BMJ Health & Care Informatics

COP27 climate change conference: urgent action needed for Africa and the world

Chris Zielinski, on behalf of the authorship group listed below

To cite: Zielinski C, . COP27 climate change conference: urgent action needed for Africa and the world. *BMJ Health Care Inform* 2022;**29**:e100690. doi:10.1136/bmjhci-2022-100690

Accepted 10 October 2022

Centre for Global Health, Faculty of Health and Wellbeing, UK Health Alliance on Climate Change, London, UK

Correspondence to
Dr Chris Zielinski, UK Health Alliance on Climate Change, London, UK;
chris.zielinski@ukhealthalliance.org

Wealthy nations must step up support for Africa and vulnerable countries in addressing past, present and future impacts of climate change

The 2022 report of the Intergovernmental Panel on Climate Change paints a dark picture of the future of life on earth, characterised by ecosystem collapse, species extinction and climate hazards such as heatwaves and floods.[1] These are all linked to physical and mental health problems, with direct and indirect consequences of increased morbidity and mortality. To avoid these catastrophic health effects across all regions of the globe, there is broad agreement—as 231 health journals argued together in 2021—that the rise in global temperature must be limited to <1.5°C compared with pre-industrial levels.

While the Paris Agreement of 2015 outlines a global action framework that incorporates providing climate finance to low-income and middle-income countries, this support has yet to materialise.[2] COP27 is the fifth Conference of the Parties (COP) to be organised in Africa since its inception in 1995. Ahead of this meeting, we—as health journal editors from across the continent—call for urgent action to ensure it is the COP that finally delivers climate justice for Africa and vulnerable countries. This is essential for the health of those countries, and for the health of the whole world.

AFRICA HAS SUFFERED DISPROPORTIONATELY, ALTHOUGH IT HAS DONE LITTLE TO CAUSE THE CRISIS

The climate crisis has had an impact on the environmental and social determinants of health across Africa, leading to devastating health effects.[3] Impacts on health can result directly from environmental shocks and indirectly through socially mediated effects.[4] Climate change-related risks in Africa include flooding, drought, heatwaves, reduced food production and reduced labour productivity.[5]

Droughts in sub-Saharan Africa have tripled between 1970–1979 and 2010–2019.[6] In 2018, devastating cyclones impacted 2.2 million people in Malawi, Mozambique and Zimbabwe.[6] In west and central Africa, severe flooding resulted in mortality and forced migration from loss of shelter, cultivated land and livestock.[7] Changes in vector ecology brought about by floods and damage to environmental hygiene has led to increases in diseases across sub-Saharan Africa, with rises in malaria, dengue fever, Lassa fever, Rift Valley fever, Lyme disease, Ebola virus, West Nile virus and other infections.[8 9] Rising sea levels reduce water quality, leading to waterborne diseases, including diarrhoeal diseases, a leading cause of mortality in Africa.[8] Extreme weather damages water and food supply, increasing food insecurity and malnutrition, which causes 1.7 million deaths annually in Africa.[10] According to the Food and Agriculture Organization of the United Nations, malnutrition has increased by almost 50% since 2012, owing to the central role agriculture plays in African economies.[11] Environmental shocks and their knock-on effects also cause severe harm to mental health.[12] In all, it is estimated that the climate crisis has destroyed a fifth of the gross domestic product of the countries most vulnerable to climate shocks.[13]

The damage to Africa should be of supreme concern to all nations. This is partly for moral reasons. It is highly unjust that the most impacted nations have contributed the least to global cumulative emissions, which are driving the climate crisis and its increasingly severe effects. North America and Europe have contributed 62% of carbon dioxide emissions since the Industrial Revolution, whereas Africa has contributed only 3%.[14]

THE FIGHT AGAINST THE CLIMATE CRISIS NEEDS ALL HANDS ON DECK

Yet it is not just for moral reasons that all nations should be concerned for Africa.

The acute and chronic impacts of the climate crisis create problems like poverty, infectious disease, forced migration and conflict that spread through globalised systems.[6 15] These knock-on impacts affect all nations. COVID-19 served as a wake-up call to these global dynamics and it is no coincidence that health professionals have been active in identifying and responding to the consequences of growing systemic risks to health. But the lessons of the COVID-19 pandemic should not be limited to pandemic risk.[16 17] Instead, it is imperative that the suffering of frontline nations, including those in Africa, be the core consideration at COP27: in an interconnected world, leaving countries to the mercy of environmental shocks creates instability that has severe consequences for all nations.

The primary focus of climate summits remains to rapidly reduce emissions so that global temperature rises are kept to below 1.5°C. This will limit the harm. But, for Africa and other vulnerable regions, this harm is already severe. Achieving the promised target of providing US$100 billion of climate finance a year is now globally critical if we are to forestall the systemic risks of leaving societies in crisis. This can be done by ensuring these resources focus on increasing resilience to the existing and inevitable future impacts of the climate crisis, as well as on supporting vulnerable nations to reduce their greenhouse gas emissions: a parity of esteem between adaptation and mitigation. These resources should come through grants not loans, and be urgently scaled up before the current review period of 2025. They must put health system resilience at the forefront, as the compounding crises caused by the climate crisis often manifest in acute health problems. Financing adaptation will be more cost-effective than relying on disaster relief.

Some progress has been made on adaptation in Africa and around the world, including early warning systems and infrastructure to defend against extremes. But frontline nations are not compensated for impacts from a crisis they did not cause. This is unfair, and drives the spiral of global destabilisation, as nations pour money into responding to disasters, but can no longer afford to pay for greater resilience or to reduce the root problem through emissions reduction. A financing facility for loss and damage must now be introduced, providing additional resources beyond those given for mitigation and adaptation. This must go beyond the failures of COP26, where the suggestion of such a facility was downgraded to 'a dialogue'.[18]

The climate crisis is a product of global inaction, and comes at great cost to disproportionately impacted African countries, and to the whole world. Africa is united with other frontline regions in urging wealthy nations to finally step up, if for no other reason than that the crises in Africa will sooner rather than later spread and engulf all corners of the globe, by which time it may be too late to effectively respond. If so far they have failed to be persuaded by moral arguments, then hopefully their self-interest will now prevail.

List of Authors

Lukoye Atwoli, Editor-in-Chief, *East African Medical Journal*; Gregory E. Erhabor, Editor-in-Chief, *West African Journal of Medicine*; Aiah A. Gbakima, Editor-in-Chief, *Sierra Leone Journal of Biomedical Research*; Abraham Haileamlak, Editor-in-Chief, *Ethiopian Journal of Health Sciences*; Jean-Marie Kayembe Ntumba, Chief Editor, *Annales Africaines de Medecine*; James Kigera, Editor-in-Chief, *Annals of African Surgery*; Laurie Laybourn-Langton, University of Exeter; Bob Mash, Editor-in-Chief, *African Journal of Primary Health Care & Family Medicine*; Joy Muhia, London School of Medicine and Tropical Hygiene; Fhumulani Mavis Mulaudzi, Editor-in-Chief, *Curationis*; David Ofori-Adjei, Editor-in-Chief, *Ghana Medical Journal*; Friday Okonofua, Editor-in-Chief, *African Journal of Reproductive Health*; Arash Rashidian, Executive Editor, and Maha El-Adawy, Director of Health Promotion, *Eastern Mediterranean Health Journal*; Siaka Sidibé, Director of Publication, *Mali Médical*; Abdelmadjid Snouber, Managing Editor, *Journal de la Faculté de Médecine d'Oran*; James Tumwine, Editor-in-Chief, *African Health Sciences*; Mohammad Sahar Yassien, Editor-in-Chief, *Evidence-Based Nursing Research*; Paul Yonga, Managing Editor, *East African Medical Journal*; Lilia Zakhama, Editor-in-Chief, *La Tunisie Médicale*; Chris Zielinski, University of Winchester.

This comment is being published simultaneously in multiple journals. For the full list of journals, see: https://www.bmj.com/content/full-list-authors-and-signatories-climate-emergency-editorial-october-2022

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 Commissioned; internally peer reviewed.

REFERENCES

1 IPCC. Climate change 2022: impacts, adaptation and vulnerability. Working group II contribution to the IPCC sixth assessment report 2022.
2 UN. The Paris Agreement: United Nations, 2022. Available: https://www.un.org/en/climatechange/paris-agreement [Accessed 12 Sep 2022].
3 Climate change and health in sub-Saharan Africa: the case of Uganda. climate investment funds 2020.
4 WHO. Strengthening health resilience to climate change 2016.
5 Trisos CH, Adelekan IO, Totin E, *et al*. Africa. In: climate change 2022: impacts, adaptation, and vulnerability, 2022. Available: https://www.ipcc.ch/report/ar6/wg2/ [Accessed 26 Sep 2022].
6 Climate change adaptation and economic transformation in sub-Saharan Africa. World Bank 2021.

7 Opoku SK, Leal Filho W, Hubert F, *et al*. Climate change and health preparedness in Africa: analysing trends in six African countries. *Int J Environ Res Public Health* 2021;18:4672.
8 Evans MS, Munslow B. Climate change, health, and conflict in Africa's Arc of instability. *Perspect Public Health* 2021;141:338–41.
9 Stawicki SP, Papadimos TJ, Galwankar SC, *et al*. Reflections on climate change and public health in Africa in an era of global pandemic. contemporary developments and perspectives in international health security. 2: Intechopen 2021.
10 Climate change and health in Africa: issues and options: African climate policy centre, 2013. Available: https://archive.uneca.org/sites/default/files/PublicationFiles/policy_brief_12_climate_change_and_health_in_africa_issues_and_options.pdf [Accessed 12 Sep 2022].
11 Climate change is an increasing threat to Africa, 2020. Available: https://unfccc.int/news/climate-change-is-an-increasing-threat-to-africa [Accessed 12 Sep 2022].
12 Atwoli L, Muhia J, Merali Z. Mental health and climate change in Africa. BJPsych international. 2022:1-4. Available: https://www.cambridge.org/core/journals/bjpsych-international/article/mental-health-and-climate-change-in-africa/65A414598BA1D620F4208A9177EED94B [Accessed 26 Sep 22022].
13 Climate vulnerable economies loss report. Switzerland: Vulnerable twenty group 2020.
14 Ritchie H. WHO has contributed most to global CO2 emissions? our world in data. Available: https://ourworldindata.org/contributed-most-global-co2 [Accessed 12 Sep 2022].
15 Bilotta N, Botti F. *Paving the way for greener central banks. current trends and future developments around the globe*. Rome: Edizioni Nuova Cultura for Istituto Affari Internazionali (IAI), 2022.
16 WHO. *COP26 special report on climate change and health: the health argument for climate action*. Geneva: World Health Organization, 2021.
17 Al-Mandhari A, Al-Yousfi A, Malkawi M, *et al*. "Our planet, our health": saving lives, promoting health and attaining well-being by protecting the planet - the Eastern Mediterranean perspectives. *East Mediterr Health J* 2022;28:247–8.
18 Evans S, Gabbatiss J, McSweeney R, *et al*. COP26: Key outcomes agreed at the UN climate talks in Glasgow. Carbon Brief [Internet], 2021. Available: https://www.carbonbrief.org/cop26-key-outcomes-agreed-at-the-un-climate-talks-in-glasgow/ [Accessed 12 Sep 2022].

BMJ Health & Care Informatics

User-centred design for machine learning in health care: a case study from care management

Martin G Seneviratne ,[1] Ron C Li,[2] Meredith Schreier,[1] Daniel Lopez-Martinez,[1] Birju S Patel,[1] Alex Yakubovich,[1] Jonas B Kemp,[1] Eric Loreaux,[1] Paul Gamble,[1] Kristel El-Khoury,[1] Laura Vardoulakis,[1] Doris Wong,[1] Janjri Desai,[3] Jonathan H Chen,[2] Keith E Morse,[2] N Lance Downing,[2] Lutz T Finger,[1] Ming-Jun Chen,[1] Nigam Shah [2]

To cite: Seneviratne MG, Li RC, Schreier M, *et al*. User-centred design for machine learning in health care: a case study from care management. *BMJ Health Care Inform* 2022;**29**:e100656. doi:10.1136/bmjhci-2022-100656

► Additional supplemental material is published online only. To view, please visit the journal online (http://dx.doi.org/10.1136/bmjhci-2022-100656).

MGS and RCL contributed equally.

Received 27 July 2022
Accepted 23 September 2022

[1]Research, Google Inc, Mountain View, California, USA
[2]Department of Medicine, Stanford University, Stanford, California, USA
[3]Division of Pharmacy, Stanford Medicine, Stanford, California, USA

Correspondence to
Dr Martin G Seneviratne; martsen@google.com

ABSTRACT

Objectives Few machine learning (ML) models are successfully deployed in clinical practice. One of the common pitfalls across the field is inappropriate problem formulation: designing ML to fit the data rather than to address a real-world clinical pain point.

Methods We introduce a practical toolkit for user-centred design consisting of four questions covering: (1) solvable pain points, (2) the unique value of ML (eg, automation and augmentation), (3) the actionability pathway and (4) the model's reward function. This toolkit was implemented in a series of six participatory design workshops with care managers in an academic medical centre.

Results Pain points amenable to ML solutions included outpatient risk stratification and risk factor identification. The endpoint definitions, triggering frequency and evaluation metrics of the proposed risk scoring model were directly influenced by care manager workflows and real-world constraints.

Conclusions Integrating user-centred design early in the ML life cycle is key for configuring models in a clinically actionable way. This toolkit can guide problem selection and influence choices about the technical setup of the ML problem.

INTRODUCTION

Despite the proliferation of machine learning (ML) in healthcare, there remains a considerable implementation gap with relatively few ML solutions deployed in real-world settings.[1] One common pitfall is the tendency to develop models opportunistically—based on availability of data or endpoint labels—rather than through ground-up design principles that identify solvable pain points for target users. There is a long history of clinical decision support tools failing to produce positive clinical outcomes because they do not fit into clinical workflows, cause alert fatigue or trigger other unintended consequences.[2 3] Li *et al* introduced a 'delivery science' framework for ML in healthcare, which is the concept that the successful integration of ML into healthcare delivery requires thinking about ML as an enabling capability of a broader set of technologies and workflows rather than the end product itself.[4] However, it is still unclear how to operationalise this framework, particularly how to select the right healthcare problems where an ML solution is appropriate. As ML becomes increasingly commoditised with advances like AutoML,[5] the real challenge shifts towards identifying and formulating ML problems in a clinically actionable way.

User-centred or human-centred design principles are recognised as an important part of ML development across a range of sectors.[6] Here, we introduce a toolkit for user-centred ML design in healthcare and showcase its application in a case study involving care managers. There were an estimated 3.5 million preventable adult hospital admissions in the USA in 2017, accounting for over US$30 billion in health care spend.[7] Care management aims to assist high-risk patients in navigating care by proactively targeting risk factors via social and medical interventions. In this case study, we provide practical guidance for 'understanding the problem' and 'designing an intervention' (stages in the Li *et al* framework) via user-centred design principles. We draw on cross-domain resources, specifically the Google People+AI Research guidebook[8] and the Stanford d.school design thinking framework (Empathise/Define/Ideate/Prototype/Test),[9] which we adapt for a clinical setting.

METHODS

Figure 1 illustrates the toolkit. First, a problem must satisfy a set of 'hurdle criteria': is it worth solving? Specifically, the problem

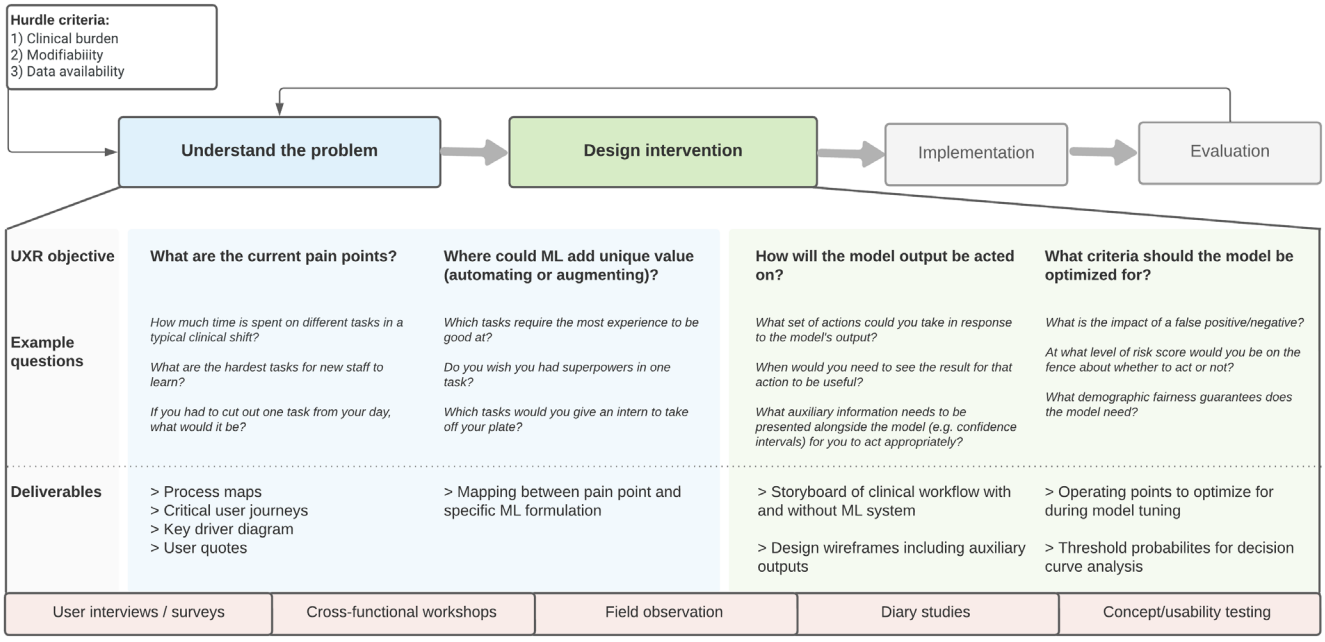

Figure 1 Toolkit for integrating user-centred design into the problem definition stages for ML development in healthcare. ML, machine learning; UXR, user-experience research.

must be associated with significant morbidity or clinical burden, have evidence of modifiability and have adequate data for ML techniques. For candidate problem areas, there are then four key user-centred questions that must be answered:

Q1. Where are the current pain points?
Q2. Where could ML add unique value?
Q3. How will the model output be acted on?
Q4. What criteria should the model be optimised for?

The above toolkit was applied through a series of six user-experience research (UXR) workshops with multidisciplinary stakeholders, including care managers, nurses, population health leaders and physicians affiliated with a managed care programme at Stanford Health Care. Workshops were conducted virtually and were approved by Stanford and Advarra Institutional Review Boards, with consent obtained from all participants.

The schedule of workshops is detailed below:

- Workshop 1 focused on mapping existing workflows. The output was a set of process maps, annotated with pain points.
- Workshop 2 focused on where ML could add unique value (Q2). This yielded a mapping between pain points and possible ML formulations, categorised into automation (replicating repetitive, time-consuming tasks) versus augmentation (adding superhuman functionality).[8]
- Workshops 3 and 4 focused on how a model output would be acted on (Q3). Low-fidelity study probes were developed—storyboards of how an ML tool might fit into a clinical workflow. These were presented to participants for feedback and refined iteratively.
- Workshops 5 and 6 explored ML evaluation metrics for the most promising concept designs. This included how care managers would expect results to be presented and any auxiliary information required alongside the main model output (Q4)

RESULTS

What are the current pain points?

The following pain points were identified:

1. Identifying and prioritising the highest risk patients.
2. Extracting relevant risk factors from the electronic health record.
3. Selecting effective interventions.
4. Evaluating intervention efficacy.

Where could ML add unique value?

Risk stratification (pain point number 1) emerged as an opportunity for 'augmentation' given the challenges in forecasting future deterioration. The ML formulation was a model to predict adverse outpatient events, with emergency department visits and unplanned chronic disease admissions chosen as the prediction endpoints (online supplemental table S1). Identifying risk factors (pain point number 2) was classed as an opportunity for 'automation' given that there is a large volume of unstructured clinical data to sift through. The proposed ML formulation was a natural language processing tool for summarising clinical notes and extracting modifiable risk factors. Selecting interventions and evaluating efficacy (pain points number 3 and 4) were also classed as augmentation opportunities. The ML formulation involved causal inference approaches to estimate individualised treatment effects.

How will the model output be acted on?

Online supplemental figure S1 shows example workflows and storyboards addressing the first two ML formulations above. The actionability pathway for risk scores and personalised risk factor summaries is that care managers can more rapidly prepare for calls and more effectively target their calls to patients with modifiable risk. These risk summaries could be presented to care managers on a monthly basis alongside the existing rule-based lists for high-risk patients. To mimic the existing workflow, the triggering frequency for inference was set as monthly and the inclusion criteria were tailored to fit the managed care population.

What criteria should the model be optimised for?

Since care managers have a limited capacity of patients whom they can contact, precision (positive predictive value) at c (where c is capacity) was selected as the primary evaluation metric. The value of c could be set either as a percentage of the total attributable population (more generalisable across health systems) or as a fixed value (more realistic given care manager staffing does not directly scale with patient load). We also selected realistic baselines to compare the ML models against—namely, rule-based risk stratification heuristics such as selecting recently discharged patients or those with high past utilisation.[10]

DISCUSSION

We applied a practical toolkit for user-centred design, involving four key questions about pain points and ML formulations, via a series of participatory design workshops with care management teams. This guided us towards the pain points of outpatient risk stratification and risk factor identification, with ML formulations involving personalised risk scoring and extraction of potentially modifiable risk factors from the notes. Critical choices about the setup of the ML model were informed by workflow considerations—namely, the endpoint definition, the triggering frequency and the inclusion criteria. Importantly, the evaluation metrics must be tailored to a care management workflow. In this case, there was a capacity constraint on how many patients a care manager can contact each day or week. Hence, the most pragmatic metric was the precision of the model on the top c highest risk patients, rather than global accuracy metrics such as area under the curve of the receiver operating characteristic (ROC-AUC) or precision recall curve (PR-AUC).

This study is limited in only focusing on a single clinical use-case and only using workshops and concept probes as a medium for UXR, given the challenges around direct field observation during the pandemic. Future work will showcase the results of the ML models generated from this UXR collaboration.

CONCLUSION

User-centred design is important for developing ML tools that address a real clinical pain point and dovetail with existing workflows. An iterative approach involving stakeholder interviews and concept feedback can be used to identify pain points, pinpoint where a model could add unique value, understand the actionability pathway and prioritise evaluation metrics.

Twitter Martin G Seneviratne @martin_sen

Acknowledgements Special thanks to the leadership and clinical staff of the Stanford University HealthCare Alliance.

Contributors MGS, RCL, NS devised the project. MS led the UX study sessions with support from MGS, LV, DL-M, DW, KE-K, LTF, JBK, EL, M-JC, PG, RCL, KEM, NLD, JD, JHC. MGS and RCL wrote the initial draft and all other authors reviewed the manuscript.

Funding This study was funded by Google (N/A).

Competing interests This research was funded by Google LLC. LF is now an employee of Marpai Inc. but work for this paper was completed at Google.

Patient consent for publication Not applicable.

Provenance and peer review Not commissioned; externally peer reviewed.

Data availability statement No data are available.

ORCID iDs

Martin G Seneviratne http://orcid.org/0000-0003-0435-3738
Nigam Shah http://orcid.org/0000-0001-9385-7158

REFERENCES

1 Seneviratne MG, Shah NH, Chu L. Bridging the implementation gap of machine learning in healthcare. *BMJ Innovations* 2019.
2 Stone EG. Unintended adverse consequences of a clinical decision support system: two cases. *J Am Med Inform Assoc* 2018;25:564–7.
3 Cohen JP, Cao T, Viviano JD, *et al*. Problems in the deployment of machine-learned models in health care. *Can Med Assoc J* 2021;193:E1391–4.
4 Li RC, Asch SM, Shah NH. Developing a delivery science for artificial intelligence in healthcare. *NPJ Digit Med* 2020;3:107.
5 Waring J, Lindvall C, Umeton R. Automated machine learning: review of the state-of-the-art and opportunities for healthcare. *Artif Intell Med* 2020;104:101822.
6 Gillies M, Fiebrink R, Tanaka A. Human-Centred machine learning. Proceedings of the 2016 CHI conference extended Abstracts on human factors in computing systems. New York, NY, USA Association for Computing Machinery; 2016: 3558–65.
7 McDermott KW, Jiang HJ. Statistical Brief #259: Characteristics and Costs of Potentially Preventable Inpatient Stays, 2017, 2020. Agency for healthcare research and quality (AHRQ). Available: https://www.hcup-us.ahrq.gov/reports/statbriefs/sb259-Potentially-Preventable-Hospitalizations-2017.jsp
8 Google. People + AI research (pair) Guidebook. Available: https://pair.withgoogle.com/guidebook
9 Stanford d.school. Resources: start with design. Available: https://dschool.stanford.edu/resources/get-started-with-design
10 Baker JM, Grant RW, Gopalan A. A systematic review of care management interventions targeting multimorbidity and high care utilization. *BMC Health Serv Res* 2018;18:65.

BMJ Health & Care Informatics

Data consistency in the English Hospital Episodes Statistics database

Flavien Hardy ,[1,2] Johannes Heyl,[1,2] Katie Tucker,[3] Adrian Hopper,[1,4] Maria J Marchã,[5] Tim W R Briggs,[1,6] Jeremy Yates,[5,7] Jamie Day,[1] Andrew Wheeler,[1] Sue Eve-Jones,[1] William K Gray [1]

To cite: Hardy F, Heyl J, Tucker K, *et al*. Data consistency in the English Hospital Episodes Statistics database. *BMJ Health Care Inform* 2022;**29**:e100633. doi:10.1136/bmjhci-2022-100633

► Additional supplemental material is published online only. To view, please visit the journal online (http://dx.doi.org/10.1136/bmjhci-2022-100633).

Received 24 June 2022
Accepted 12 October 2022

[1]Getting It Right First Time, NHS England and NHS Improvement London, London, UK
[2]Department of Physics and Astronomy, University College London, London, UK
[3]Innovation and Intelligent Automation Unit, Royal Free London NHS Foundation Trust, London, UK
[4]Ageing and Health, Guy's and St Thomas' NHS Foundation Trust, London, UK
[5]Science and Technology Facilities Council Distributed Research Utilising Advanced Computing High Performance Computing Facility, London, UK
[6]Royal National Orthopaedic Hospital NHS Trust, Stanmore, UK
[7]Department of Computer Science, University College London, London, UK

Correspondence to
Dr Flavien Hardy;
flavien.hardy.17@ucl.ac.uk

ABSTRACT

Background To gain maximum insight from large administrative healthcare datasets it is important to understand their data quality. Although a gold standard against which to assess criterion validity rarely exists for such datasets, internal consistency can be evaluated. We aimed to identify inconsistencies in the recording of mandatory International Statistical Classification of Diseases and Related Health Problems, tenth revision (ICD-10) codes within the Hospital Episodes Statistics dataset in England.

Methods Three exemplar medical conditions where recording is mandatory once diagnosed were chosen: autism, type II diabetes mellitus and Parkinson's disease dementia. We identified the first occurrence of the condition ICD-10 code for a patient during the period April 2013 to March 2021 and in subsequent hospital spells. We designed and trained random forest classifiers to identify variables strongly associated with recording inconsistencies.

Results For autism, diabetes and Parkinson's disease dementia respectively, 43.7%, 8.6% and 31.2% of subsequent spells had inconsistencies. Coding inconsistencies were highly correlated with non-coding of an underlying condition, a change in hospital trust and greater time between the spell with the first coded diagnosis and the subsequent spell. For patients with diabetes or Parkinson's disease dementia, the code recording for spells without an overnight stay were found to have a higher rate of inconsistencies.

Conclusions Data inconsistencies are relatively common for the three conditions considered. Where these mandatory diagnoses are not recorded in administrative datasets, and where clinical decisions are made based on such data, there is potential for this to impact patient care.

WHAT IS ALREADY KNOWN ON THIS TOPIC

⇒ Large-scale administrative healthcare datasets are increasingly being used to support decision-making, but very little work has been done to assess the quality and consistency of the data.

WHAT THIS STUDY ADDS

⇒ The study offers a novel assessment and analysis of the data quality of the Hospital Episode Statistics dataset in the recording of mandatory diagnoses for patients with autism, type II diabetes mellitus with peripheral complications and Parkinson's disease dementia.

HOW THIS STUDY MIGHT AFFECT RESEARCH, PRACTICE OR POLICY

⇒ Data inconsistencies are relatively common for the conditions considered. Where these mandatory diagnoses are not recorded, there is potential for this to impact on the care provided. This study should motivate the improvement of clinical coding for all conditions with mandatory diagnosis recording.

INTRODUCTION

Decision-making by clinicians and healthcare service managers is increasingly being informed by large-scale administrative healthcare data.[1] Although such data are observational and often lack clinical details, they can support decision-making, particularly in cases where other research methods (eg, randomised controlled trial) may be considered unethical or impractical. Where such data cover an entire population of interest, they can also help minimise the risk of bias due to unrepresentative patient selection criteria (collider bias).[2] However, it is important to have a clear understanding of data quality and the strengths and limitations of any dataset prior to analysis.[3 4]

In England, the Getting It Right First Time (GIRFT) programme is an National Health Service (NHS) England and NHS Improvement initiative with a remit to reduce unwarranted variation in clinical practice that negatively impacts on patient outcomes. The GIRFT programme is one of the largest users of administrative healthcare data for clinical outcome measurement in the UK and has a particular interest in data quality. A key data resource for the GIRFT programme is the Hospital Episodes Statistics (HES) dataset, which contains data for all hospital admissions of NHS patients in England.

The aim of this exploratory study was to identify the extent of, and data features associated with, data inconsistencies within the HES administrative dataset for England.[5]

METHODS

Study design and data collection

This was a retrospective exploratory analysis of HES data. HES data are collected by NHS Digital for all NHS-funded patients admitted to hospitals in England. Hospital trusts run all NHS hospitals in England. A hospital trust is an administrative unit of, typically, one to four hospitals which provides secondary and/or tertiary care for all people living in a geographically defined catchment area. HES includes data for patients funded by the NHS but receiving treatment in a non-NHS hospital. Data collection and reporting is mandatory for NHS funded patients. Data are taken from clinical notes and discharge summarises and data are entered by trained clinical coders at each trust working to a national data standard.[6] Extracts from HES data are audited against clinical audit in a small number of truss each year.

Data regarding pre-existing diagnoses would only be recorded by a coder if detailed in the medical notes or discharge summary, and all clinicians receive training in the importance of accurately recording data. Although autocoding of data is becoming more common in the NHS, its use in the period covered by our study was very limited.

HES data are primarily collected for the purposes of reimbursement. However, their value as a research resource and to inform policy decisions is being increasing recognised.[7]

In HES, a hospital spell is defined as a continuous period in hospital from admission to discharge. A spell can include multiple smaller episodes of care in various hospital settings and under different consultants. As an example, following an emergency department attendance, a patient may initially be under the care of acute medicine (episode one), then transferred to a critical care setting (episode two) and then to a care of the elderly ward (episode three) prior to discharge. Spells involving transfers to other trusts were analysed as separate spells.

Timing, case ascertainment, inclusion and exclusion criteria

Data were taken from HES for all patient discharges during the period 1 April 2013 to 31 March 2021. Using International Statistical Classification of Diseases and Related Health Problems, tenth revision (ICD-10) codes three separate exemplar datasets were extracted for patients with a diagnosis of: childhood autism (F84.0), atypical autism (F84.1) and Asperger's syndrome (F84.5); type II diabetes mellitus with peripheral circulatory complications (DMPC; E11.5) and Parkinson's disease dementia (PDD; F02.3). ICD-10 codes allow data to be captured and defined consistently over time and across settings. There have been no major changes in ICD-10 coding guidance for these conditions over the study period.

DMPC and PDD were selected as representative of patients within the broader disease categories of diabetes mellitus and dementia, respectively.

These conditions were chosen for several reasons:

1. Recording of these conditions is mandated by NHS Digital and NHS England for all subsequent hospital episodes once a diagnosis has been made.[8]
2. The conditions have typical onset in childhood (autism), midlife (DMPC) and late-life (PDD) and so cover a range of demographic groups.
3. All tend to be lifelong once present, accepting that DMPC and PDD are representatives of broader conditions and that the details of the diagnosis may change within these broad definitions.

The first use of the specified code in the diagnostic record for a hospital spell during the study period was identified (index spell) and data for all subsequent spells for the same person extracted.

Spells were removed from the datasets if:

1. The only ICD-10 code present in the record was R69 (unknown and unspecified causes of morbidity) or there was no valid entry in the diagnostic code field.
2. The spell was a regular attendance for renal or liver dialysis (Office of Population Censuses and Surveys Classification of Interventions and Procedures version 4 code X40 or X43; or other regular attendance with ICD-10 code N185 (chronic renal failure) present. Regular day-attendances are usually for a specific procedure, and in most cases only that procedure and the related diagnosis is coded. Inclusion of these spells would unduly bias the dataset.
3. Patients were in age bands where the initial coding diagnoses were most likely miscoded: we removed patients with PDD younger than 40 years, and patients with DMPC younger than 18 years. The data extraction and cleaning procedure for each dataset is summarised in online supplemental figures S1–S3.

Identification of data inconsistencies

All data inconsistencies are reported at the spell level. A subsequent spell was considered consistent with the first spell if at least one of its constituent episodes mentioned the ICD-10 codes listed below for that condition:

Autism: F84.0, F84.1 or F84.5.

DMPC (representing the broader disease category of diabetes mellitus): E10-, E11-, E14-.

PDD (representing the broader disease category of dementia): F00-, F01-, F02-, F03-, F05.1, G30.1, G30.8, G309.

Further details on the definitions of these codes are summarised in online supplemental table S1. In the case of DMPC and PDD, a broader definition of the condition was used for subsequent spells than for the first spell. This was in recognition of the fact that details may not be recorded regarding the diabetes subtype or its presentation or the exact role of Parkinson's disease in the development of dementia.

Covariates and data features/characteristics

Patient characteristics: sex, age in years, ethnicity (white, black or black British, Asian or Asian British, mixed, other and not stated), comorbidities (Charlson Comorbidity

Index,[9] frailty (Hospital Frailty Risk Score (HFRS)[10] and the Global Frailty Score,[11] and deprivation (Index of Multiple Deprivation scores).[12]

Features of hospital stay: Spell length of stay, admission method (emergency or elective), main specialty, number of days since the first spell with the diagnosis recorded (reported as the difference between the discharge date of the first spell and the admission date of the subsequent spell), change of trust between the first and subsequent spell, change of clinical specialty between the first and subsequent spell.

Coding of underlying conditions: We identified spells where a related condition would be expected to also be diagnosed. For PDD this was Parkinson's disease (ICD-10 code G20), and for autism, whether learning disability (ICD-10 codes F70-, F71-, F72-, F73-, F78-, F79-, F80-, F81-, F82- or F83-) was also mentioned in the diagnostic record. The Parkinson's disease code is not mandatory, although the learning disability codes are mandatory.

Outcome (target) variable

For each condition, the target was described by a binary flag indicating whether a code was recorded in the subsequent spell.

Data analysis

Data were extracted onto a secure encrypted server controlled by NHS England and NHS Improvement. Analysis within this secure environment took place using Alteryx 2019.3 (Alteryx, Irvine, California, USA), Python V.3.9.6 and the scikit-learn machine learning library V.1.0.1 (Python Software Foundation, Beaverton, Oregon, USA).[13]

Important predictors associated with data inconsistencies were identified using a random forest classifier algorithm (briefly described in online supplemental figure S4). Missing data values were handled by imputation with the mean or mode in each class. The datasets were separated into a training, validation and test sets with 70%, 15% and 15% of data respectively. Machine learning algorithms require the data to be randomly split so that the algorithm can learn the relationships between the data points and then apply this learning to an unseen part of the data set. The algorithm parameters were determined using the validation set by performing a randomised search on a grid of values and choosing the ones that led to the highest value for the area under the precision recall curve. The classifiers were then trained on the training set and evaluated on the withheld test set. The final parameters of each classifier are summarised in online supplemental table S2.

The models' most important predictors were identified using the SHapley Additive exPlanation (SHAP) feature importance[14] to minimise bias towards high-cardinality variables. Positive or negative correlations of predictors with coding inconsistencies were estimated by calculating the Kendall Tau-b correlation coefficients between the values of the variables, and their estimated Shapley values. These were calculated using TreeSHAP, an efficient estimation approach for tree-based models.[15] Model performance was evaluated using the area under the receiver operating characteristics (AUROC) curve, precision-recall curves and precision gain—recall gain curves.[16] CIs for the areas under the curves were computed using a python implementation of the DeLong method.[17 18]

In subanalyses, we evaluated the impact of time from the first spell on the proportion of inconsistencies. Time from admission for the first spell where the diagnostic code was used to admission for a subsequent spell was calculated in days for the subset of patients where the first spell was prior to 1 April 2018. The follow-up period was set at 3 years for all patients. This was done to avoid a potential bias due to varying maximum follow-up periods for each patient.

RESULTS

Data were available for 172 324 unique patients with autism, 106 943 unique patients with DMPC and 27 794 unique patients with PDD. The characteristics of these patients on their first spell during the study period are summarised in table 1 together with the number of patients without data recorded for each feature. Autism patients had the youngest and patients with PDD the oldest age structure. The autism and DMPC dataset had a high proportion of patients from more deprived areas.

The number of subsequent spells for each patient within a 3-year follow-up period are shown in online supplemental figure S5 for each condition. High numbers of patients (more than 50% for patients with autism) had no subsequent spells within 3 years of their first spell. Patients with DMPC had the highest numbers of subsequent spells. Figure 1 summarises the number of data inconsistencies in these subsequent spells up to 3 years from the first spell where the diagnostic code was used. The number of data inconsistencies increased with time from the first spell, although the trend was less obvious after approximately 20 weeks. Figure 2 illustrate the percentage of subsequent spells with missing mandatory codes in the 3 years after the first spell. The consistency of the coding for PDD appeared to broadly improve over the study period, while for autism patients, consistency appears to have decreased slightly over time.

The number of subsequent spells with data inconsistencies were 170 447 (43.7%) for patients with autism, 46 679 (8.6%) for DMPC patients and 18 975 (31.2%) for patients with PDD. The number of subsequent spells with inconsistencies according to patient characteristics is summarised in table 2. For people with autism, data inconsistencies became more common with greater age. However, for PDD inconsistencies became less common with greater age. Females with autism and PDD had a noticeably higher proportion of inconsistencies than males. There was a modest trend towards a higher proportion of data inconsistencies in autism patients with

Table 1 Table of patient characteristics on first spell within the study period

	Autism	Diabetes mellitus with peripheral complications	Parkinson's disease dementia
No of patients	172324	106943	27794
Age band			
0–17	98591 (57.2 %)	8 (0.01 %)	0 (0.0 %)
18–39	50682 (29.4 %)	1085 (1.0 %)	12 (0.04 %)
40–59	16060 (9.3 %)	21745 (20.3 %)	279 (1.0%)
60–79	6171 (3.6 %)	55050 (51.5 %)	12375 (44.5 %)
80 years and over	820 (0.5 %)	28938 (27.1 %)	15111 (54.4 %)
Not recorded	0	117	17
Sex			
Female	49414 (28.7 %)	32854 (30.7 %)	9828 (35.4 %)
Male	122616 (71.2 %)	74089 (69.3 %)	17961 (64.6 %)
Not recorded	294	0	5
Deprivation quintile			
1 (most deprived)	48539 (29.1 %)	27136 (25.4 %)	4475 (16.1 %)
2	38254 (22.9 %)	23419 (21.9 %)	5248 (18.9 %)
3	31311 (18.28%)	20714 (19.4 %)	5815 (20.9 %)
4	26332 (15.8 %)	17008 (15.9 %)	6084 (21.9 %)
5 (least deprived)	22275 (13.4 %)	13757 (12.9 %)	5790 (20.8 %)
Not recorded	5613	4909	382
Ethnicity			
White	113146 (77.9 %)	89084 (84.8 %)	21402 (93.6 %)
Asian	6916 (4.8 %)	4778 (4.5 %)	730 (3.2 %)
Black	4964 (3.4 %)	3371 (3.2 %)	426 (1.9 %)
Mixed	3695 (2.5 %)	435 (0.4 %)	61 (0.3 %)
Other ethnic groups	16537 (11.4%)	7325 (7.0 %)	240 (1.1 %)
Not recorded	27066	1950	4935
Most common specialties	Paediatrics (23.5 %)	General medicine (33.4 %)	General medicine (33.1%)
	General surgery (7.3 %)	General surgery (31.9 %)	Geriatrics medicine (20.7 %)

Where data are not recorded for deprivation, this is due to the lower super output area of residence not being recorded. In most cases this is due to the patient not having a permanent residence in England (typically they be residents of other parts of the UK). Percentages for each recorded category are calculated excluding any unrecorded data.

increasing deprivation. White patients had the highest rate of inconsistencies for autism.

The variation in data inconsistencies across trusts in England is summarised in online supplemental figure S6. There was substantial spread in terms of data inconsistencies across trusts.

Three random forest classifiers were optimised and trained to identify coding inconsistencies for each

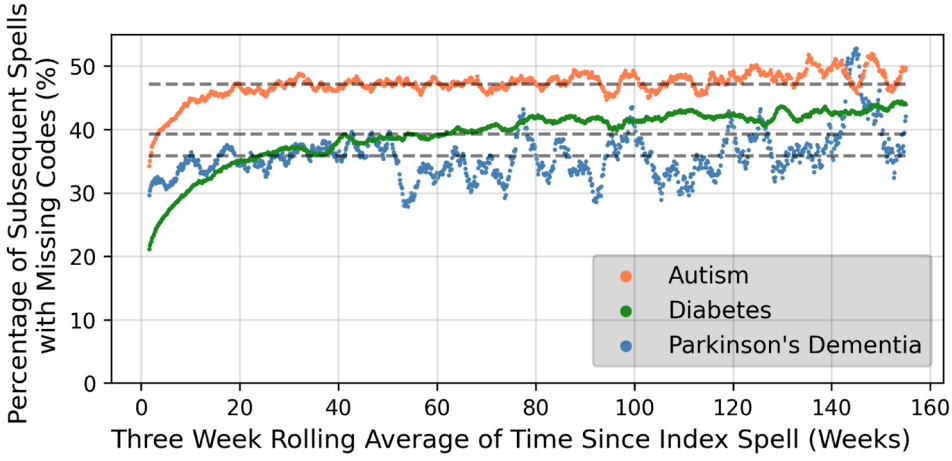

Figure 1 Proportion of subsequent spells with inconsistencies over time up to three years after the index spell

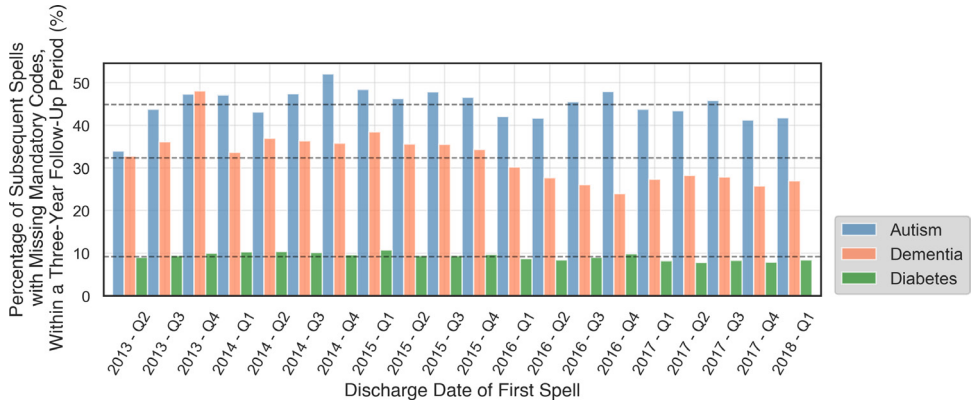

Figure 2 Percentage of spells with missing mandatory codes within 3 years of the first diagnosis, for the discharge date of the first spell ranging from Q2-2013 to Q1-2018.

condition. The relative importance of each feature is shown in figure 3. Across all three conditions, features strongly associated with data inconsistencies included a change in specialty, a change in provider, shorter spell length of stay and female sex. Data inconsistencies were also associated with older patient age for autism and DMPC and younger patients age for PDD. Although deprivation score was an important predictor for all three conditions, the directionality of the relationship was unclear. For patients with PDD, emergency admissions and the absence of the diagnostic code for Parkinson's disease were the most important features. AUROC curve values were 0.80 (95% CI 0.80 to 0.81) for autism, 0.76 (95% CI 0.76 to 0.77) for DMPC and 0.75 (95% CI 0.73 to 0.76) for PDD. online supplemental figure S7 reports the areas under the precision-recall curves and precision gain—recall gain curves, also suggesting the classifiers to have good performance. The performance of each model in Black, Asian, male and female patient subgroups is summarised in Online supplemental table S3o and indicates no significant drop in performance for these groups.

DISCUSSION

We used machine learning algorithms to analyse three large datasets to investigate the consistency of clinical coding of three mandatory health conditions within a large administrative healthcare dataset. Clinical coding of DMPC as a mandatory condition was relatively consistent. However, over two-fifths of subsequent spells for autism patients and almost a quarter of subsequent spells for patients with PDD had data inconsistencies. There was a high level of variation in the proportion of data inconsistencies between trusts, and there was no evidence that trusts are consistently poor at reporting mandatory codes across the three conditions studied.

In the HES dataset, inconsistencies related to mandatory clinical codes can arise from two main sources. A failure of the clinician to record the diagnosis in the medical notes or a failure of the clinical coder to code a diagnosis recorded in medical notes. In our analysis, data inconsistencies could also be due to misuse of the code of interest on the first spell (ie, a false positive in the index spell), although the numbers involved are likely to be small.

From the random forest classifier algorithms, age was strongly associated with data inconsistencies. A greater proportion of data inconsistencies were associated with increasing age for autism and DMPC, and with decreasing age for PDD. This confirms the pattern seen in the descriptive data and is likely to be due to expectations around the likelihood that a patient has the condition. This may also explain the relative importance of the association between female sex and more inconsistencies in the autism dataset. Although we identified a relationship between deprivation score and data inconsistencies in all three datasets, the nature of the relationship was unclear. This may suggest a bias towards continuous variables in the algorithms used.[19 20]

Change in provider, change in main specialty and time from first spell to the subsequent admission were also associated with a higher proportion of data inconsistencies across all datasets. Initiatives to allow easier cross-referencing of information across providers and settings and over an extended period of time should be encouraged.

For the PDD dataset, coding of Parkinson's disease and emergency admission were associated with lower rates of inconsistencies. Elective admissions are generally of short duration and the case notes are likely to focus on the elective procedure being conducted, with limited coding depth.

Large scale, administrative datasets, such as HES, are being increasingly used to inform decision-making in healthcare.[21 22] Such data have helped inform the response to the COVID-19 pandemic[23 24] and are being used to inform service structure postpandemic.[25–27] Having data which is as reliable as possible will be invaluable. Understanding the source and structure of coding inconsistencies may also help the development of new quality improvement programmes, as well as inform the work of researchers, clinical coders and policy analysts.[22 28] The

Table 2 Characteristics of subsequent spells with data inconsistencies

	Autism	Diabetes mellitus with peripheral complications	Parkinson's disease dementia
Total no of spells	583 873	651 458	91 328
No of subsequent spells	390 220	544 341	60 822
No of subsequent spells with missing mandatory codes	170 447 (43.7 %)	46 679 (8.6 %)	18 975 (31.2 %)
Data inconsistencies by overnight stays			
Overnight stay	66 251 (38.9 %)	16 792 (5.0 %)	11 134 (25.2 %)
Day case	104 196 (47.3 %)	29 887 (14.2 %)	7841 (46.8 %)
Data inconsistencies by method of admission			
Elective	84 845 (44.4 %)	19 393 (5.6 %)	6214 (57.8 %)
Emergency	85 380 (43.0 %)	27 247 (13.9 %)	12 752 (25.5 %)
Not recorded	222 (42.4 %)	39 (24.7 %)	9 (32.1 %)
Data inconsistencies by age band			
0–17	57 317 (35.2 %)	0	0
18–39	71 936 (47.3 %)	594 (9.9 %)	0
40–59	27 441 (54.4 %)	9065 (7.3 %)	502 (60.5 %)
60–79	12 025 (53.7 %)	25 409 (8.6 %)	9514 (34.0 %)
80 years and over	1728 (65.2 %)	11 597 (9.7 %)	8955 (28.0 %)
Not recorded	0	14 (6.6 %)	7 (23.5 %)
Data inconsistencies by sex			
Female	64 650 (46.7 %)	14 150 (8.8 %)	6554 (32.9 %)
Male	105 797 (42.0 %)	32 529 (8.5 %)	12 421 (30.4 %)
Not recorded/other	0	0	0
Data inconsistencies by deprivation quintile			
1 (most deprived)	50 739 (45.1 %)	11 298 (7.3 %)	3466 (29.2 %)
2	40 305 (44.8 %)	10 383 (8.1 %)	3622 (30.0 %)
3	30 392 (41.5 %)	9493 (8.8 %)	5000 (36.1 %)
4	26 390 (43.0 %)	8428 (9.9 %)	3587 (30.0 %)
5 (least deprived)	20 478 (41.8 %)	6811 (10.1 %)	3213 (29.9 %)
Not recorded	2143 (50.1 %)	266 (9.7 %)	87 (30.5 %)
Data inconsistencies by ethnicity			
White	114 938 (42.9 %)	39 129 (8.5 %)	14 301 (31.4 %)
Asian	4733 (36.3 %)	1811 (7.0 %)	720 (35.1 %)
Black	4195 (41.3 %)	1301 (7.5 %)	295 (25.5 %)
Mixed	2362 (36.3 %)	385 (14.8 %)	48 (34.0 %)
Other ethnic groups	10 155 (41.8 %)	2968 (9.8 %)	183 (32.6 %)
Not recorded/stated	34 064 (42.9 %)	1086 (12.8 %)	3428 (30.0 %)

Where data are not recorded for deprivation, this is due to the lower super output area of residence not being recorded. In most cases this is due to the patient not having a permanent residence in England (typically they would be residents of other parts of the UK).

impact of the data inconsistencies identified in this paper will vary in importance depending on the nature and aims of the data analysis being undertaken. However, we recommend that researchers using HES and interested in long-term comorbidities should not rely on the coding of the index spell alone, but should look at prior spells for the same patient. Frailty/comorbidity indices, such as the Charlson Comorbidity Index and HFRS, if constructed from HES data, perform this function (to an extent) by looking back over 1 and 2 years of prior hospital spells, respectively.

The performance of the algorithms used to identify key features of data inconsistencies was similar in smaller subgroups of ethnicity and sex. There are concerns that artificial intelligence (AI) techniques can accentuate known biases against representation of smaller subpopulations of a dataset.[29 30] Although the problem of fair data analysis is not unique to AI techniques, and can occur with more traditional forms of data processing and analysis, the 'black-box' element of AI methodology leads naturally to concerns over 'fair AI' and data equity. We used random forest classifiers in our analysis, allowing us

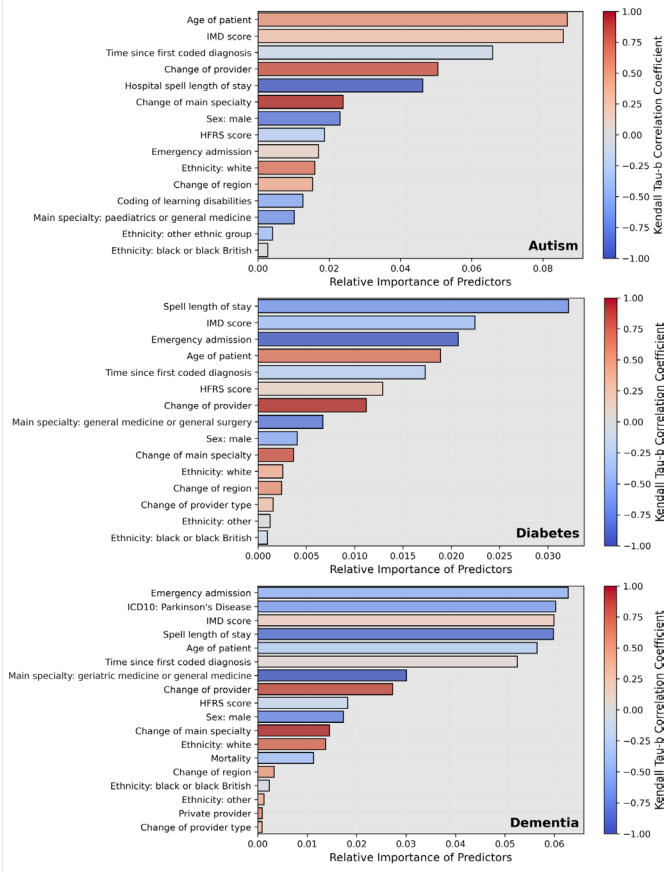

Figure 3 Relative permutation importance of predictors contributing to the identification of coding inconsistencies at the spell level for diagnoses of autism (top), diabetes mellitus with peripheral complications (middle) and Parkinson's disease dementia (bottom). Note: The length of each bar indicates how strongly the classifiers rely on each variable to predict coding consistency at the spell level in the test sets; it is a measure of the relative importance of each predictor. The colour bars indicate the values of the Kendall tau-b correlation coefficient between the values of each variable and the estimated Shapley values. Coefficients close to 1 or -1 correspond to strong positive or negative correlations with coding inconsistencies respectively. HFRS, Hospital Frailty Risk Score; ICD-10, International Statistical Classification of Diseases and Related Health Problems, 10th revision; IMD, Index of Multiple Deprivation.

to understand the key features represented in our algorithms and allowing a degree of transparency.

Our study has a number of strengths and limitations. We had access to one of the most extensive and complete healthcare datasets anywhere in the world. However, this meant that there was no 'gold standard' against which to externally validate the dataset. Difference in coding practice across trusts will have affected our assessment of data quality on the national scale, and we highlight the variation across trusts. We were not able to identify whether an inconsistency was related to a mandatory code being misused in a first spell or being missing in all subsequent spells. We recognise that patients with diabetes mellitus can go into remission, but the number involved across the time period investigated are likely to be very small indeed. We also acknowledge that some forms of dementia and autism may be mild and not impact on the clinical care. Nevertheless, all the conditions studied are mandatory and should still be recorded once diagnosed. Given the potential variability in the source and proportions of coding inconsistencies across all three conditions, the performance of the three classifiers should not be assessed by one single metric alone. For that reason, we opted to also use the precision-recall curves and the recall-aware precision gain—recall gain curves, particularly relevant for the coding of diabetes where the number of inconsistencies is much lower (ie, higher class imbalance). Our analysis highlights that the characteristics of coding inconsistencies can be particular to the condition under investigation. Although we selected conditions that tend to be present across the lifetime, extrapolation to other disease groups should be done with caution. More broadly, although we investigated inconsistent use of mandatory diagnostic codes in this study, it would be possible to investigate other types of inconsistences using similar methods.

CONCLUSIONS

We have identified the extent of, and features associated with, data inconsistencies in the HES database for the three conditions studied, with autism having the highest rate of data inconsistencies. With the likely increased use of administrative data to inform healthcare decision-making, data quality will be of central importance if outcomes for patients are to be optimised. As such, improving data quality should be a priority.

Machine learning techniques, as well as providing insight into the characteristics associated with data inconsistencies, may also be of value in identifying potential data inconsistencies during data input, allowing inconsistencies to be corrected prior to finalisation of the data submission.

Acknowledgements We acknowledge NHS Digital for permission to use their data in this report. The GIRFT programme is providing a framework for examining contemporary clinical practice in unprecedented detail and breadth. We also thank all staff within individual NHS trusts who collected and entered the data used in this study.

Contributors This study was designed and organised by FH, JH, WKG, KT, MM, AH and JY. Data cleaning, and analysis was by FH, supported by JH, MM, KT, AH, AW, SE-J and WKG. Writing of the first draft was by FH and WKG. WKG acts as guarantor

Funding This study was funded by Science and Technology Facilities Council (Grant numbers: Fellowship from DiRAC, ST/S003916/1, ST/W002760/1).

Competing interests None declared.

Patient consent for publication Not applicable.

Ethics approval Consent from individuals involved in this study was not required for this analysis of the HES dataset. Data were pseudonymised for analysis. The analysis and presentation of data follows current NHS Digital guidance for the use of HES data for research purposes. Reported data are anonymised to the level required by ISB1523 Anonymisation Standard for Publishing Health and Social Care Data.

Provenance and peer review Not commissioned; externally peer reviewed.

Data availability statement No data are available. Requests for any underlying data cannot be granted by the authors because the data were acquired from data under licence/data sharing agreement from NHS Digital, for which conditions of use (and further use) apply. Individuals and organisations wishing to access HES data can make a request directly to NHS Digital.

ORCID iDs
Flavien Hardy http://orcid.org/0000-0003-1939-8744
William K Gray http://orcid.org/0000-0002-9597-5446

REFERENCES

1 Agrawal R, Prabakaran S. Big data in digital healthcare: lessons learnt and recommendations for general practice. *Heredity* 2020;124:525–34.
2 Griffith GJ, Morris TT, Tudball MJ, *et al*. Collider bias undermines our understanding of COVID-19 disease risk and severity. *Nat Commun* 2020;11:1–12.
3 Stulberg JJ, Haut ER. Practical guide to surgical data sets: healthcare cost and utilization project national inpatient sample (NIS). *JAMA Surg* 2018;153:586–7.
4 Benchimol EI, Smeeth L, Guttmann A, *et al*. The reporting of studies conducted using observational Routinely-collected health data (record) statement. *PLoS Med* 2015;12:e1001885.
5 Oswald M. *Anonymisation standard for publishing health and social care data specification (process standard*. Leeds, UK: Information Standards Board for Health and Social Care, 2013.
6 NHS Digital. *National clinical coding standards ICD-10Accurate data for quality information*. 5th Edition. Leeds, UK: Terminology and Classifications Delivery Service, 2021.
7 Herbert A, Wijlaars L, Zylbersztejn A, *et al*. Data resource profile: Hospital episode statistics admitted patient care (Hes APC). *Int J Epidemiol* 2017;46:1093–1093i.
8 NHS Digital. *National clinical coding standards ICD-10*. 2021. 5th edn, 2021. https://classbrowser.nhs.uk/ref_books/ICD-10_2021_5th_Ed_NCCS.pdf
9 Sundararajan V, Henderson T, Perry C, *et al*. New ICD-10 version of the Charlson comorbidity index predicted in-hospital mortality. *J Clin Epidemiol* 2004;57:1288–94.
10 Gilbert T, Neuburger J, Kraindler J, *et al*. Development and validation of a hospital frailty risk score focusing on older people in acute care settings using electronic Hospital records: an observational study. *Lancet* 2018;391:1775–82.
11 Soong JTY, Kaubryte J, Liew D, *et al*. Dr foster global frailty score: an international retrospective observational study developing and validating a risk prediction model for hospitalised older persons from administrative data sets. *BMJ Open* 2019;9:e026759.
12 Ministry of Housing and Communities and Local Government. English indices of deprivation, 2019. Available: https://www.gov.uk/government/collections/english-indices-of-deprivation [Accessed 25 Aug 2021].
13 Pedregosa F, Varoquaux G, Gramfort A, *et al*. Scikit-learn: machine learning in python. *Journal of machine Learning research* 2011;12:2825–30.
14 Lundberg SM, Lee S-I. A unified approach to interpreting model predictions. *Proceedings of the 31st international conference on neural information processing systems* 2017;2017:4768–77.
15 Lundberg SM, Erion GG, Lee S-I. Consistent individualized feature Attribution for tree ensembles. *arXiv preprint arXiv* 2018:180203888.
16 Flach P, Kull M. Precision-recall-gain curves: PR analysis done right. *Adv Neural Inf Process Syst* 2015;28.
17 DeLong ER, DeLong DM, Clarke-Pearson DL. Comparing the areas under two or more correlated receiver operating characteristic curves: a nonparametric approach. *Biometrics* 1988;44:837–45.
18 Sun X, Xu W. Fast implementation of DeLong's algorithm for comparing the areas under correlated receiver operating characteristic curves. *IEEE Signal Process Lett* 2014;21:1389–93.
19 Strobl C, Boulesteix A-L, Zeileis A, *et al*. Bias in random forest variable importance measures: illustrations, sources and a solution. *BMC Bioinformatics* 2007;8:1–21.
20 Strobl C, Boulesteix A-L, Kneib T, *et al*. Conditional variable importance for random forests. *BMC Bioinformatics* 2008;9:1–11.
21 Gray WK, Day J, Horrocks M. Editor's Choice - Volume-Outcome Relationships in Elective Abdominal Aortic Aneurysm Surgery: Analysis of the UK Hospital Episodes Statistics Database for the Getting It Right First Time (GIRFT) Programme. *Eur J Vasc Endovasc Surg* 2020;60:509–17.
22 Nouraei SAR, Mace AD, Middleton SE, *et al*. A stratified analysis of the perioperative outcome of 17623 patients undergoing major head and neck cancer surgery in England over 10 years: towards an Informatics-based outcomes surveillance framework. *Clin Otolaryngol* 2017;42:11–28.
23 Navaratnam AV, Gray WK, Day J, *et al*. Patient factors and temporal trends associated with COVID-19 in-hospital mortality in England: an observational study using administrative data. *Lancet Respir Med* 2021;9:397–406.
24 Gray WK, Navaratnam AV, Day J, *et al*. Changes in COVID-19 in-hospital mortality in hospitalised adults in England over the first seven months of the pandemic: an observational study using administrative data. *Lancet Reg Health Eur* 2021;5:100104.
25 Model Health System. Model health system: supporting NHS teams to provide high quality patient care and continuous improvement, 2022. Available: https://model.nhs.uk/ [Accessed 02 Jul 2022].
26 National Clinical Improvement Programme. National clinical improvement programme, 2022. Available: https://www.gettingitrighthfirsttime.co.uk/associated-projects/ncip/ [Accessed 02 Jul 2022].
27 Gray WK, Takhar AS, Navaratnam AV, *et al*. Safety of day-case paediatric tonsillectomy in England: an analysis of administrative data for the getting it right first time programme. *Anaesthesia* 2022;77:277–85.
28 Jewell A, Broadbent M, Hayes RD, *et al*. Impact of matching error on linked mortality outcome in a data linkage of secondary mental health data with Hospital episode statistics (Hes) and mortality records in South East London: a cross-sectional study. *BMJ Open* 2020;10:e035884.
29 Reddy S, Fox J, Purohit MP. Artificial intelligence-enabled healthcare delivery. *J R Soc Med* 2019;112:22–8.
30 Panch T, Mattie H, Atun R. Artificial intelligence and algorithmic bias: implications for health systems. *J Glob Health* 2019;9:010318.

BMJ Health & Care Informatics

Three-year trends in literature on artificial intelligence in ophthalmology and vision sciences: a protocol for bibliometric analysis

Hayley Monson,[1] Jeff Demaine,[2] Laura Banfield,[2] Tina Felfeli [3,4]

To cite: 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. doi:10.1136/bmjhci-2022-100594

► Additional supplemental material is published online only. To view, please visit the journal online (http://dx.doi.org/10.1136/bmjhci-2022-100594).

Received 10 May 2022
Accepted 14 September 2022

[1]Department of Mathematics and Statistics, McMaster University, Hamilton, Ontario, Canada
[2]Library Department, McMaster University, Hamilton, Ontario, Canada
[3]Department of Ophthalmology and Vision Sciences, University of Toronto, Toronto, Ontario, Canada
[4]The Institute of Health Policy, Management and Evaluation, University of Toronto, Toronto, Ontario, Canada

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

ABSTRACT

Introduction The aim of this study is to provide an insight into the literature at the intersection of artificial intelligence and ophthalmology.

Methods and analysis The project will be performed in four key stages: formulation of search terms, literature collection, literature screening and literature analysis. A comprehensive search of databases including Scopus, Web of Science, Dimensions and Cochrane will be conducted. The Distiller SR software will be used for manual screening all relevant articles. The selected articles will be analysed via R Bibliometrix, a program for mathematical analysis of large sets of literature, and VOSviewer, which creates visual representations of connections between articles.

Ethics and dissemination This study did not require research ethics approval given the use of publicly available data and lack of human subjects. The results will be presented at scientific meetings and published in peer-reviewed journals.

WHAT IS ALREADY KNOWN ON THIS TOPIC

⇒ The bibliometric research in ophthalmology, vision research and artificial intelligence is sparse, with many studies looking only at small cross-sections of research or a small volume of papers.

WHAT THIS STUDY ADDS

⇒ This is the first study to use articles across multiple different databases and perform well-established types of analysis to obtain a clear view of the field of vision research and artificial intelligence and its direction.

HOW THIS STUDY MIGHT AFFECT RESEARCH, PRACTICE OR POLICY

⇒ This study will provide a clear view into the present state of ophthalmology and artificial intelligence research and will make predictions about the future of the field. This will allow clinicians to adjust their practices as the field changes and integrate new technologies into their practices as they become available.

INTRODUCTION

Since the term artificial intelligence (AI) was first coined in 1956 by McCarthy and Minsky, its wide-reaching applications to medicine and research have grown in recent years.[1] To date, several studies on the use of AI in ophthalmology have used deep learning technology and machine learning algorithms, which allow for unsupervised programming and training of computer algorithms to make diagnosis of common eye diseases including diabetic retinopathy, macular degeneration, retinopathy of prematurity and glaucoma.[2 3]

Given that the popularity of research in AI and its applications in medicine has grown over recent years, it is important to characterise the field in order to predict future applications of the technology. A bibliometric analysis is a statistical analysis of a large set of research pertaining to a chosen topic. Within ophthalmology, bibliometric analyses have been conducted on the general body of ophthalmological literature and some subspecialties such as glaucoma.[4] Currently, there is no existing bibliometric analysis on the topic of AI in ophthalmology.

The objective of this study is to give a comprehensive view of the impact and importance of AI technology in ophthalmology and vision research through a bibliometric analysis of existing publications in this field from demographic, geographical and topical perspectives. This will allow the medical community to adapt to new technologies and their integration into the future model of patient care.

METHODS

This is a bibliometric analysis of articles relating to AI technology and ophthalmology and vision research. This study will follow the Preferred Reporting Items for Systematic reviews and Meta-Analyses charts reporting guidelines.

Database selection

The aim with database selection was to both capture as much relevant data as possible while also maintaining software compatibility and manageability of the sizes of the datasets. As such, four databases were selected including Web of Science (WoS), Scopus, Dimensions and Cochrane. Note that PubMed, Embase and MEDLINE are subsets of Scopus, so searching Scopus should yield the results from both platforms. Furthermore, the Dimensions database also includes PubMed data. The 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 analytical software including VOSviewer, R Studio and Distiller (https://www.vosviewer.com/).[5–8]

Main outcomes

The main study outcomes will include linkage by coauthorship, co-occurrence, co-citation, citation and bibliographic coupling. In the context of this study, coauthorship networks will offer information about the demographics of the publishing population as well as countries of publication, while co-citation, citation and bibliographic coupling networks will show where collaborations are taking place among authors as well as help to determine which publications had the highest impact; highly cited articles will be counted as more impactful.

Search strategy

A systematic search was conducted on the selected databases from 1 January 2006 until 4 August 2021. To choose a time period, a preliminary curve was graphed using all the results which met the search criteria from the Scopus database (figure 1). A 3-year timeline for the citation analysis was chosen with regard to feasibility of analyses as well as its focused overview of the latest and most relevant technology in AI and ophthalmology.

Keywords have been carefully selected to ensure only relevant documents are analysed. Keywords are separated into two categories, including those relating to AI, and those relating to ophthalmology; these are listed in the table below. The keywords were collected first via combing through of articles deemed highly relevant to the topic, then more were added by referring to ophthalmological and AI vocabulary appendices. Finally, preliminary co-occurrence networks were created with the collected and uncleaned data to determine if any relevant keywords were missing. Table 1 represents the collected keywords, and these will be used to perform the final search. Relevant keywords will also be searched both in their British spellings and American spellings and searched in both capitalised and lowercase forms. Only English articles will be selected for as co-occurrence analysis relies on the measurement of the frequency of keywords. All words in the paper's bodies must be in one language for this analysis to be successful.

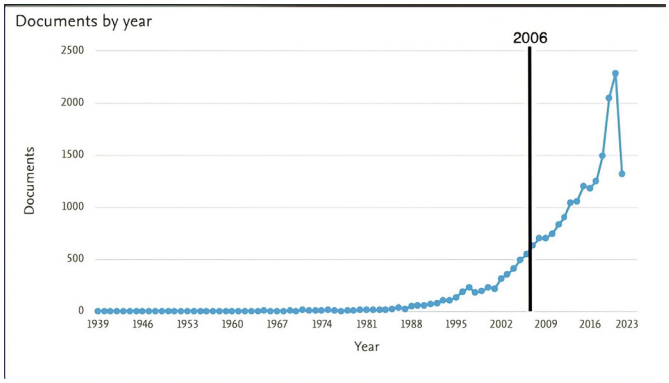

Figure 1 Graph illustration of all the peer-reviewed article hits on utilisation of artificial intelligence and ophthalmology meeting the search inclusion and exclusion criteria from the Scopus database.

Software used

The databases will be searched using the above outlined criteria. The first stage of the search will include those articles which are compatible with the VOSviewer software, these being articles from WoS, Scopus and Dimensions. Duplicates and articles deemed irrelevant will be removed using the Distiller software. These will then be imported into the VOSviewer software and analysis will be performed as outlined in the Methods section: first on each individual dataset and then on the data from all three compatible databases. The second stage will involve downloading articles from all four chosen databases. Duplicates and irrelevant articles will once again be removed using the Distiller software and then R studio software will be used for data analysis.

Data analysis

Networks linking articles will be created based on the following characteristics: countries of publication, author, co-citation and bibliographic linkage. A comparison will be drawn between trends in general ophthalmology research and AI-focused ophthalmology research and investigation conducted into the implications of these statistics as well as determination of the extent of scientific impact from each group. All literature from WoS, Dimensions and Scopus will be amalgamated into one super-network which is less specific, and then networks for each of these databases will be created individually and analysed on a more specific level.

Given that the VOSviewer software does not support the Cochrane database, all documents will be analysed with respect to a number of mathematical informatics models including Bradford's Law which predicts that only a few journals will account for a large proportion of literature in a field[9 10]; Lotka's Law, which predicts an inverse square correlation between the number of authors publishing and the number of articles published, specifically, the number of authors publishing N papers is proportional to the inverse square of that number of papers[11 12]; and Price's Law, which predicts that the growth of productivity in an area of scientific research can be fitted to an exponential curve, levelling off asymptotically after a

Table 1 Summary of keywords and search terms used in systematic search of the selected databases

Ophthalmology	Artificial intelligence
General terms: ▶ Ophthalmology ▶ Ocular ▶ Eye ▶ Intraocular ▶ Iridology ▶ Visual field Anatomical terms: ▶ Retina ▶ Macula ▶ Fovea ▶ Uvea ▶ Sclera ▶ Cornea ▶ Conjunctiva ▶ Iris ▶ Vitreous body ▶ Vitreous humor ▶ Vitreous fluid ▶ Vitreo ▶ Aqueous humor ▶ Retinal ganglion cells ▶ Fundus oculi Imaging terms: ▶ Optical coherence tomography ▶ OCT ▶ Color fundus photography ▶ CFP ▶ Slit lamp ▶ Confocal microscopy ▶ Confocal scanning microscopy ▶ Confocal laser scanning microscopy ▶ Ultrasound biomicroscopy ▶ Fundus fluorescein angiography ▶ Indocyanine green angiography ▶ Scanning laser ophthalmoscopy ▶ Ocular ultrasonography ▶ Microperimetry ▶ Multifocal visual-evoked potentials ▶ Perimetry ▶ Retinal functional imaging ▶ Retinal vessel segmentation ▶ Iris recognition ▶ Visual field tests Disease terms: ▶ Diabetic retinopathy ▶ Retinopathy ▶ Retinopathy of prematurity ▶ Macular degeneration ▶ Retinal vein occlusion ▶ Cataracts ▶ Glaucoma ▶ Retinoblastoma ▶ Uveitis ▶ Iritis ▶ Choroiditis ▶ Retinitis ▶ Chorioretinitis ▶ Conjunctivitis ▶ Endophthalmitis ▶ Optic neuropathy ▶ Optic atrophy ▶ Diabetic macular edema ▶ Mellitus ▶ Myopia ▶ Visual disorder ▶ Vision disorder Procedure terms: ▶ Vitrectomy ▶ Phacoemulsification ▶ Paracentesis ▶ Trabeculectomy ▶ Canaloplasty ▶ Laser iridotomy ▶ Baerveldt valve ▶ Iridotomy ▶ Iridectomy ▶ Goniotomy ▶ Scleral buckle ▶ Pneumatic retinopexy ▶ Phacoemulsification ▶ Extracapsular ▶ Photocoagulation ▶ Selective laser trabeculoplasty ▶ Canthotomy ▶ Brachytherapy ▶ Catholysis ▶ Closure of cyclodialysis cleft ▶ Corneal transplantation ▶ Decompression of dacryocele ▶ Decompression of orbit ▶ Pars plana lensectomy ▶ Retrobulbar injection ▶ Strabismus surgery ▶ Synechiolysis ▶ Tarsorrhaphy ▶ Transscleral cyclophotocoagulation	▶ Artificial intelligence ▶ Deep learning ▶ Deep learning system ▶ Convolutional neural network ▶ Massive training artificial neural network ▶ Neural network ▶ Machine learning ▶ Image processing ▶ Long short term memory ▶ Supervised clustering ▶ Unsupervised learning ▶ Semi-supervised learning ▶ Backpropagation ▶ Feed forward ▶ Feature learning ▶ Decision tree ▶ Transfer learning ▶ Big data ▶ Natural language processing ▶ Computer vision ▶ Image recognition ▶ Semantic analysis ▶ Unsupervised learning ▶ Cognitive computing ▶ Entity annotation ▶ Entity extraction ▶ Machine intelligence ▶ Predictive analysis ▶ k-nearest neighbour ▶ Lattice neural network ▶ Random forest ▶ Feature extraction ▶ Neural nets ▶ Feature fusion ▶ Deep belief fusion ▶ Image segmentation ▶ Computer-aided detection ▶ Optic cup segmentation ▶ Data mining

period of time.[13 14] For this data analysis, the R Bibliometrix package will be used. Comparison of ratios between these numbers with the expected informetric models will further elucidate anomalies in the data and contribute to the objective of developing an understanding of the impact and trajectory of research in AI technology and ophthalmology.

DISCUSSION

We anticipate that the field of AI in ophthalmology has grown at an exponential rate over the past 3 years per Price's Law. Furthermore, we predict that most of the identified articles will be related to diagnostics rather than to direct patient care technology, such as surgical robots. Diagnostic algorithms are more realistically and immediately applicable to patient care; they are low cost and easy to create and implement. Surgical robots are costly, require more professional skill to develop and have narrower applications in ophthalmology.

It is anticipated that the bulk of the literature will be produced by more populated countries such as the USA and China, though extensive collaboration between these countries is not predicted because of their geographical locations. Collaboration between neighbouring countries, such as Canada and the USA, is more likely. Furthermore, we predict that publication volume will drop in 2020 with some doctors diverting their research to the SARS-CoV-2 virus.

Due to the specificity of the field, the bulk of the research will be found in a few non-specific journals, with fewer and fewer articles being found in increasingly specific journals. This would align with the Bradford zones outlined in the analysis. Inverse correlation between the topicality of the journal and the number of articles is predicted given that the field is narrow and still emerging.

Limitations

The authors would like to acknowledge the limitations of this bibliometric study. First, only English articles will be selected for in order to produce the most effective analysis, and this may limit the scope of the search. Second, only three of four of the selected databases are supported by the VOSviewer software and as such network analyses can only be performed on documents from these. The availability of information is also largely dependent on database indexing; PubMed documents will not export accompanying citation information and so only co-occurrence and coauthorship networks can be made with these data. In order to address and overcome these limitations, meta-networks will be created with all the data from Scopus, WoS and Dimensions. Then, each dataset will be analysed individually using all available techniques in order to glean more detailed information. All data will be analysed with the above outlined informetric models using the R Bibliometrix package.

Twitter Tina Felfeli @TinaFelfeli

Contributors Conception and design—TF. Acquisition of data—HM, LB and TF. Data analysis—HM and JD. Interpretation of data—HM, JD and TF. First draft of

the article—HM, JD and TF. Critical revision—HM, JD and TF. Final approval of the version to be published—HM, JD, LB and TF. 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. Felfeli.

Competing interests None declared.

Patient consent for publication Not required.

Ethics approval The Institutional Review Board of the University of Ontario deemed that ethical approval was not necessary, as this was a pure bibliometric study.

Provenance and peer review Not commissioned; externally peer reviewed.

ORCID iD

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. https://sitn.hms.harvard.edu/flash/2017/history-artificial-intelligence/
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* 2020;34:451–60.
4 López-Muñoz F, Weinreb RN, Moghimi S, *et al*. A bibliometric and mapping analysis of glaucoma research between 1900 and 2019. *Ophthalmol Glaucoma* 2022;5:16–25.
5 Web of Science: Summary of Coverage - Web of Science platform - LibGuides at Clarivate Analytics [Internet]. Available: https://clarivate.libguides.com/webofscienceplatform/coverage [Accessed 28 Jul 2021].
6 Elsevier. *Content coverage guide*. Elsevier, 2010: 1–24. https://www.elsevier.com/__data/assets/pdf_file/0017/114533/Scopus_GlobalResearch_Factsheet2019_FINAL_WEB.pdf
7 Cochrane Library. About the Cochrane Database of Systematic Reviews [Internet]. Available: https://www.cochranelibrary.com/cdsr/about-cdsr [Accessed 28 Jul 2021].
8 Dimensions. A Guide to the Dimensions Data Approach [Internet]. Available: https://www.dimensions.ai/resources/a-guide-to-the-dimensions-data-approach/ [Accessed 10 Aug 2021].
9 BS C. Sources of information on specific subjects 1934. *Engineering* 1985;10:85–6 https://cir.nii.ac.jp/crid/1570854175272181632
10 Garfield E. Bradford's Law and Related Statistical Patterns. *Curr Contents* 1980;19:5–12.
11 Lotka AJ. The frequency distribution of scientific productivity. *J Washingt Acad Sci* 1926;16:317–23.
12 Kawamura M, Thomas CD, Tsurumoto A, *et al*. Lotka's law and productivity index of authors in a scientific Journal. *J Oral Sci* 2000;42:75–8.
13 Price DJ. Quantitative measures of the development of science. *Arch Int dHistoire des Sci* 1951;4:85–93.
14 Fernández-Cano A, Torralbo M, Vallejo M. Reconsidering Price's model of scientific growth: An overview. *Scientometrics* 2004;61:301–21.