An interactive dashboard to track themes, development maturity, and global equity in clinical artificial intelligence research

Interest in the application of artificial intelligence (AI) to human health continues to grow, but widespread translation of academic research into deployable AI devices has proven more elusive.[1,2] There is increasing recognition of limitations in how AI research is carried out, from methods of model validation that do not emulate real-world conditions,[3] to characteristics of data[4] and inadequate inclusion of researchers and populations from diverse global regions.[5] Systematic reviews of clinical AI criticise widespread risk of bias and lack of downstream clinical utility, and research waste is an increasing concern.[6,7]

One problem is the lack of a unifying perspective over the colossal-sized landscape of global AI research. Continual quantification of research characteristics can enable identification and monitoring of shortcomings in this heterogeneous landscape. However, the sheer quantity of published research (>150 000 papers on MEDLINE under broad terms; appendix p 1) makes this a substantial challenge. Literature database searches have poor specificity, and cannot directly identify original research in model development, or pinpoint research representing advanced stages of model validation. Literature reviews only describe a portion of research at a single timepoint, are laborious to conduct and reproduce, and are quickly outdated in a rapidly changing landscape.

In response to these requirements, we produced an end-to-end Natural Language Processing (NLP) pipeline that performs real-time identification, classification, and characterisation of AI research abstracts extracted from MEDLINE, outputting results to an interactive dashboard, creating a live view of global AI development. We identified four primary aims: first, to directly discover original research in clinical AI model development; second, to identify research at more advanced development stages using mature evaluation methodology, ie comparative evaluation of AI algorithms versus a reference standard[8] or prospective real-world testing (appendix p 13); third, to map, in real-time, global distribution and equity in AI research production on a per-author basis; and fourth, to track the main active research themes across clinical specialties, diseases, algorithms, and data types.

Development was done using Python (version 3.8) and Tensorflow (version 2.5). To achieve the required performance, we employed transfer learning, using state-of-the-art Bi-directional Encoder Representations from Transformers NLP models with pre-training on medical corpuses.[9] Models were fine-tuned on manually labelled abstracts indexed on MEDLINE before 2020 and tested prospectively on abstracts indexed after pipeline completion. The final pipeline and methods described are available in the appendix (pp 1–5, 13). In a prospective evaluation, the classifier for research discovery achieves an F1 score of 0·96 and Matthews correlation coefficient (MCC) of 0·94. The classifier for maturity achieves an F1 of 0·91 and MCC of 0·90. The multi-class classifier for labelling themes achieves a macro-average F1 of 0·97. When evaluated against publications discovered by recent systematic reviews, the pipeline correctly classified 98% for inclusion and maturity. Performance metrics are reported in the appendix (pp 8–12).

See **Online** for appendix

The dashboard allows all discovered research to be visualised by development maturity, medical specialty, data type, algorithm, research location, publication date, or different combinations of attributes. Datasets containing labelled abstracts and metadata are refreshed every 24 h and made available to download, as an aid to literature reviewers, or for reproducible analysis of research progress across any cross-section of characteristics.

For the **interactive dashboard** see https://aiforhealth.app

Using dashboard datasets, we illustrate heterogeneity in research maturity across major specialties and diseases over the past decade using a horizon chart (appendix p 17).[10] Respiratory medicine, breast cancer, and retinopathy demonstrate greatest production of mature research relative to total research production. Distribution of data type usage across major subspecialties are shown as heatmaps (appendix p 14), showing increased prevalence of mature validation methodology using radiomics (and other computer vision tasks) across all specialties. Notably, only 1·3% of research, and 0·6% of mature research, involved an author from a low to low-middle income country (as per World Bank definitions), with 93·6% of such research published after 2016 (appendix p 15). Live visualisations are found on the dashboard website.

While demonstrating state-of-the-art NLP performance, classifier limitations include imperfect accuracy compared with careful human reviewers (the trade-off against time required for manual characterisation). We use only MEDLINE due to their unique application programming interface. Finally, prediction using full articles could increase performance, but this was hindered by a paywalled access to most publications.

The interactive dashboard was published in November, 2021. Given its popularity and utility to date, we plan to continue enhancement of this resource. We consider immediate downstream use-cases to be analysis of drivers for AI maturity and translation, reviewing features of mature AI research, and ongoing characterisation of AI development in developing countries. Codes and data are made public, with the hope that functionality can be expanded in collaboration with the global AI community.

For the **codes and data** see https://github.com/whizzlab

This publication did not receive any direct funding. Views expressed are authors' own. JZ receives funding from the Wellcome Trust (203928/Z/16/Z) and acknowledges support from the National Institute for Health Research Biomedical Research Centre based at Imperial College NHS Trust and Imperial College London. SB receives funding from the Wellcome Trust (566701). LAC receives funding from the National Institute of Health (NIBIB R01 EB017205). We would like to thank the team at SparkNLP for academic licensing rights to use their named entity recognition engine. All data and codes are available for the public through the open access Github repository. Datasets are available to download through the online dashboard.

*Joe Zhang, Stephen Whebell, Jack Gallifant, Sanjay Budhdeo, Heather Mattie, Piyawat Lertvittayakumjorn, Maria del Pilar Arias Lopez, Beatrice J Tiangco, Judy W Gichoya, Hutan Ashrafian, Leo A Celi, James T Teo
joe.zhang@imperial.ac.uk

Institute of Global Health Innovation (JZ, HA) and Department of Computing (PL), Imperial College London, London, UK; Department of Critical Care (JZ) and Department of Neurology (JTT), King's College Hospital NHS Foundation Trust, London, UK; Department of Critical Care, Townsville University Hospital, Queensland Health, Townsville, QLD, Australia (SW); Department of Surgery, Imperial College Healthcare NHS Foundation Trust, London, UK (JG); Centre for Human and Applied Physiological Sciences, King's College London, London, UK (JG); Department of Neurology, National Hospital for Neurology and Neurosurgery, London, UK (SB); Department of Clinical and Movement Neurosciences, University College London, London, UK (SB); Department of Biostatistics, Harvard T H Chan School of Public Health, Harvard University, Cambridge, MA, USA (HM, LAC); SATI-Q Program, Argentine Society of Intensive Care, Buenos Aires, Argentina (MPAL); National Institute of Health, College of Medicine, University of the Philippines, Metro Manila, Philippines (BJT); Division of Medicine, The Medical City, Pasig City, Philippines (BJT); Department of Radiology, Emory University School of Medicine, Atlanta, Georgia, USA (JWG); Preemptive Medicine and Health Security Initiative, Flagship Pioneering, Cambridge, MA, USA (HA); Institute for Medical Engineering & Science, Massachusetts Institute of Technology, Cambridge, MA, USA (LAC); Department of Medicine, Beth Israel Deaconess Medical Centre, Boston, MA, USA (LAC); London Medical Imaging & AI Centre, Guy's and St Thomas' Hospital, London, UK (JTT)

[1] Muehlematter UJ, Daniore P, Vokinger KN. Approval of artificial intelligence and machine learning-based medical devices in the USA and Europe (2015–20): a comparative analysis. *Lancet Digit Health* 2021; **3:** e195–203.

[2] Lyell D, Coiera E, Chen J, Shah P, Magrabi F. How machine learning is embedded to support clinician decision making: an analysis of FDA-approved medical devices. *BMJ Health Care Inform* 2021; **28:** e100301.

[3] Panch T, Mattie H, Celi LA. The "inconvenient truth" about AI in healthcare. *NPJ Digit Med* 2019; **2:** 77.

[4] Ibrahim H, Liu X, Zariffa N, Morris AD, Denniston AK. Health data poverty: an assailable barrier to equitable digital health care. *Lancet Digit Health* 2021; **3:** e260–65.

[5] Wawira Gichoya J, McCoy LG, Celi LA, Ghassemi M. Equity in essence: a call for operationalising fairness in machine learning for healthcare. *BMJ Health Care Inform* 2021; **28:** e100289.

[6] Navarro CLA, Damen JAA, Takada T, et al. Risk of bias in studies on prediction models developed using supervised machine learning techniques: systematic review. *BMJ* 2021; **375:** n2281.

[7] Wilkinson J, Arnold KF, Murray EJ, et al. Time to reality check the promises of machine learning-powered precision medicine. *Lancet Digit Health* 2020; **2:** e677–80.

[8] Chen PC, Mermel CH, Liu Y. Evaluation of artificial intelligence on a reference standard based on subjective interpretation. *Lancet Digit Health* 2021; **3:** e693–95.

[9] Gu Y, Tinn R, Cheng H, et al. Domain-specific language model pretraining for biomedical natural language processing. *ACM Trans Comput Healthcare* 2022; **3:** 1–23.

[10] Heer J, Kong N, Agrawala M. Sizing the horizon: the effects of chart size and layering on the graphical perception of time series visualizations. In: Olsen DR, Arthur RB, eds. Proceedings of the SIGCHI conference on human factors in computing systems. Boston, MA: Association for Computing Machinery, 2009: 1303–12.