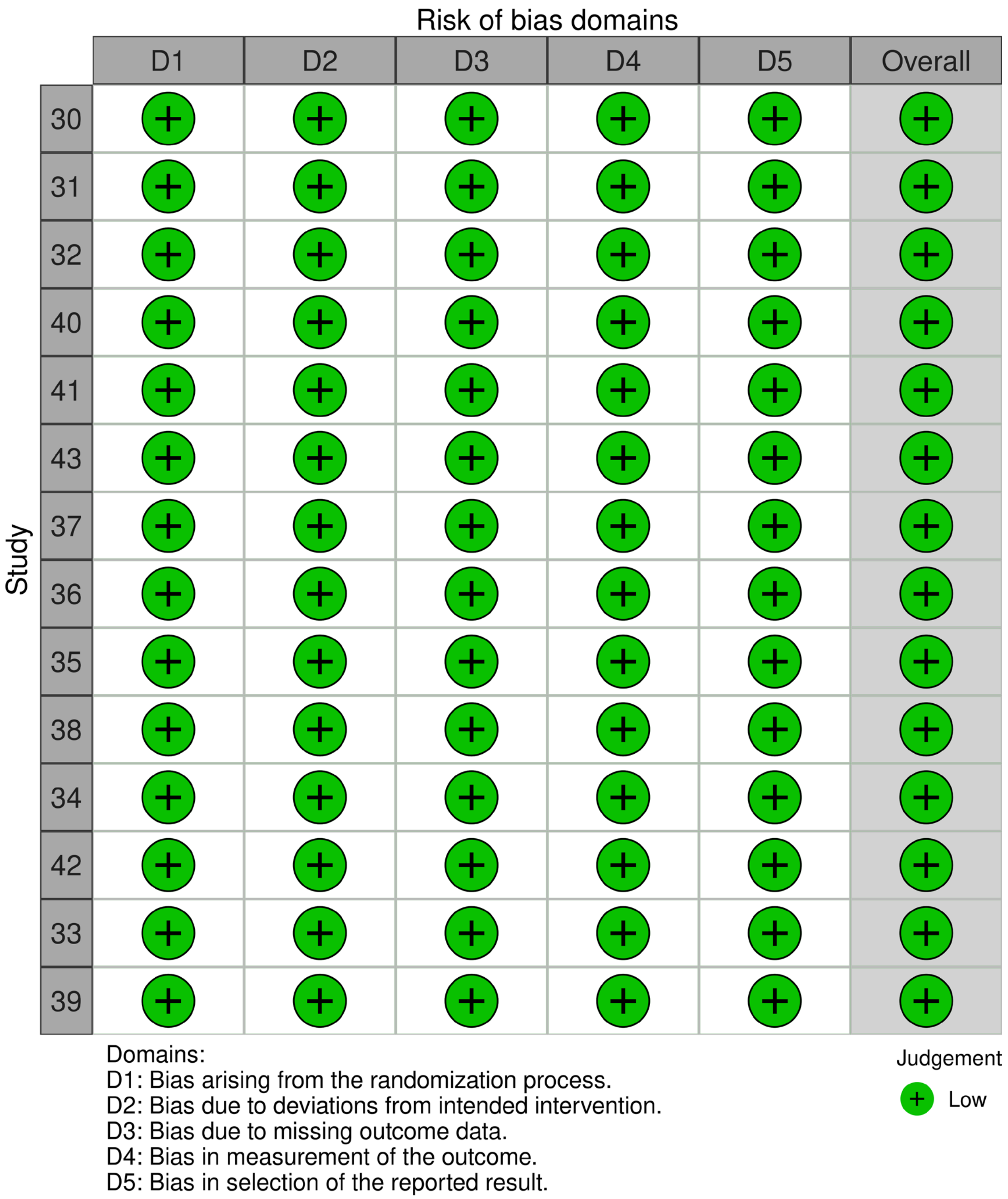

**Figure 3** Traffic light plot for risk of bias.

In Lee *et al*'s study,[36] accuracy, sensitivity, specificity and AUC were used. Simple linear iterative clustering superpixel segmentation method and the LIME explanation algorithm were employed to explain how the model makes decisions.

The area under the ROC of machine learning and an average of 10 board-certified breast radiologists were compared.[35] In this case, radiologists decreased their false-positive rates with the help of XAI. They also evaluated an independent external test dataset to prove the potential of XAI in improving the accuracy, consistency and efficiency of breast US diagnosis worldwide. The study [35] discuss accuracy of the VGG backbone to ResNet50 and EfficientNet B0 backbone was evaluated and BI-RADS descriptors were used to evaluate.[37]

In Zhang *et al*'s study,[38] accuracy, sensitivity, specificity, F1 score, R2, Mean Squared Error (MSE), Root Mean Square Error (RMSE),d shape orientation and margin were used to test the likelihood of malignancy. Explainer I was used to explain the classification results semantically. Explainer II constructs a quantitative explanation based on the classifier and Explainer I.

The study by Amanova *et al*[32] proposes and applies a new explainability method: OMIG method. The study proved that the proposed approach yields substantially more expressive and informative results for our specific use case. To avoid issues like limited meaning and confirmation bias due to low-fidelity explanations unnecessarily, Gurmessa and Jimma[8] suggest four metrics based on performance (D, R, F and S), but none of the selected studies used these metrics.

Bad stuff (bad decision, bad medical diagnosis and bad prediction) is the most common drawback of AI algorithms today. However, XAI could resolve this drawback. Robustness is also a characteristic expected from XAI. The study by Song *et al*[29] tested the robustness of the proposed framework. This study puts explainability as not only related to AI performance but also to responsibility and risk in medical diagnosis. XAI proves that the performance of algorithms is complementary but not enough alone. The complementing of both performance and explainability satisfaction increases the system's acceptance of legal and personal recognition.

## XAI and ethical challenges

XAI overcomes ethical challenges[37 38 42 43] by providing confidence, trustworthiness, transparency, accountability and interpretability in the decision-making process. It provides an opportunity to know the reason behind the prediction for patients, clinicians and doctors.[37]

The study by Song *et al*[29] recommends focusing on augmenting AI systems to extract relevant information from past US examinations as future research. Another limitation of this work is the design of the reader study.[29] A limitation of the method proposed by Ortega-Martorell *et al*[39] is that the calculation of the FI distances when creating the embedding might be slow depending on the number of data points and the sizes of the images. However, existing implementations can be used in a high-performance computing cluster which can reduce the time considerably.[39] Future studies could re-examine the cost-effectiveness of using AI to guide breast cancer screening not just among women aged 40–49 years but also in women across the entire candidate age range, including those over age 50 years.[40] To further enhance the applicability and accuracy parameters of the model, a larger dataset across multiple centres is necessary to enhance the data quality.[42] While Sun *et al*'s study[42] focuses on age groups with the highest incidence of breast cancer, future analysis encompassing older age groups would yield significant conclusions, especially about the postmenopausal population.[42] The retrospective nature of the study[42] makes it prone to selection bias[42] and also a small size dataset used.[36]

The study by Shen *et al*[35] did not provide an evaluation of patient cohorts stratified by risk factors such as family history of breast cancer and breast and ovarian cancer are the breast cancer (BRCA) gene test results and it was only provided with US images, patients' ages and notes from the operating technician.

It is important to investigate how the experience of working with these algorithms impacts the way radiologists make decisions.[34] The image's 'low-resolution' restriction remained a limitation. In future work, it is recommended to conduct a study for qualitative assessment of the level of explainability of this approach with BUS clinicians via structured interviews and questionnaires.[37] The study by Zhang *et al*[37] stated that using a more diverse dataset, trying different convolutional neural network architectures, building a multimodal model and implementing denoising algorithms can be done to improve this research.[33] It also states that combining convolutional networks with decision trees is an interesting future work.[41] To do so OMIG is used. OMIG reveals a complex pattern behind the prediction; this pattern could also be the subject of future work.[32]

Future research can also focus on augmenting AI systems to extract relevant information from past US examinations. Another limitation of this work is the design of the reader study.[29] A limitation of the method proposed by Ortega-Martorell *et al*[39] is that the calculation of the FI distance when creating the embedding might be slow depending on the number of data points and the sizes of the images. However, existing implementations can be used in a high-performance computing cluster which can reduce the time considerably.[39] Re-examine the cost-effectiveness of using AI to guide breast cancer screening not just among women aged 40–49 years but also in women across the entire candidate age range, including those over age 50 years.[40] To further enhance the applicability and accuracy parameters of their model, a larger dataset across multiple centres is necessary to enhance the data quality.[36 42] The study by Addala[33] recommended a more diverse dataset, trying different convolutional neural network architectures, building a multimodal model and implementing denoising algorithms as a future work, combining convolutional neural networks with decision trees.[41] OMIG reveals a complex pattern behind the prediction; this pattern was the subject of future work by the study.[32]

Shen *et al*'s study[35] recommends focusing on augmenting AI systems to extract relevant information from past US examinations as future research. In addition, Shen *et al*'s study[35] did not provide an evaluation of patient cohorts stratified by risk factors such as family history of BRCA gene test results. To provide a fair comparison with the AI system, readers in the study were only provided with US images, patients' ages and notes from the operating technician.[35]

Finally, it is important to investigate how the experience of working with these algorithms impacts the way radiologists make decisions.[34] The study by Zhang *et al*[37] recommended conducting a study for qualitative assessment of the level of explainability with Breast ultrasound (BUS) clinicians via structured interviews and questionnaires.

## XAI toolkits

The most popularly used toolkits that we can access from this review are DALEX and AIX360. DALEX[21 22] is a library used by R Studio. It only supports a few functionalities (ie, local post-hoc and global post-hoc), whereas AIX360[12] is a library used by Python. This toolkit supports all functionalities (ie, data explanations, directly interpretable, local post-hoc, global post-hoc and persona-specific explanations) including the evaluation matrix.

## CONCLUSION

In addition to increasing accuracy, reducing human error and technological advancement, XAI for breast cancer diagnosis overcomes ethical challenges by providing the right to know, robustness, transparency, accountability and interpretability in the decision-making process of machine learning models. However, it is not approved that it increases users' and doctors' trust in the system. Effective and systematic evaluation of its usefulness in this scenario is also lacking. Additionally, further work is needed to enhance the interpretability of deep learning algorithms through overcoming explainable to accuracy trade-offs, as well as to investigate the potential insights they can provide for clinicians' decision-making.

**Twitter** Daraje kaba Gurmessa @darajejaba

**Funding** The authors have not declared a specific grant for this research from any funding agency in the public, commercial or not-for-profit sectors.

**Competing interests** None declared.

**Patient consent for publication** Not applicable.

**Ethics approval** Not applicable.

**Provenance and peer review** Not commissioned; externally peer reviewed.

**Data availability statement** Data are available upon reasonable request.

**ORCID iD**

Daraje kaba Gurmessa http://orcid.org/0000-0002-1526-7547

## REFERENCES

1. Han H-J, Chu Y-C, Wang J, *et al*. Characteristics of breast cancers detected by screening mammography in Taiwan: a single institute's experience. *BMC Womens Health* 2023;23:330.
2. Arnold M, Morgan E, Rumgay H, *et al*. Current and future burden of breast cancer: Global statistics for 2020 and 2040. *Breast* 2022;66:15–23.
3. Lawrence RA. *2 - Anatomy of the Breast', in Breastfeeding*. Ninth Edition. Philadelphia: Elsevier, 2022: 38–57.
4. Berg WA, Bandos AI, Mendelson EB, *et al*. Ultrasound as the Primary Screening Test for Breast Cancer: Analysis From ACRIN 6666. *J Natl Cancer Inst* 2016;108:4.
5. Meenalochini G, Ramkumar S. Survey of machine learning algorithms for breast cancer detection using mammogram images. *Materials Today: Proceedings* 2021;37:2738–43.
6. Zhang Y-D, Satapathy SC, Guttery DS, *et al*. Improved Breast Cancer Classification Through Combining Graph Convolutional Network and Convolutional Neural Network. *Information Processing & Management* 2021;58:102439.
7. Zhang Y-D, Pan C, Chen X, *et al*. Abnormal breast identification by nine-layer convolutional neural network with parametric rectified linear unit and rank-based stochastic pooling. *Journal of Computational Science* 2018;27:57–68.
8. Gurmessa DK, Jimma W. A comprehensive evaluation of explainable Artificial Intelligence techniques in stroke diagnosis: A systematic review. *Cogent Engineering* 2023;10:2273088.
9. Markus AF, Kors JA, Rijnbeek PR. The role of explainability in creating trustworthy artificial intelligence for health care: A comprehensive survey of the terminology, design choices, and evaluation strategies. *J Biomed Inform* 2021;113:103655.
10. Pfeuffer N*et al*. Explanatory Interactive Machine Learning: Establishing an Action Design Research Process for Machine Learning Projects. *Business and Information Systems Engineering* 2023.
11. Carvalho DV, Pereira EM, Cardoso JS. Machine Learning Interpretability: A Survey on Methods and Metrics. *Electronics* 2019;8:832.
12. Gutti G, Arya K, Singh SK. Latent Tuberculosis Infection (LTBI) and Its Potential Targets: An Investigation into Dormant Phase Pathogens. *Mini Rev Med Chem* 2019;19:1627–42.
13. Graziani M, Dutkiewicz L, Calvaresi D, *et al*. A global taxonomy of interpretable AI: unifying the terminology for the technical and social sciences. *Artif Intell Rev* 2023;56:3473–504.
14. Klaise J, Van Looveren A, Vacanti G. Alibi explain: Algorithms for explaining machine learning models Alexandru Coca. 2021. Available: http://jmlr.org/papers/v22/21-0017.html
15. Wu L, Huang R, Tetko IV, *et al*. Trade-off Predictivity and Explainability for Machine-Learning Powered Predictive Toxicology: An in-Depth Investigation with Tox21 Data Sets. *Chem Res Toxicol* 2021;34:541–9.
16. Ribeiro PH, Orzechowski P, Wagenaar J, *et al*. Benchmarking Automl Algorithms on a collection of synthetic classification problems. 2022. Available: http://arxiv.org/abs/2212.02704
17. Ledell E, Poirier S. H2O Automl: Scalable automatic machine learning. 2020. Available: https://scinet.usda.gov/user/geospatial/#tools-and-software
18. Nori H, Jenkins S, Koch P, *et al*. Interpretml: A unified framework for machine learning Interpretability. 2019. Available: http://arxiv.org/abs/1909.09223
19. Maxwell AE, Sharma M, Donaldson KA. Explainable Boosting Machines for Slope Failure Spatial Predictive Modeling. *Remote Sensing* 2021;13:4991.
20. Rasheed K, Qayyum A, Ghaly M, *et al*. Explainable, trustworthy, and ethical machine learning for healthcare: A survey. *Comput Biol Med* 2022;149:106043.
21. Baniecki H*et al*. Dalex: responsible machine learning with interactive Explainability and fairness in python monitoring of AI regulations view project Explainable machine learning view project Dalex: responsible machine learning with interactive Explainability and fairness in python. 2021. Available: http://jmlr.org/papers/v22/20-1473.html
22. Baniecki H, Kretowicz W, Piatyszek P, *et al*. Dalex: responsible machine learning with interactive Explainability and fairness in python. 2021. Available: http://jmlr.org/papers/v22/20-1473.html
23. Egger R. Applied data science in tourism. In: Applications RE, ed. *Interpretability of Machine Learning Models', in Applied Data Science in Tourism: Interdisciplinary Approaches, Methodologies*. Cham: Springer International Publishing, 2022: 275–303.
24. Kawakura S, Hirafuji M, Ninomiya S, *et al*. Adaptations of Explainable Artificial Intelligence (XAI) to Agricultural Data Models with ELI5, PDPbox, and Skater using Diverse Agricultural Worker Data. *EJAI* 2022;1:27–34.
25. Meng C, Trinh L, Xu N, *et al*. Interpretability and fairness evaluation of deep learning models on MIMIC-IV dataset. *Sci Rep* 2022;12:7166.
26. Page MJ, McKenzie JE, Bossuyt PM, *et al*. The PRISMA 2020 statement: an updated guideline for reporting systematic reviews. *BMJ* 2021;372:n71.
27. Gurmessa WJD. Explainable machine learning for breast cancer diagnosis from Mammography and ultrasound images: A systematic review; 2023.
28. McGuinness LA, Higgins JPT. Risk-of-bias VISualization (robvis): An R package and Shiny web app for visualizing risk-of-bias assessments. *Res Synth Methods* 2021;12:55–61.
29. Song D, Yao J, Jiang Y, *et al*. A new xAI framework with feature explainability for tumors decision-making in Ultrasound data: comparing with Grad-CAM. *Comput Methods Programs Biomed* 2023;235:107527.
30. Oh J-H, Kim H-G, Lee KM, *et al*. Reliable quality assurance of X-ray mammography scanner by evaluation the standard mammography phantom image using an interpretable deep learning model. *Eur J Radiol* 2022;154:110369.
31. Qian X, Pei J, Zheng H, *et al*. Prospective assessment of breast cancer risk from multimodal multiview ultrasound images via clinically applicable deep learning. *Nat Biomed Eng* 2021;5:522–32.
32. Amanova N, Martin J, Elster C. Explainability for deep learning in mammography image quality assessment. *Mach Learn: Sci Technol* 2022;3:025015.
33. Addala V. BREAST AI: low cost, Explainable artificial intelligence based App for efficient diagnosis of breast cancer in developing areas. 2023 IEEE 3rd International Conference on Electronic

Communications, Internet of Things and Big Data (ICEIB); Taichung, Taiwan.2023:164–7

34 Rezazade Mehrizi MH, Mol F, Peter M, *et al*. The impact of AI suggestions on radiologists' decisions: a pilot study of explainability and attitudinal priming interventions in mammography examination. *Sci Rep* 2023;13:1.

35 Shen Y, Shamout FE, Oliver JR, *et al*. Artificial intelligence system reduces false-positive findings in the interpretation of breast ultrasound exams. *Nat Commun* 2021;12:1.

36 Lee Y-W, Huang C-S, Shih C-C, *et al*. Axillary lymph node metastasis status prediction of early-stage breast cancer using convolutional neural networks. *Comput Biol Med* 2021;130:104206.

37 Zhang B, Vakanski A, Xian M. BI-RADS-NET: AN EXPLAINABLE MULTITASK LEARNING APPROACH FOR CANCER DIAGNOSIS IN BREAST ULTRASOUND IMAGES. IEEE Int Workshop Mach Learn Signal Process 2021;2021:1–6.

38 Zhang B, Vakanski A, Xian M. BI-RADS-NET-V2: A Composite Multi-Task Neural Network for Computer-Aided Diagnosis of Breast Cancer in Ultrasound Images With Semantic and Quantitative Explanations. *IEEE Access* 2023;11:79480–94.

39 Ortega-Martorell S, Riley P, Olier I, *et al*. Breast cancer patient characterisation and visualisation using deep learning and fisher information networks. *Sci Rep* 2022;12:14004.

40 Mital S, Nguyen HV. Cost-effectiveness of using artificial intelligence versus polygenic risk score to guide breast cancer screening. *BMC Cancer* 2022;22:501.

41 Rezazadeh A, Jafarian Y, Kord A. Explainable Ensemble Machine Learning for Breast Cancer Diagnosis Based on Ultrasound Image Texture Features. *Forecasting* 2022;4:262–74.

42 Sun J, Sun C-K, Tang Y-X, *et al*. Application of SHAP for Explainable Machine Learning on Age-Based Subgrouping Mammography Questionnaire Data for Positive Mammography Prediction and Risk Factor Identification. *Healthcare (Basel)* 2023;11:2000.

43 Dong F, She R, Cui C, *et al*. One step further into the blackbox: a pilot study of how to build more confidence around an AI-based decision system of breast nodule assessment in 2D ultrasound. *Eur Radiol* 2021;31:4991–5000.

BMJ Health & Care Informatics

# Performance of large language models on advocating the management of meningitis: a comparative qualitative study

Urs Fisch ,[1] Paulina Kliem,[2] Pascale Grzonka,[2] Raoul Sutter[1,2,3]

**To cite:** Fisch U, Kliem P, Grzonka P, *et al*. Performance of large language models on advocating the management of meningitis: a comparative qualitative study. *BMJ Health Care Inform* 2024;**31**:e100978. doi:10.1136/bmjhci-2023-100978

► Additional supplemental material is published online only. To view, please visit the journal online (http://dx.doi.org/10.1136/bmjhci-2023-100978).

Received 24 November 2023
Accepted 15 January 2024

[1]Department of Neurology, University Hospital Basel, Basel, Switzerland
[2]Clinic for Intensive Care Medicine, University Hospital Basel, Basel, Switzerland
[3]Medical Faculty, University Basel, Basel, Switzerland

**Correspondence to**
Dr Urs Fisch; urs.fisch@usb.ch

## ABSTRACT

**Objectives** We aimed to examine the adherence of large language models (LLMs) to bacterial meningitis guidelines using a hypothetical medical case, highlighting their utility and limitations in healthcare.

**Methods** A simulated clinical scenario of a patient with bacterial meningitis secondary to mastoiditis was presented in three independent sessions to seven publicly accessible LLMs (Bard, Bing, Claude-2, GTP-3.5, GTP-4, Llama, PaLM). Responses were evaluated for adherence to good clinical practice and two international meningitis guidelines.

**Results** A central nervous system infection was identified in 90% of LLM sessions. All recommended imaging, while 81% suggested lumbar puncture. Blood cultures and specific mastoiditis work-up were proposed in only 62% and 38% sessions, respectively. Only 38% of sessions provided the correct empirical antibiotic treatment, while antiviral treatment and dexamethasone were advised in 33% and 24%, respectively. Misleading statements were generated in 52%. No significant correlation was found between LLMs' text length and performance (r=0.29, p=0.20). Among all LLMs, GTP-4 demonstrated the best performance.

**Discussion** Latest LLMs provide valuable advice on differential diagnosis and diagnostic procedures but significantly vary in treatment-specific information for bacterial meningitis when introduced to a realistic clinical scenario. Misleading statements were common, with performance differences attributed to each LLM's unique algorithm rather than output length.

**Conclusions** Users must be aware of such limitations and performance variability when considering LLMs as a support tool for medical decision-making. Further research is needed to refine these models' comprehension of complex medical scenarios and their ability to provide reliable information.

### WHAT IS ALREADY KNOWN ON THIS TOPIC

⇒ Large language models (LLMs) have demonstrated proficiency in responding to medical licensing examination-level queries and shown aptitude in accurate medical triage decision-making. However, performance with knowledge-testing scenarios is not necessarily indicative of effectiveness in real-world medical contexts.

### WHAT THIS STUDY ADDS

⇒ This investigation presents a qualitative analysis of the performance of seven publicly accessible LLMs, using a stepwise presentation of a hypothetical bacterial meningitis case reflecting a real-world scenario. While LLMs generally offered helpful triage and diagnostic advice, there were significant discrepancies in their recommendations for treatment and specific diagnostic work-ups. Moreover, the generation of misleading statements and variability in performances between different sessions were observed among individual LLMs.

### HOW THIS STUDY MIGHT AFFECT RESEARCH, PRACTICE OR POLICY

⇒ This study highlights the current capabilities of LLMs in handling real-world medical emergency situations and identifies areas of future research, such as enhancing LLMs' understanding of complex medical scenarios and their capacity for delivering reliable and deterministic information.

## INTRODUCTION

Large language models (LLMs) are powerful artificial intelligence (AI) models trained on extensive text data to generate human-like text. They can interpret user-generated textual instructions (prompts) and respond immediately with the contextually most appropriate response based on probabilistic computations learnt during their training. Lately, several LLMs were released to the public, attracting substantial attention for their chat-like interfaces requiring no technical prerequisites.

Recently, both trained and untrained LLMs have shown proficiency in handling medical licensing examination-level questions and demonstrated the ability to make rapid and accurate judgments in medical triage and diagnosing or provide helpful information to patients, underscoring their potential applicability in the healthcare sector.[1–6] However, the ability to perform well in knowledge-testing vignettes does not fully reflect the needs of real-world medical settings which demand parallel work-up and nuanced decision-making on the basis of sometimes incomplete information. Considering that physicians already frequently use internet resources for diagnostic decisions and treatment options and that not all hospitals may have free access to the medical literature, it is likely that LLMs will be increasingly used as potential aids in clinical practice.[7–9] However, a deeper understanding of their potential and limitations is essential for an appropriate use.[10–12]

This study explored the potentials and limitations of current LLMs by presenting these models with a predefined hypothetical but typical scenario of a patient with acute bacterial meningitis. The aim was to analyse their performance and alignment with good clinical practice and established medical guidelines regarding suggested diagnostic and treatment measures. Bacterial meningitis was chosen for its life-threatening nature, urgency required in diagnosis and treatment and the range of differential diagnoses it involves, making it ideal for assessing the performance of LLMs in a realistic and high stakes medical scenario.

## METHODS

Seven publicly accessible LLMs were evaluated between 5 and 8 August 2023: Bard by Google, Bing by Microsoft, generative pre-trained transformer (GTP)-3.5 by OpenAI, GTP-4 by OpenAI (accessed via Poe (Quora)), Claude-2 by Anthropic PBC (accessed via Poe), pathways language model (PaLM) 2 chat-bison-001 by Google (accessed via Poe) and Llama-2-70b by Meta Platforms (accessed via Poe).

Each LLM was presented with the same hypothetical scenario of a patient presenting with symptoms of acute bacterial meningitis (as outlined below) three times within 3 days. The actual diagnosis was not provided. For the LLM Bard, the settings were chosen to inhibit intersession information storage. All other LLMs claimed that they are incapable of storing user information between sessions. Each session was initiated with a context clearance of previous conversations.

### Hypothetical scenario of a patient with acute bacterial meningitis

The patient vignette described a clinical scenario of a patient with acute symptoms due to pneumococcal meningitis secondary to mastoiditis without providing definite diagnosis. The text of the inputted case vignette and the subsequent follow-up queries consisted of five text blocks that were predefined and presented unchanged to each LLM in every session (online supplemental table 1). Given that the performance of LLMs is heavily influenced by prompting,[13] the initial question began with a contextualisation wherein the LLM was asked to act as an ‘experienced medical assistant’ and the user was identified as a ‘junior medical doctor’ seeking advice for a 52-year-old female patient suffering from severe headache and confusion, followed by an open-ended question about the next steps. This prompt engaged all LLMs in a conversation about the hypothetical case. Second, a detailed vignette was presented, depicting the medical history (notably acute headache and confusion, a history of diabetes type 2 and migraine), vital signs (tachycardia and fever) and prominent abnormal clinical findings (ie, a Glasgow Coma Scale (GCS) of 12 with lethargy, disorientation, fast downward drift of extremities, absence of stiff neck, signs of inflammatory skin of the right mastoid), followed by the open-ended request for a detailed step-by-step recommendation of how to proceed. Third, two closed-ended questions were asked: (1) if a computer tomography (CT) scan of the head needs to be awaited before lumbar puncture (LP) and (2) if administration of antibiotics should be delayed until LP has been performed. Fourth, the exact dosages of antibiotics were asked. Fifth, an open-ended question was asked about any other considerations regarding the treatment or work-up.

The case was created to reflect clinical reality and not a medical license examination question, meaning that information was presented stepwise and reflected a realistic clinical case where not all typical signs and symptoms are necessarily present from the beginning. For example, neck stiffness has shown to have a low sensitivity and as such, its absence cannot rule out meningitis.[14] A search for an infectious focus is crucial and patients should be examined for otitis media or mastoiditis.[15] By this design we aimed to challenge the LLMs in multiple aspects, including good clinical practice, possible differential diagnoses and consideration of risk factors and comorbidities, such as age, diabetes and migraine, for diagnosis and treatment.

### Evaluation of LLM performance

The Infectious Diseases Society of America (IDSA) and the European Society of Clinical Microbiology and Infectious Diseases (ESCMID) guidelines were chosen as references as they have previously both been shown in a systematic review to be excellent clinical management guidelines for bacterial meningitis with multinational validity (online supplemental table 2, right column).[14 16 17]

Individual responses from the LLMs underwent two temporally separated qualitative assessments (accomplished vs unaccomplished) of predefined tasks (online supplemental table 2, middle column) in adherence with good clinical practice and the reference guidelines.[14–18] Accomplished tasks were summarised to a qualitative performance summary. Response consistency was defined as the percentage of responded tasks that were assessed identically (regardless of accomplished or unaccomplished) across all sessions of an individual LLM. In cases where an LLM declined to respond to a question, the corresponding tasks were excluded from the assessment.

As the two reference guidelines differently define criteria for imaging before LP (ie, according to the IDSA guideline, a scan of the brain would be required as

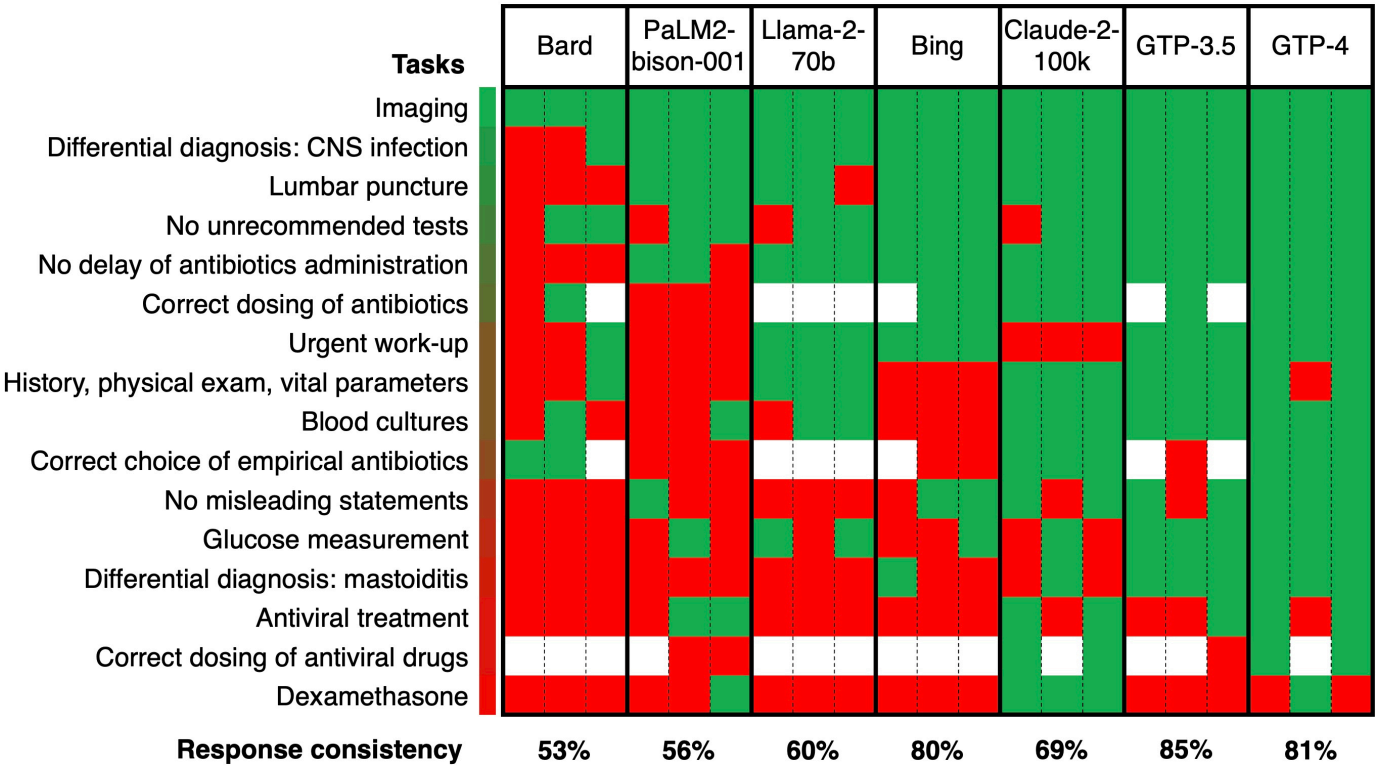

**Figure 1** Qualitative assessment of large language models (LLMs) performance on a case of bacterial meningitis. Each LLM was tested three times with a standardised case vignette (individual sessions separated by dashed lines). Accomplished tasks are marked in green in decreasing order of agreement among all LLMs, while unaccomplished tasks are highlighted in red. White boxes represent tasks where the model either declined to respond or no additional information could be provided due to gaps in previous responses. Response consistency was defined as identically assessed responded tasks across different sessions of a single LLM. CNS, central nervous system.

the patient expresses any altered mental status and has downward drift of extremities, whereas according to the ESCMID guideline, a scan of the brain is not mandatory with a GCS>10) and maximal allowed delay to start antibiotics, these aspects were not included in the qualitative performance summary.[14 16]

## Statistics

Descriptive statistics with numbers and percentages and the two-sided Pearson correlation coefficient were used where appropriate (R, V.4.3.1). Due to the principally qualitative aim of this study, a statistical comparison between the LLMs was not intended.

## RESULTS

The individual responses of all 21 sessions of the seven LLMs are summarised in figure 1. We noticed marked differences in the qualitative performance summary between different LLMs and to a lesser extent also between different sessions of individual LLMs. The response consistency ranged from 53% to 85%. LLMs with low numbers of accomplished tasks also had low response consistency. Among all the LLMs evaluated, GPT-4 demonstrated the most consistent performance, effectively addressing almost all tasks and having a high response consistency across all tasks and responses. Exemplary transcripts of the first conversations with Bard and GTP-4 are shown in online supplemental material.

The word count of individual LLMs sessions varied significantly, ranging from 325 (PaLM 2 chat-bison-001) to 2045 (GTP-3.5), with an average of 1270 words (standard deviation 477). There was no significant correlation (r=0.29, p=0.20) between the total length of individual LLM responses and the summative performance of accomplished tasks, indicating that simply generating more text output does not necessarily lead to improved performance.

### Suggested differential diagnoses and recommended diagnostic work-up

In 62% of the sessions, LLMs suggested an urgent work-up without direct prompting. In 57% of sessions, they recommended measuring vital parameters, taking the patient's history and performing a physical examination as initial steps. Furthermore, in 90% of the sessions, the LLMs accurately suspected a central nervous system (CNS) infection as a possible cause of the patient's symptoms. However, only 38% of the responses mentioned mastoiditis as a potential underlying cause or suggested correspondent diagnostic procedures (imaging with purpose of investigating mastoiditis, otoscopy, ear–nose–throat consultation). The most frequently mentioned differential diagnoses were stroke (86%), followed by intracranial/subarachnoid haemorrhage and brain tumour (both 48%). Other proposed differential diagnoses were migraine (19%), metabolic/endocrine disbalances (19%), medication side effects (10%), non-CNS infections (10%), severe hypertension (5%), drug intoxication (5%) and neurodegenerative disorders (5%).

Regarding diagnostic work-up, cranial imaging was recommended in 100% of sessions, LP in 81% and blood cultures in 62%. Blood glucose measurement in the diabetic patient with altered mental status was suggested in 53%. Unrecommended tests by the IDSA and ESCMID guidelines (eg, electroencephalogram, electrocardiogram, chest radiography) were proposed in 19% of sessions as an initial work-up.

In 43% of responses, LLMs stated that a cranial CT scan is necessary before LP, while 14% suggested to perform an LP without CT scan and another 43% gave unclear answers. Only three LLMs (GTP-3.5, Claude-2, GTP-4) provided a case-specific rationale for their recommendation (92% responses suggested CT scan before LP). Due to different definitions of criteria for cranial imaging before LP in the reference guidelines and maximal allowed delay to start antibiotics,[14 16] these aspects were not included in the qualitative performance summary displayed in figure 1.

### Recommended treatment

Regarding treatment, 81% of responses stated that rapid administration of antibiotics is necessary. The correct choice of empirical antibiotic treatment, consisting of a third-generation cephalosporin with ampicillin (alternatives: amoxicillin, penicillin G) with or without vancomycin, was provided in 38%, and of those, almost 90% with correct dosing.[14 16] Another 29% provided an incomplete choice of antibiotic treatment and 33% declined to comment on any choice of antibiotics. In 33% of the sessions, antiviral treatment was considered with approximately half of them providing correct dosing. Dexamethasone administration was recommended in 24% of all responses.

### Misleading statements

Misleading statements were identified in 52% of the sessions, such as performing an LP to relieve intracranial pressure or carrying it out prior to imaging in order to facilitate image interpretation; administering prophylactic antiseizure medication or giving benzodiazepines for sedation; adjusting ceftriaxone dosage based on age, weight and kidney function or administering dexamethasone for meningococcal meningitis; the presence of a stiff neck and Kernig's sign (while the vignette stated that these were absent); or the misinterpretation of mastoiditis as herpes zoster ophthalmicus.

## DISCUSSION

This study investigated qualitative performance characteristics of different LLMs when challenged with a hypothetical clinical case of an adult patient with bacterial meningitis and revealed marked discrepancies between the LLMs. This reflects both the potentials and limitations of these models when used as a guidance for medical work-up and treatment.[9] The differences in qualitative performances observed among the LLMs did not demonstrate a correlation with the length of their respective outputs. This suggests that the performance variations can be attributed to the unique algorithmic designs of each LLM rather than their quantitative output.

CNS infection was identified as a probable cause among other differential diagnosis in the majority of cases and almost all LLMs succeeded in identifying and recommending appropriate investigations, including cranial imaging and LP. A fair proportion underscored the need for urgent diagnostics and antibiotic treatment. These results align with previous findings demonstrating a satisfactory performance of GTP-3 (the predecessor of GTP-3.5) in terms of triage and reasoning on differential diagnoses and the high performance of GTP-4 in diagnostic case challenges.[4 19–21] Our study expands on these findings by examining an additional five LLMs which were not available at the time of the previous studies.

Our investigation also highlights limitations of most LLMs regarding their understanding of case complexity and their ability to link different disease entities. For instance, the identification of mastoiditis as an underlying cause was mentioned infrequently, as were blood glucose measurements, drawing blood cultures, considerations of empirical antiviral treatment and the administration of dexamethasone. The considerable heterogeneity in the responses of individual LLMs, despite standardised prompts, raises further concerns about their reliability and consistency. The presentation of misleading statements in more than half of the LLM sessions underscores the potential risk that comes along with their use for critical medical decision-making, especially in complex, life-threatening and time-sensitive situations, such as with bacterial meningitis. Such challenges must be addressed in future research when developing tools on the basis of LLMs for medical purposes.[10–12]

Most LLMs' inability to provide definitive guidance on whether to conduct a cranial CT scan before an LP might be due to the differences in the guidelines.[14 16] However, the lack of clear direction in many LLM responses could also suggest an insufficiency in handling complex clinical situations where there is a need for reasoned decision-making. This finding may be viewed in the context of the research gap between healthcare AI development and the challenge of its validation and implementation in real-world clinical settings.[22–24]

### Limitations

Our study has several limitations. Most importantly, none of the LLMs was designed to assist in medical diagnostics and treatment and most correctly included respective disclaimers. However, as LLMs are powerful, new and easily accessible AI tools, it is highly probable that they will find increasing use in the health sector, which justifies studying their reliability and applicability.[1–6] Further, prompting has significant influence on the result.[13] While our study did not explore the impact of different prompting strategies, we used standardised prompts, which included contextualisation and step-by-step reasoning, to ensure comparability between LLMs. Although we evaluated the LLMs' intuitive assessment of the scenario's urgency, we did not directly inquire this in the prompts. In addition, the selection of tasks for the qualitative assessment was unweighted and focused on important initial management steps, while other aspects, such as laboratory testing procedures or duration of antimicrobial treatment, were not investigated. Lastly, the study was limited to a single case scenario, and the results may not be generalisable to other clinical scenarios. Thus, we refrained from an absolute ranking of the LLMs.

## CONCLUSIONS

The latest versions of LLMs show potential in helping healthcare professionals. Our study underscores the need for cautious and informed use of most of these models as demonstrated by the limitations in providing specific information and potentially misleading information for diagnostic work-up and treatment of adult patients with bacterial meningitis. Users should be aware of the variability in their performance.

Further research is needed to refine these models and enhance their understanding of complex medical scenarios and their ability to provide deterministic, reliable information regardless of prompt nuances. Concurrently, efforts are necessary to mitigate the potential for disseminating erroneous content.

**Contributors** UF and RS planned and designed the study. UF acquired the data and wrote the first draft of the manuscript and is responsible for the overall content as guarantor. All authors interpreted the data, revised the manuscript for important intellectual content and approved the final submitted version.

**Funding** The authors have not declared a specific grant for this research from any funding agency in the public, commercial or not-for-profit sectors.

**Competing interests** None declared.

**Patient consent for publication** Not applicable.

**Ethics approval** As no real patients were involved in this study, ethical approval was not required.

**Provenance and peer review** Not commissioned; externally peer reviewed.

**Data availability statement** Data are available upon reasonable request.

**ORCID iD**

Urs Fisch http://orcid.org/0000-0003-1557-9062

## REFERENCES

1 Singhal K, Azizi S, Tu T, *et al*. Large language models encode clinical knowledge. *Nature* 2023;620:172–80.
2 Kung TH, Cheatham M, Medenilla A, *et al*. Performance of ChatGPT on USMLE: potential for AI-assisted medical education using large language models. *PLOS Digit Health* 2023;2:e0000198.
3 Gilson A, Safranek CW, Huang T, *et al*. How does ChatGPT perform on the United States medical licensing examination? The implications of large language models for medical education and knowledge assessment. *JMIR Med Educ* 2023;9:e45312.
4 Levine DM, Tuwani R, Kompa B, *et al*. The diagnostic and triage accuracy of the GPT-3 artificial intelligence model. *Health Informatics* [Preprint] 2023.
5 Ayers JW, Poliak A, Dredze M, *et al*. Comparing physician and artificial intelligence chatbot responses to patient questions posted to a public social media forum. *JAMA Intern Med* 2023;183:589–96.
6 Haver HL, Ambinder EB, Bahl M, *et al*. Appropriateness of breast cancer prevention and screening recommendations provided by ChatGPT. *Radiology* 2023;307:e230424.
7 Tang H, Ng JHK. Googling for a diagnosis--use of Google as a diagnostic aid: internet based study. *BMJ* 2006;333:1143–5.
8 Russell-Rose T, Chamberlain J. Expert search strategies: the information retrieval practices of healthcare information professionals. *JMIR Med Inform* 2017;5:e33.
9 Lee P, Bubeck S, Petro J. Benefits, limits, and risks of GPT-4 as an AI Chatbot for medicine. *N Engl J Med* 2023;388:1233–9.
10 Howard A, Hope W, Gerada A. ChatGPT and antimicrobial advice: the end of the consulting infection doctor? *Lancet Infect Dis* 2023;23:405–6.
11 Liévin V, Hother CE, Motzfeldt AG, *et al*. Can large language models reason about medical questions? *arXiv:220708143* 2023. Available: https://doi.org/10.48550/arXiv.2207.08143
12 Norori N, Hu Q, Aellen FM, *et al*. Addressing bias in big data and AI for health care: a call for open science. *Patterns (N Y)* 2021;2:100347.
13 Wang J, Shi E, Yu S, *et al*. Prompt engineering for healthcare: methodologies and applications. 2023. Available: https://doi.org/10.48550/arXiv.2304.14670
14 van de Beek D, Cabellos C, Dzupova O, *et al*. ESCMID guideline: diagnosis and treatment of acute bacterial meningitis. *Clin Microbiol Infect* 2016;22 Suppl 3:S37–62.
15 Dyckhoff-Shen S, Koedel U, Pfister H-W, *et al*. SOP: emergency workup in patients with suspected acute bacterial meningitis. *Neurol Res Pract* 2021;3:2.
16 Tunkel AR, Hartman BJ, Kaplan SL, *et al*. Practice guidelines for the management of bacterial meningitis. *Clin Infect Dis* 2004;39:1267–84.
17 Sigfrid L, Perfect C, Rojek A, *et al*. A systematic review of clinical guidelines on the management of acute, community-acquired CNS infections. *BMC Med* 2019;17:170.
18 Steiner I, Budka H, Chaudhuri A, *et al*. Viral meningoencephalitis: a review of diagnostic methods and guidelines for management. *Eur J Neurol* 2010;17:999–e57.
19 Nori H, King N, McKinney SM, *et al*. Capabilities of GPT-4 on medical challenge problems. 2023. Available: https://doi.org/10.48550/arXiv.2303.13375
20 Eriksen AV, Möller S, Ryg J. Use of GPT-4 to diagnose complex clinical cases. *NEJM AI* 2023;1.
21 Kanjee Z, Crowe B, Rodman A. Accuracy of a generative artificial intelligence model in a complex diagnostic challenge. *JAMA* 2023;330:78–80.
22 Yin J, Ngiam KY, Teo HH. Role of artificial intelligence applications in real-life clinical practice: systematic review. *J Med Internet Res* 2021;23:e25759.
23 Susanto AP, Lyell D, Widyantoro B, *et al*. Effects of machine learning-based clinical decision support systems on decision-making, care delivery, and patient outcomes: a scoping review. *J Am Med Inform Assoc* 2023;30:2050–63.
24 Gama F, Tyskbo D, Nygren J, *et al*. Implementation frameworks for artificial intelligence translation into health care practice: scoping review. *J Med Internet Res* 2022;24:e32215.