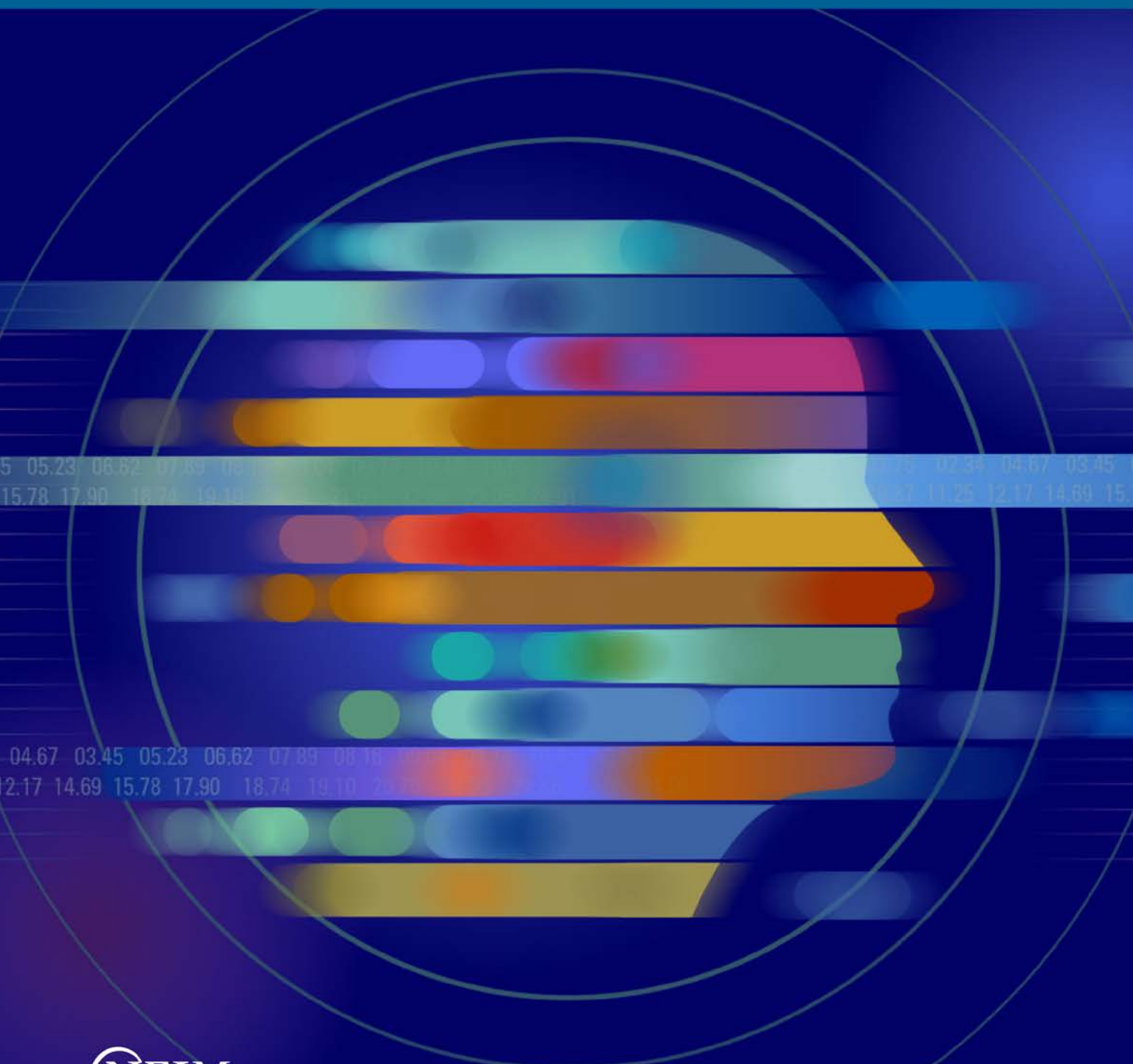
NEJM AI

Vol. 1, No. 1 | January 2024



NEJM GROUP

NEJM AI

NEJM AI Editors

NEJM AI

TABLE OF CONTENTS

Volume 1 | No. 1 | January 2024

NEJM AI is a publication of NEJM Group, a division of the Massachusetts Medical Society.

NEJM AI

NEJM AI 2023; 1 (1)
DOI: 10.1056/AIe2300197

EDITORIAL

Injecting Artificial Intelligence into Medicine

Isaac S. Kohane, M.D., Ph.D.[1]

Received: October 17, 2023; Accepted: October 19, 2023; Published: December 11, 2023

Abstract

Introducing a new journal from NEJM Group: *NEJM AI*, a platform aimed at informing readers and guiding the responsible development of artificial intelligence (AI) to enhance the quality of health care. This editorial reflects on AI's historical milestones and current capabilities, particularly large language models, and the imperative for their rigorous clinical evaluation. It discusses the crucial need for AI in medicine to undergo the same level of scrutiny as any clinical intervention, predominantly through randomized controlled trials, despite the challenges posed by the technology's complexity. Looking forward, *NEJM AI* commits to fostering a multidisciplinary discourse and the development of a transparent, patient-centered approach to AI in health care, with an emphasis on the critical importance of diverse and accessible datasets. The editorial concludes by envisioning a future where *NEJM AI* not only informs but actively shapes the ethical integration of AI into a health care system that respects patient autonomy and upholds the highest standards of care.

Welcome to a collaborative journey as we launch *NEJM AI*. Whether you are a reader or author, we hope that you share our primary goal: augmenting the capabilities of clinicians, patients, and their larger community using the latest entrant to our ecosystem — AI — to deliver safe and effective health care to the highest of our collective standards. We approach this goal with optimism yet bear a constant awareness that misuse or careless implementation of these technologies will precipitate systemic harms to all parties in the health care system, most importantly, patients.

The spotlight on AI has been intensifying since its breakthrough in image recognition in 2012,[1] achieving performance rivaling human capability across various medical disciplines. By 2020, as revealed by a PubMed query, clinical evaluations of AI and machine learning had transformed from mere topics into a burgeoning field, that year alone boasting over 300 articles appraising these technologies. Of course, the aspiration to augment clinicians with AI, whether for safety or to overcome our cognitive limitations, is hardly new and

The author affiliation is listed at the end of the article.

Dr. Kohane can be contacted at Isaac_Kohane@hms.harvard.edu or at Harvard Medical School, Department of Biomedical Informatics, 10 Shattuck Street, Boston, MA 02115.

Back to Table of Contents

dates back to at least the 1950s.[2] In 1970, in the pages of the *New England Journal of Medicine*, the noted endocrinologist William Schwarz wrote that "computing science will probably exert its major effects by augmenting and, in some cases, largely replacing the intellectual functions of the physician …[and will influence] in a fundamental fashion the problems of both physician manpower and quality of medical care, it will also inevitably exact important social costs — psychologic, organizational, legal, economic and technical."[3] Fifty years later, the development of large online datasets, including medical texts and health records, the soaring performance of graphical processing units, and advances in neural network architectures gave ample reason for optimism that Schwarz's forecast might be realized in this century.[4]

Enter large language models (LLMs), which catapulted onto the stage and into the lives of patients and doctors in the fall of 2022, fueled by a social media frenzy with stories spanning from the transformative to the cautionary.[5] A retrospective glance to 1812, when the *New England Journal of Medicine and Surgery* made its debut, reminds us that randomized trials were then an unimaginable 130 years away. Yet, 208 years later, in 2020, of the aforementioned clinical AI evaluations published, less than 1% were prospective, randomized controlled trials. This makes clear our urgent mandate: In addition to needed technical advances, AI must meet the same bar for clinical evidence that is expected from other clinical interventions. For a given AI tool to be used, evidence that it will perform in a safe and effective manner must be demonstrated, preferably using randomized controlled trials designed to test the tool against an established standard.

Randomized controlled trials with LLMs will not be easy. The breadth of these programs' capabilities and unknowns about what data they have already "seen" makes their evaluation on narrowly defined tasks somewhat artificial and not entirely reflective of their usage by clinicians or patients. Necessarily, ensuring that pluripotent AI programs are "clinical grade" and safe will require a multifaceted, collaborative approach among clinicians, patient advocacy groups, and a myriad of stakeholders. Transparency will be the principal tool required to win the trust of our readers, regardless of the form this evaluation takes.

Evaluation is but one dimension of AI's future in health care. Questions loom large: Who will steer the innovation ship and toward which horizon? Our various article types, from Perspectives to Policy Corner, will serve as a vibrant forum, dissecting these questions and more, with a commitment to spotlighting not just successes but also informative failures and cautionary tales. Moreover, in the world of medical AI, well-annotated, representative, diverse, and freely available datasets arguably stand among the most precious resources, enabling investigators to assess the mettle of AI programs in realistic, representative tasks. Thus, articles unveiling new datasets, benchmarks, and innovative, reproducible protocols will not merely be welcomed but celebrated (see ai.nejm.org for a full list of article types).

So, who do we envision contributing to the pages of *NEJM AI*? Clinicians? AI researchers? In our eyes, the most impactful articles will blossom from the fertile ground of multidisciplinary teams, reflecting the vibrance at the intersection of computer science, clinician–patient dynamics, and biomedical research. Our editorial board, a meld of pioneers from the realms of AI, life sciences, ethics, and policy, is genuinely thrilled to engage with our community to find how best to navigate these complex, uncharted waters with precision, curiosity, and integrity.

Fast-forward a decade, and should you inquire about *NEJM AI*'s journey and its measure of success, many responses might serve. However, the paramount answer will hinge on this essential query: Did we help guide our health care system through the maelstrom of rapidly evolving technology, ensuring that patients are afforded an unprecedented standard of health care, all while safeguarding the autonomy and dignity they (i.e., we) inherently deserve? Although second in importance, answering the following query affirmatively is dear to all of us: Did we effectively advocate for the design and implementation of AI so that the practice of medicine is as emotionally and intellectually rewarding as some earnest medical school applications seem to anticipate?

NEJM AI's activities will extend considerably beyond the publication of a journal. We already host a popular podcast, *NEJM AI Grand Rounds*, a newsletter, and two to three seminars per year in a hybrid in-person/webcast format. We will organize international meetings with a specific thematic focus as the need arises. For example, we organized the Symposium for Responsible AI for Social and Ethical Healthcare (RAISE) that met at the end of October, and the white paper will be shared on our website and those of collaborating journals. Although the advent of AI in medicine is momentous, we are only at the very

Back to Table of Contents

beginning. I invite those of you who share our enthusiasm to engage with us in these activities as well as within the pages of our journal.

Disclosures

Author disclosures are available at ai.nejm.org.

Author Affiliation

[1] Harvard Medical School, Department of Biomedical Informatics, Boston

References

1. Krizhevsky A, Sutskever I, Hinton GE. ImageNet classification with deep convolutional neural networks. NeurIPS Proceedings. 2012 (https://papers.nips.cc/paper_files/paper/2012/hash/c399862d3b9d6b76c8436e924a68c45b-Abstract.html).
2. Haward LRC. The robot anaesthetist; an introduction to the automatic control of anaesthesia by means of an electro-encephalographic intermediary. Med World 1952;76:624-626.
3. Schwartz WB. Medicine and the computer. The promise and problems of change. N Engl J Med 1970;283:1257-1264. DOI: 10.1056/NEJM197012032832305.
4. Rajkomar A, Dean J, Kohane I. Machine learning in medicine. N Engl J Med 2019;380:1347–58. DOI: 10.1056/NEJMra1814259.
5. Lee P, Bubeck S, Petro J. Benefits, limits, and risks of GPT-4 as an AI chatbot for medicine. N Engl J Med 2023;388:1233-1239. DOI: 10.1056/NEJMsr2214184.

Back to Table of Contents

NEJM AI

NEJM AI 2023; 1 (1)
DOI: 10.1056/AIe2300128

EDITORIAL

Why We Support and Encourage the Use of Large Language Models in *NEJM AI* Submissions

Daphne Koller, Ph.D.,[1] Andrew Beam, Ph.D.,[2,3,4] Arjun Manrai, Ph.D.,[3,4] Euan Ashley, M.B., Ch.B., D.Phil.,[5] Xiaoxuan Liu, M.B.Ch.B., Ph.D.,[4,6,7] Judy Gichoya, M.B.Ch.B., M.S.,[8] Chris Holmes, Ph.D.,[9,10] James Zou, Ph.D.,[11] Noa Dagan, M.D., Ph.D., M.P.H.,[12,13] Tien Y. Wong, M.D., Ph.D.,[14,15] David Blumenthal, M.D., M.P.P.,[16] Isaac Kohane, M.D., Ph.D.,[3,4] on behalf of the editors and editorial board of *NEJM AI**

Received: September 12, 2023; Accepted: October 17, 2023; Published: December 11, 2023

Abstract

Large language models (LLMs) promise to revolutionize many aspects of the creation and dissemination of scientific knowledge; however, their use in scientific writing remains controversial, because of concerns about authorship, originality, factual inaccuracies, and "hallucinations" or confabulations. As a result, several publication venues have explicitly prohibited their use. At *NEJM AI*, we have elected instead to allow the use of LLMs for submissions, as long as authors take complete responsibility for the content and properly acknowledge the use of LLMs. However, this policy does not allow an LLM to be listed as a coauthor. We believe that the use of LLM tools can help scientists enhance the quality of their scientific work and democratize both the creation and consumption of scientific knowledge, thereby helping us maximally enable the scientific workforce to produce robust, novel scientific findings and disseminate them broadly.

Large language models (LLMs) have recently emerged as a powerful tool across many areas of biomedicine. They are able to rapidly summarize large amounts of text, generate high-quality text from a short description, create code that can help support data analysis, produce images on the basis of a verbal description, and much more. On the surface, it appears plausible that an LLM can generate an entire scientific paper, which can then be submitted, as is, for publication. This possibility complicates proper attribution of a text's authorship, and it raises the specter of a potential flood of low-quality scientific work that was not originated or overseen by a human but nonetheless was submitted for peer-reviewed publication. As a consequence, some publication venues have elected to prohibit the use of LLMs in submissions. Most notably, at the time of writing, *Science* has stated a policy whereby "[t]ext generated from AI, machine learning, or similar

*A complete list of the editors and editorial board of NEJM AI is available at ai.nejm.org.

The author affiliations are listed at the end of the article.

Dr. Koller can be contacted at koller@gmail.com or insitro, 279 East Grand Ave., Suite 200, South San Francisco, CA 94080.

Back to Table of Contents

algorithmic tools cannot be used in papers published in *Science* journals, nor can the accompanying figures, images, or graphics be the products of such tools, without explicit permission from the editors. In addition, an AI program cannot be an author of a *Science* journal paper."

Historically, there have been multiple occasions where people have resisted the use of "overly powerful" tools in various contexts. For years, calculators were banned in schools (and in some cases, still are) because of a perceived imperative to have students do their own calculations. In the early days of information technology, even word processors were banned in some organizations because they might reduce typing skills. There are clearly some contexts — such as education, where students are learning and being evaluated on foundational skills — where access to LLMs might be counterproductive. However, our primary goal at *NEJM AI* is to increase the quality of scientific publications, which includes several aspects such as novelty, rigor, and accessibility to others. If powerful tools, such as LLMs, help us achieve those goals, we should welcome their use.

Moreover, we are, as of yet, far from a world in which an LLM can generate an original and correct piece of scientific research — human involvement is still paramount. To use an LLM effectively, one needs to suggest the core premise of the work, identify the most relevant resources, often generate new data that did not exist, explore different analysis approaches, distill conclusions, and engage in multiple iterations before new and interesting scientific output is created.

At *NEJM AI*, we have therefore elected to allow the use of LLMs. Our two key conditions are first, that the use of LLMs is appropriately acknowledged by the authors. This standard is the same as for any tool or resource that is used in a substantive way by authors in their scientific work, including experimental reagents, animal models, data sets, software systems, or third-party copyediting services. Second, we require that the authors be completely accountable for the correctness and originality of the submitted work. Likewise, the same quality standards for clarity, exposition, and strength of the scientific arguments will be applied to all papers submitted to *NEJM AI*, regardless of how the text was generated. Using an LLM does not absolve one of the responsibility to write well and to avoid plagiarism. Above all, the insights in any paper we consider must be original, novel, and clearly articulated. It is important to note, however, that this policy does not allow an LLM to be listed as a coauthor on any submission. This is because LLMs cannot be held accountable for the content of their work. To properly disclose the use of an LLM, authors should include a statement describing how the LLM was used in the acknowledgments section of the submission. For more details on the proper ways to disclose the use of an LLM, please see our guidance to authors.

We believe that the potential benefits of LLMs in scientific writing are considerable. LLMs can help scientists contextualize their work, democratize knowledge, enhance data analysis, and produce better scientific output. They can also help non-English native speakers and those with language disabilities express their ideas more effectively. As such, the use of LLMs could help reduce the language barrier for scientists around the world and improve the quality of the scientific literature. This is increasingly important as science becomes more globalized and insights are produced by a diverse range of individuals from different cultural, linguistic, and educational backgrounds.

Moreover, it is important to recognize that a ban on the use of LLMs is likely not enforceable. Although some tools for recognizing LLM-generated text have been developed, they are, as of yet, too inaccurate to be reliable. Several efforts to detect the use of an LLM in the classroom have resulted in instances of students being falsely accused of using an LLM when it was prohibited. As such, a prohibition on the use of LLMs would disadvantage those authors who are law-abiding by preventing them from benefiting from all the advantages outlined above.

The existence of LLMs (whether approved or not) does pose risks. As is the case for many technologies — ranging from nuclear power and computers to stem cell research and genetic engineering to cryptography — LLMs may enable and accelerate behaviors both good and bad. We live in a time of rapid progress for AI tools, and as such, editorial policies must be sufficiently agile. We will continue to reevaluate the proper use of AI tools in all parts of the scientific process to update this policy, as the landscape will almost surely change dramatically over the coming years. At *NEJM AI*, we believe that our fundamental goal in scientific discovery and publication should be to maximally enable humankind to produce robust, novel scientific findings and disseminate them broadly. The better the tools that we provide to scientists, the greater their ability to do so.

Back to Table of Contents

Disclosures

Author disclosures are available at ai.nejm.org.

The endorsement list is as follows.

Daphne Koller, Ph.D., insitro, South San Francisco, CA.

Andrew Beam, Ph.D., Department of Epidemiology, Harvard T.H. Chan School of Public Health, Boston; Department of Biomedical Informatics, Harvard Medical School, Boston; and *NEJM AI*, Boston.

Arjun K. Manrai, Ph.D., Department of Biomedical Informatics, Harvard Medical School, Boston; and *NEJM AI*, Boston.

Ziad Obermeyer, M.D., University of California, Berkeley, Berkeley, CA.

Euan Ashley, M.B., Ch.B., D.Phil., Stanford, CA.

Marinka Zitnik, Ph.D., Department of Biomedical Informatics, Harvard Medical School, Boston.

Jianfei Zhao, Ph.D., Jiahui Medical Research and Education Group, Shanghai, China; and *NEJM AI*, Boston.

Isaac Kohane, M.D., Ph.D., Department of Biomedical Informatics, Harvard Medical School, Boston; and *NEJM AI*, Boston.

Pranav Rajpurkar, Ph.D., Department of Biomedical Informatics, Harvard Medical School, Boston.

Chris Holmes, Ph.D., Department of Statistics and Nuffield Department of Medicine, University of Oxford, Oxford, United Kingdom.

Carey Goldberg, independent journalist formerly at *Bloomberg News*, *WBUR*, *The New York Times*, and *The Boston Globe*.

James Zou, Ph.D., Department of Biomedical Data Science, Stanford University, Stanford, CA.

Xiaoxuan Liu, M.B.Ch.B., Ph.D., University Hospitals Birmingham NHS Foundation Trust, Birmingham, United Kingdom; College of Medical and Dental Sciences, University of Birmingham, Birmingham, United Kingdom; and *NEJM AI*, Boston.

Noa Dagan, M.D., Ph.D., M.P.H., Clalit Research Institute, Innovation Division, Clalit Health Services, Tel Aviv, Israel and Department of Software and Information Systems Engineering, Ben Gurion University, Be'er Sheva, Israel.

Judy Gichoya, M.B.Ch.B., M.S., Department of Radiology & Imaging Sciences, Emory University School of Medicine, Atlanta.

Tien Y. Wong, M.D., Ph.D., Tsinghua Medicine, Tsinghua University, Beijing; and Singapore Eye Research Institute, Singapore National Eye Center, Singapore.

David Ouyang, M.D., Department of Cardiology, Cedars–Sinai Medical Center, Los Angeles; and *NEJM AI*, Boston.

Lily Peng, M.D., Ph.D., Verily Life Sciences, South San Francisco, CA; and *NEJM AI*, Boston.

Charlotte Haug, M.D., Ph.D., SINTEF Digital Health, Oslo, Norway; Stanford Health Policy, Stanford University, Palo Alto, CA; and *NEJM AI*, Boston.

Author Affiliations

[1] insitro, South San Francisco, CA
[2] Department of Epidemiology, Harvard T.H. Chan School of Public Health, Boston
[3] Department of Biomedical Informatics, Harvard Medical School, Boston
[4] *NEJM AI*, Boston
[5] Department of Medicine, Stanford University, Stanford, CA
[6] University Hospitals Birmingham NHS Foundation Trust, Birmingham, United Kingdom
[7] College of Medical and Dental Sciences, University of Birmingham, Birmingham, United Kingdom
[8] Department of Radiology and Imaging Sciences, Emory University School of Medicine, Atlanta
[9] Department of Statistics, University of Oxford, Oxford, United Kingdom
[10] Nuffield Department of Medicine, University of Oxford, Oxford, United Kingdom
[11] Department of Biomedical Data Science, Stanford University, Stanford, CA
[12] Clalit Research Institute, Innovation Division, Clalit Health Services, Tel Aviv, Israel
[13] Department of Software and Information Systems Engineering, Ben Gurion University, Be'er Sheva, Israel
[14] Tsinghua Medicine, Tsinghua University, Beijing
[15] Singapore Eye Research Institute, Singapore National Eye Center, Singapore
[16] Department of Health Policy and Management, Harvard T.H. Chan School of Public Health, Boston

Back to Table of Contents

NEJM AI

NEJM AI 2023; 1 (1)
DOI: 10.1056/AIp2300031

PERSPECTIVE

Use of GPT-4 to Diagnose Complex Clinical Cases

Alexander V. Eriksen, M.D.,[1,2] Sören Möller, M.Sc., Ph.D.,[3,4] and Jesper Ryg, M.D., Ph.D.[1,2]

Received: July 10, 2023; Revised: September 15, 2023; Accepted: September 29, 2023; Published: November 9, 2023

Abstract

We assessed the performance of the newly released AI GPT-4 in diagnosing complex medical case challenges and compared the success rate to that of medical-journal readers. GPT-4 correctly diagnosed 57% of cases, outperforming 99.98% of simulated human readers generated from online answers. We highlight the potential for AI to be a powerful supportive tool for diagnosis; however, further improvements, validation, and addressing of ethical considerations are needed before clinical implementation. (No funding was obtained for this study.)

Introduction

The combination of a shortage of physicians and the increased complexity in the medical field partly due to the rapidly expanding diagnostic possibilities already constitutes a significant challenge for the timely and accurate delivery of diagnoses. Given demographic changes, with an aging population this workload challenge is expected to increase even further in the years to come, highlighting the need for new technological development. AI has existed for decades and previously showed promising results within single modal fields of medicine, such as medical imaging.[1] The continuous development of AI, including the large language model (LLM) known as the Generative Pretrained Transformer (GPT), has enabled research in exciting new areas, such as the generation of discharge summaries[2] and patient clinical letters. Recently, a paper exploring the potentials of GPT-4 showed that it was able to answer questions in the U.S. Medical Licensing Examination correctly.[3] However, how well it performs on real-life clinical cases is less well understood. For example, it remains unclear to what extent GPT-4 can aid in clinical cases that contain long, complicated, and varied patient descriptions and how it performs on these complex real-world cases compared with humans.

We assessed the performance of GPT-4 in real-life medical cases by comparing its performance with that of medical-journal readers. Our study utilized available complex clinical case challenges with comprehensive full-text information published online between January 2017 and January 2023.[4] Each case presents a medical history and a poll with six options for the most likely diagnosis. To solve the case challenges, we provided GPT-4

The author affiliations are listed at the end of the article.

Dr. Eriksen can be contacted at alexander.viktor.eriksen@rsyd.dk or at University of Southern Denmark Faculty of Health Sciences, Department of Clinical Research, Geriatric Research Unit, Kløvervænget 10, Odense, Syddanmark, Denmark 5000.

Back to Table of Contents

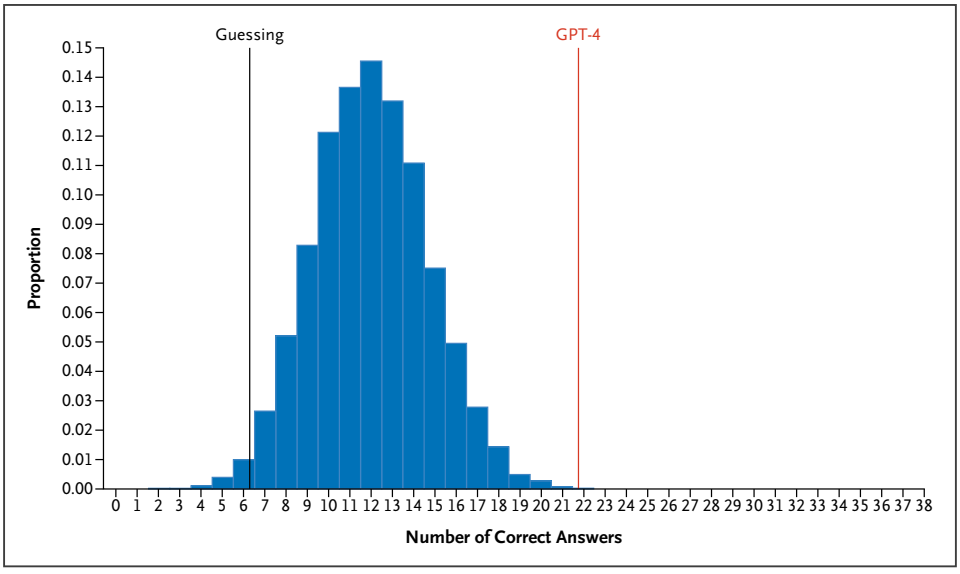


Figure 1. Number of Correct Answers of GPT-4 Compared with Guessing and a Simulated Population of Medical-Journal Readers.

Number of correct answers of GPT-4 (red line) to 38 multiple-choice real-world clinical case challenges compared with what would be expected by purely guessing with uniform probability for all answer possibilities (black line) and to the proportion of correct answers by a simulated population of 10,000 medical-journal readers (blue histogram).

with a prompt and a clinical case (see Supplementary Methods 1 in the Supplementary Appendix). The prompt instructed GPT-4 to solve the case by answering a multiple-choice question followed by the full unedited text from the clinical case report. Laboratory information contained in tables was converted to plain text and included in the case. The version of GPT-4 available to us could not accept images as input, so we added the unedited image description given in the clinical cases to the case text. The March 2023 edition of GPT-4 (maximum determinism: temp=0) was provided each case five times to assess reproducibility across repeated runs. This was also performed using the current (September 2023) edition of GPT-4 to test the behavior of GPT-4 over time. Because the applied cases were published online from 2017 to 2023 and GPT-4's training data include online material until September 2021, we furthermore performed a temporal analysis to assess the performance in cases before and after potentially available training data. For medical-journal readers, we collected the number and distribution of votes for each case. Using these observations, we simulated 10,000 sets of answers to all cases, resulting in a pseudopopulation of 10,000 generic human participants. The answers were simulated as independent Bernoulli-distributed variables (correct/incorrect answer) with marginal distributions as observed among medical-journal readers (see Supplementary Methods 2).

We identified 38 clinical case challenges and a total of 248,614 answers from online medical-journal readers.[4] The most common diagnoses among the case challenges were in the field of infectious disease, with 15 cases (39.5%), followed by 5 cases (13.1%) in endocrinology and 4 cases (10.5%) in rheumatology. Patients represented in the clinical cases ranged in age from newborn to 89 years old (median [interquartile range], 34 [18 to 57]), and 37% were female. The number of correct diagnoses among the 38 cases occurring by chance would be expected to be 6.3 (16.7%) due to the six poll options. The March 2023 edition of GPT-4 correctly diagnosed a mean of 21.8 cases (57%) with good reproducibility (55.3%, 57.9%, 57.9%, 57.9%, and 57.9%), whereas the medical-journal readers on average correctly diagnosed 13.7 cases (36%) (see Supplementary Table 1 and Supplementary Methods 1). GPT-4 correctly diagnosed 15.8 cases (52.7%) of those published up to September 2021 and 6 cases (75.0%) of those published after September 2021. Based on the simulation, we found that GPT-4 performed better than 99.98% of the pseudopopulation (Fig. 1). The September 2023 edition of GPT-4 correctly diagnosed 20.4 cases (54%).

Limitations

An important study limitation is the use of a poorly characterized population of human journal readers with unknown levels of medical skills. Moreover, we cannot assess whether the responses provided for the clinical cases reflect their maximum effort. Consequently, our results may represent a best-case scenario in favor of GPT-4. The assumption of independent answers on the 38 cases in our pseudopopulation is somewhat unrealistic, because some readers might consistently perform differently from others and the frequency at which participants respond correctly to the cases might depend on the level of medical skills as well as the distribution of these. However, even in the extreme case of maximally correlated correct answers among the medical-journal readers, GPT-4 would still perform better than 72% of human readers.

Conclusions

In this pilot assessment, we compared the diagnostic accuracy of GPT-4 in complex challenge cases to that of journal readers who answered the same questions on the Internet. GPT-4 performed surprisingly well in solving the complex case challenges and even better than the medical-journal readers. GPT-4 had a high reproducibility, and our temporal analysis suggests that the accuracy we observed is not due to these cases' appearing in the model's training data. However, performance did appear to change between different versions of GPT-4, with the newest version performing slightly worse. Although it demonstrated promising results in our study, GPT-4 missed almost every second diagnosis. Furthermore, answer options do not exist outside case challenges. However, a recently published letter reported research that tested the performance of GPT-4 on a closely related data set, demonstrating diagnostic abilities even without multiple-choice options.[5]

Currently, GPT-4 is not specifically designed for medical tasks. However, it is expected that progress on AI models will continue to accelerate, leading to faster diagnoses and better outcomes, which could improve outcomes and efficiency in many areas of health care.[1] Whereas efforts are in progress to develop such models, our results, together with recent findings by other researchers,[5] indicate that the current GPT-4 model may hold clinical promise today. However, proper clinical trials are needed to ensure that this technology is safe and effective for clinical use. Additionally, whereas GPT-4 in our study worked only on written records, future AI tools that are more specialized are expected to include other data sources, including medical imaging and structured numerical measurements, in their predictions. Importantly, future models should include training data from developing countries to ensure a broad, global benefit of this technology and reduce the potential for health care disparities. AI based on LLMs might be relevant not only for in-patient hospital settings but also for first-line screening that is performed either in general practice or by patients themselves. As we move toward this future, the ethical implications surrounding the lack of transparency by commercial models such as GPT-4 also need to be addressed,[1] as well as regulatory issues on data protection and privacy. Finally, clinical studies evaluating accuracy, safety, and validity should precede future implementation. Once these issues have been addressed and AI improves, society is expected to increasingly rely on AI as a tool to support the decision-making process with human oversight, rather than as a replacement for physicians.[1,3]

Disclosures

Author disclosures and other supplementary materials are available at ai.nejm.org.

No funding was obtained for the present study.

Author Affiliations

[1] Geriatric Research Unit, Department of Clinical Research, University of Southern Denmark, Odense, Denmark
[2] Department of Geriatric Medicine, Odense University Hospital, Odense, Denmark
[3] Open Patient data Explorative Network, OPEN, Odense University Hospital, Odense, Denmark
[4] Department of Clinical Research, University of Southern Denmark, Odense, Denmark

References

1. Haug CJ, Drazen JM. Artificial intelligence and machine learning in clinical medicine, 2023. N Engl J Med 2023;388:1201-1208. DOI: 10.1056/NEJMra2302038.
2. Patel SB, Lam K. ChatGPT: the future of discharge summaries? Lancet Digit Health 2023;5:e107-e108. DOI: 10.1016/S2589-7500(23)00021-3.
3. Lee P, Bubeck S, Petro J. Benefits, limits, and risks of GPT-4 as an AI chatbot for medicine. N Engl J Med 2023;388:1233-1239. DOI: 10.1056/NEJMsr2214184.
4. The New England Journal of Medicine. Case challenges (https://www.nejm.org/case-challenges).
5. Kanjee Z, Crowe B, Rodman A. Accuracy of a generative artificial intelligence model in a complex diagnostic challenge. JAMA 2023;330:78-80. DOI: https://doi.org/10.1001/jama.2023.8288.

Back to Table of Contents

NEJM AI

NEJM AI 2023; 1 (1)
DOI: 10.1056/AIp2300189

PERSPECTIVE

Patient Portal

Carey Goldberg [1]

Received: October 13, 2023; Accepted: October 17, 2023; Published: December 11, 2023

Abstract

A patient fantasizes about a letter on generative AI from her provider.

Introduction

Time to get your flu vaccine. There is an updated Covid-19 vaccine, too. And you may have heard about a shift in hospital partnerships for cancer care — don't worry, it won't affect current treatment.

These are all helpful notes I have received recently from the Boston-centered medical behemoth that provides my care. Lately, I have been fantasizing about one more message I would like to get. Something like this:

Dear Patient:

We would like you to know about a powerful new tool that could help you achieve better health. It's called generative artificial intelligence (AI). Chances are you've heard about ChatGPT, Bard, and other examples in recent months, whether hype about a discovery as important as fire or doom about the potential end of humanity.

We are writing with more practical, immediate news: you can already use this new type of AI in your health care yourself. It is as easy as going online. And we are committed to getting as much benefit out of it for you as possible. There are a few things you need to know first, however.

Most importantly, you cannot trust it to be factually correct. Repeat: Do not trust it. This type of AI makes things up when it does not know an answer. Never ever act on what it tells you without checking.

So why are we even bothering to write to you about it? Because despite that glaring flaw — which researchers are racing to solve — the new AI tools can help you be a better patient and us to practice better medicine. We and other health care systems are working on all kinds of ways to use it; for now, however, the following sections offer some uses you may want to explore for yourself.

The author affiliation is listed at the end of the article.

Ms. Goldberg can be contacted at careyg@comcast.net.

Back to Table of Contents

Explaining and Summarizing

After you see your doctor, a long note describing the visit is generated, and it may be hard to plow through and understand fully. You can ask an AI program to summarize it for you and keep asking it questions until all is clear.

Cancer survivor Dave deBronkart — also known as "e-Patient Dave" for his years of activism to empower patients — tried that recently. He pasted in a visit note of more than 2000 words and asked ChatGPT to "produce a summary of my current problems, and a list of action items." After a bit of back and forth, he got a concise list of seven issues, including post–Covid-19 symptoms and high cholesterol levels, and what to do about them.

Similarly, you can use the new AI tools to explain the health insurance "Explanation of Benefits" that, let's face it, no one really understands. You can paste that form's text into an AI program and ask it to summarize the treatment you received in lay language. Or in a language other than English.

You should know, though, that if you paste medical information into an online AI program, it loses our privacy protections.

Thinking through Medical Issues

Hugo Campos, a colleague of deBronkart's in patient advocacy, says that because of the potential for chatbots to provide inaccurate answers, he would not rely on them entirely for answers to his questions. Rather, he told deBronkart, he uses them to augment his thinking and decision-making.

For example, Campos asked AI to analyze the pros and cons of using one blood pressure medication rather than another in a detailed analysis that a doctor likely did not have time for. Another friend of deBronkart's who wanted to try a wearable blood pressure monitor used AI to marshal the arguments to persuade his doctor that it made sense — and succeeded.

deBronkart encourages patients to learn how to use generative AI just as we learned how to use Google searches. "The practice of medicine is a function, among other things, of how much the professionals know and how much the patients bring to the table," he says. For all their flaws, the stunning capabilities of the new AI tools let you bring much more to the table.

As a patient, you have a legal right to all your medical data; now you have a better tool than ever before to help you make sense of it.

Tips and Tricks

All of this is very new, and we are all still learning. This new AI has only been out in the world since late 2022, and medicine usually takes many years to put new discoveries to widespread use. But we know that, just as our patients started using Google, you will start using AI, and we aim to help you use it as well as possible.

You may want to use AI for many other purposes, including help with possible diagnoses and support when facing daunting medical issues. One of the biggest surprises of the new AI is that when you interact with it, it can come across much like a deeply compassionate doctor — one with unlimited time to talk and also able to answer any question, at length, within seconds.

We are attaching a tip sheet, including advice on how to "prompt" the AI program to give you what you want. [Note to *NEJM AI* readers: This tip sheet is completely imaginary at this point — but not for long!] Sometimes, whether for obvious reasons such as safety concerns or for totally mysterious reasons, the AI tool does not fulfill requests. It can be so tricky to get what you want from it that there is a whole new field labeled "prompt engineering." A leading thinker on AI, Eliezer Yudkowsky, recently told an AI program to play the role of a smart, thoughtful medical student before asking it why he was having uncomfortable clicks in his shoulder.

"In today's strange new world," Yudkowsky wrote, "it's vital to know how to instruct imaginary doctors to give you straight answers."

Back to Table of Contents

That is a skill we will all be learning now, and we hope that all of us — patients and providers — will be learning it together as we bring a new kind of intelligence into our care. We will do our best to share what we learn that could be of use to you.

Sincerely,
Your hospital

[Comments? Feedback? What you would most like to say to patients?]

Health/science journalist Carey Goldberg co-authored the 2023 book The AI Revolution in Medicine: GPT-4 and Beyond. *She covers public and patient perspectives on generative AI and can be contacted at* careyg@comcast.net.

Disclosures

Author disclosures are available at ai.nejm.org.

Author Affiliation

[1] MIT, Cambridge, MA

NEJM AI

NEJM AI 2023; 1 (1)
DOI: 10.1056/AIoa2300030

ORIGINAL ARTICLE

Characterizing the Clinical Adoption of Medical AI Devices through U.S. Insurance Claims

Kevin Wu, M.S.,[1] Eric Wu, M.S.,[2] Brandon Theodorou,[3] Weixin Liang, M.S.,[4] Christina Mack, Ph.D.,[5] Lucas Glass, Ph.D.,[5] Jimeng Sun, Ph.D.,[3,6] and James Zou, Ph.D.[1,2,4]

Received: July 9, 2023; Revised: September 15, 2023; Accepted: September 29, 2023; Published: November 9, 2023

Abstract

There are now over 500 medical artificial intelligence (AI) devices that are approved by the U.S. Food and Drug Administration. However, little is known about where and how often these devices are actually used after regulatory approval. In this article, we systematically quantify the adoption and usage of medical AI devices in the United States by tracking Current Procedural Terminology (CPT) codes explicitly created for medical AI. CPT codes are widely used for documenting billing and payment for medical procedures, providing a measure of device utilization across different clinical settings. We examined a comprehensive nationwide claims database of 11 billion CPT claims between January 1, 2018, and June 1, 2023 to analyze the prevalence of medical AI devices based on submitted claims. Our results indicate that medical AI device adoption is still nascent, with most usage driven by a handful of leading devices. For example, only AI devices used for assessing coronary artery disease and for diagnosing diabetic retinopathy have accumulated more than 10,000 CPT claims. Furthermore, we found that zip codes that had a higher income level, were metropolitan, and had academic medical centers were much more likely to have medical AI usage. Our study sheds light on the current landscape of medical AI device adoption and usage in the United States, underscoring the need to further investigate barriers and incentives to promote equitable access and broader integration of AI technologies in health care.

Introduction

As artificial intelligence (AI) has rapidly progressed in recent years, significant investments have been devoted to developing and commercializing AI in medicine. As of 2023, over 500 medical AI devices have undergone U.S. Food and Drug Administration (FDA) evaluation and received approval across areas such as radiology, neurology, and pathology.[1] During an FDA submission, device manufacturers are required to report evidence of the efficacy and safety of their products, providing crucial

The author affiliations are listed at the end of the article.

Dr. Zou can be contacted at jamesyzou@gmail.com or at 350 Jane Stanford Way, Room 369, Stanford, CA 94305.

Read Article at ai.nejm.org

Back to Table of Contents

insight into how AI algorithms are evaluated before being used on patients.[2] However, after approval, companies rarely share where and when their products are used. As such, despite the proliferation of medical AI approvals, little is known about their real-world usage.

The usage and adoption patterns of medical AI devices can significantly affect their clinical impact. First, the performances of AI algorithms are notoriously susceptible to changes in health care settings and fluctuate during deployment.[3,4] For instance, despite initial studies indicating up to a 20% improvement in detection rates, computer-aided detection (CAD) products for mammography approved in the early 2000s have been found to provide no tangible benefits to women.[5] This discrepancy has been attributed to adoption and usage factors such as changes in clinician interaction with the software and the transition from film to digital mammograms.[6] Consequently, although AI medical devices may demonstrate strong performance under specific evaluation conditions, variations in real-world applications can yield drastically different outcomes. Second, the impact of medical AI devices is mediated by economic forces. After FDA approval, companies need to find sustainable revenue streams for the promises of AI-driven health care to be realized. Different reimbursement approaches can affect how often and on whom these devices are used, and it is still unclear which model is optimal for the new AI devices.[7,8] Studying the empirical usage of medical AI devices is a crucial step in characterizing the landscape of medical innovations and can provide a more holistic view of the translational pipeline from algorithm to patient.

Recently, Current Procedural Terminology (CPT) codes have been created specifically for medical AI devices.[7,8] CPT codes are designated by the U.S. Department of Health and Human Services under the Health Insurance Portability and Accountability Act as a national coding set for physicians and other health care professional services and procedures to be used by the Centers for Medicare & Medicaid Services (CMS).[9] The codes are regularly created, updated, and modified by the American Medical Association (AMA) and are the most widely accepted medical nomenclature under public and private health insurance programs.[9] Health care providers use these codes to generate itemized bills detailing the specific services delivered to a patient during a medical encounter. Subsequently, these bills are submitted to insurance companies, who use the coded information to determine the appropriate reimbursement for the services rendered. As such, CPT codes play a crucial role in ensuring the accuracy and uniformity of medical billing, as well as promoting accountability and transparency within the health care system.

CMS also provides coverage for medical AI devices through a new technology add-on payment (NTAP), which is specifically designed to encourage health care providers to adopt new technologies.[7] However, the NTAP program specifically focuses on inpatient payments, whereas CPT codes apply to both inpatient and outpatient settings.[10,11] In this article, we focus on CPT codes because they are most widely adopted and standardized across both public and private insurance programs,[9] whereas the NTAP approach is specifically used within Medicare,[11] presenting only a partial view of national AI usage. Additionally, because of its extensive and long-term adoption by health care payers, CPT is also an informative resource for comparing baseline usage rates of non-AI devices.

Although an increasing number of CPT codes have been made available for medical AI devices, these codes are generally spread across various medical domains and reserved for medical coders and insurance companies. As such, there currently does not exist a single database of AI-related CPT codes or a systematic analysis of their usage. In this article, we identify and organize a comprehensive list of CPT codes that apply to medical AI devices. We analyze the usage of these codes on a large national claims database and present their temporal and geographic trends.

RELATED WORKS

Previous analyses have focused on translational roadblocks for medical AI devices stemming from model evaluation, ethics, and reporting.[2,12] Specific studies have shown how AI algorithms can perform worse in clinical practice despite promising retrospective evaluations.[13,14] A variety of studies have analyzed the emergence of reimbursement mechanisms for medical AI products. For example, researchers have highlighted Viz.ai's NTAP model and its potential impact on stroke care, as well as the economic challenges of adopting LumineticsCore from a cost–benefit perspective.[11,15] Current payment models for AI have been previously analyzed along with examples of reimbursable AI devices.[7] More specifically, a recent study has proposed a framework for analytically determining the value and cost of each unique AI service in order to encourage ethical and optimal deployment.[8]

Although our work is the first to analyze AI usage through CPT codes, several studies have analyzed geographic distributions present in AI development. For example, researchers have analyzed PubMed for the training datasets used in various medical AI algorithms and found that the data are disproportionately located in California, Massachusetts, and New York.[16] Datasets used in AI skin cancer diagnosis have also been exclusively found to be from Europe, North America, and Oceania.[17] The usage of CPT codes for digital health technologies like remote physiologic monitoring, e-consults, and e-visits have also been systematically studied by reporting the total number of claims in Medicare data.[18] Our work focuses specifically on the subset of digital health relevant to AI and machine learning (ML).

Methods

Our analysis consists of two main parts: the organization of medical AI device CPT codes and the analysis of their usage. First, to find CPT codes used for medical AI devices, we used a combination of official sources, Web resources, and insurance company policies. Second, we searched a large national claims database to quantify the usage of each code.

COLLECTING CPT CODES FOR MEDICAL AI DEVICES

Official AMA Sources

The AMA develops CPT codes and is responsible for the development of new billing codes for medical AI products. The CPT Editorial Panel has issued guidance for classifying AI applications, which includes assistive, augmentative, and autonomous work,[19] but only a few examples of AI codes are referenced. For a comprehensive list of new CPT codes, we processed the AMA's list of Category III codes (accessed March 1, 2023[20]), which are a set of temporary codes assigned to emerging technologies, services, and procedures.[21] Although these codes are billed like all other codes, Category III codes are intended to be used primarily for data collection to substantiate widespread usage before granting reimbursement. After 5 years, they are reevaluated and replaced with a Category I code if deemed qualified. We analyzed each of the AMA's Category III codes (long descriptors) for the terms *artificial intelligence* and *machine learning* and their variants. Next, for Category I and II codes, we performed a comprehensive search using Codify by AAPC, a search engine for CPT codes.[22]

CPT code long descriptors provide limited information on the underlying technology behind the procedure and the product name. Therefore, we complemented the CPT code descriptions with details provided by insurance companies in policy documents. Such documents provide detailed descriptions of a given procedure, as well as any medical evidence that might support the case for its reimbursement. Additionally, the policies often reference specific product names that the CPT codes refer to. We analyzed the policies of Premera, Amerigroup, and Blue Cross and noted products that were referred to as AI or ML devices.[23-25]

Determining AI Devices

We determined whether each candidate CPT code bills for an AI medical device if either of the following criteria was met: the device manufacturer makes explicit marketing claims that its product uses AI and/or ML, or a third party (e.g., insurance company or news publication) refers to the product as powered by AI and/or ML. Additionally, we excluded CPT codes that are also used for billing non-AI devices, because this dilutes the number of AI-specific occurrences. For example, recently, AI has been integrated into a continuous glucose monitoring device, but other non-AI devices are billed under the same code. Another example includes mammography with CAD, which is largely dominated by traditional CAD and should be differentiated from modern CAD products.[26] As a whole, radiology AI devices are underrepresented in our analysis relative to their share of all FDA-approved AI devices[27] because they are commonly billed using existing CPT codes that are not specific to AI. However, more procedures in areas like cardiology (e.g., HeartFlow's FFRCT analysis) do not have non-AI counterparts, allowing for the creation of new CPT codes that are AI specific. Next, several CPT codes exist for ML-based proprietary laboratory tests (identified with the letter "U"), but are excluded from this study because they are typically designed and deployed in specific laboratories and are outside the FDA's purview.[28] Finally, to focus our analysis on the usage of recently developed AI, we include only CPT codes developed after 2015.

Grouping CPT Codes

Multiple CPT codes may be related to the same underlying medical procedure but describe different aspects of

the procedure. For example, both 0648T and 0649T are used to report quantitative magnetic resonance (MR) analysis of tissue composition, but 0649T is used when diagnostic magnetic resonance imaging (MRI) is also completed, whereas 0648T is used when it is not. In our analysis, we organized codes that refer to the same underlying medical AI procedure into a CPT code group. To this end, we computed the sum total of all codes in that code group when reporting the number of claims for each procedure.

CLAIMS DATA

IQVIA PharMetrics® Plus

We used the IQVIA PharMetrics® Plus for MedTech dataset, a longitudinal health plan database of medical and pharmacy claims.[29] The dataset consists of more than 210 million unique U.S. enrollees and comprises largely commercial health plans. The data are compliant with the Health Insurance Portability and Accountability Act and are representative of the commercially insured U.S. national population for patients under 65 years of age.[30] The IQVIA dataset is commonly used for analyses of medical trends in areas like infectious diseases,[31-35] cardiology,[36] dermatology,[37] pulmonology,[38] oncology,[39] and neurology.[40-42] The unit measurement we used in our analysis is a medical claim that uses a CPT code associated with a medical AI procedure. We analyzed usage in all 50 U.S. states from January 1, 2015, to June 1, 2023; the dataset consists of 16 billion claims in total, with 11 billion claims after 2018. We have included a table that details the number of claims in our dataset for each year between 2015 and 2023 (Table S3). As a point of reference, CMS reports that there are a total of 5 billion claims processed in the United States per year,[43] which suggests that our dataset has approximately 40% coverage of all U.S. claims.

Finding Associated Device Names and FDA Approvals

To provide the commercial context for each CPT code, we also located specific device names associated with each AI CPT code by searching through insurance policies as well as company websites. Although we were able to locate at least one product for each procedure, the list may not be comprehensive if a product was not indicated by the company or a third-party source. For the top products we found, we also located their corresponding FDA approval (if applicable) to provide a timeline context for the overall translational pipeline for each product.

Geographic Analysis

For each medical AI procedure, we aggregated all unique zip codes that contained an occurrence of at least one code. First, we searched for each zip code's median income and classified it as high income if it exceeded $100,000 per year, consistent with the IRS's classification.[44] Next, we determined whether it was in a metropolitan area by referencing the U.S. Department of Agriculture's Rural-Urban Commuting Area Codes.[45] Finally, we computed the percentage of all unique zip codes that had a high median income level and were metropolitan. We compared these rates with the rates found for all U.S. zip codes, as well as unique zip codes found in a random sample of 1 million claims (across all CPT codes).

Insurance Pricing

In addition to CPT code billing frequencies, we collected public and private pricing information. First, when available, we looked up Medicare pricing for each CPT code that had been made publicly available each year.[46] Second, we gathered negotiated pricing rate data from Anthem Healthcare in California and New York, focusing specifically on in-network rates as of November 2022. These data are made available as part of the Transparency in Coverage regulation, which was introduced by the Tri-Agencies (U.S. Departments of Health and Human Services, Labor, and Treasury) on November 12, 2020.[47] The regulation requires health plans to publish their negotiated rates for all items and services for commercial coverage, including in-network files, in machine-readable formats, with monthly updates starting from July 1, 2022. We utilized the November 2022 version of the in-network rate files, which are provided in the CMS-defined JSON (Java Script Object Notation) format.

Results

BILLABLE MEDICAL AI DEVICES

Given our methodology, we found a total of 16 medical AI procedures billable under CPT codes. Several procedures can be reimbursed through multiple codes, comprising a total of 32 unique CPT codes that are associated with AI. These procedures are detailed in Table 1, alongside the total number of claims containing the codes, product name, and effective date of the codes. The procedures fall within a wide range of health care areas, such as cardiology, radiology, and ophthalmology, and were created very

Table 1. Summary of AI CPT Codes.*

Total Claims	Condition or Medical AI Procedure	CPT Code(s)	Example Product Name	Effective Date
67,306	Coronary artery disease	0501T–0504T	HeartFlow Analysis[48]	June 1, 2018
15,097	Diabetic retinopathy	92229	LumineticsCore[49]	January 1, 2021
4,459	Coronary atherosclerosis	0623T–0626T	Cleerly[50]	January 1, 2021
2,428	Liver MR	0648T–0649T	Perspectum LiverMultiScan[51]	January 1, 2021
591	Multiorgan MRI	0697T–0698T	Perspectum CoverScan[52]	January 1, 2022
552	Breast ultrasound	0689T–0690T	Koios DS[53]	January 1, 2022
435	ECG cardiac dysfunction	0764T–0765T	Anumana[50]	January 1, 2023
331	Cardiac acoustic waveform recording	0716T	CADScor[50]	July 1, 2022
237	Quantitative MR cholangiopancreatography	0723T–0724T	Perspectum MRCP+[54]	July 1, 2022
67	Epidural infusion	0777T	CompuFlo[55]	January 1, 2023
4	Quantitative CT tissue characterization	0721T–0722T	Optellum Virtual Nodule Clinic[56]	July 1, 2022
1	Autonomous insulin dosage	0740T–0741T	d-Nav[57]	January 1, 2023
1	CT vertebral fracture assessment	0691T	HealthVCF[50]	January 1, 2022
1	Noninvasive arterial plaque analysis	0710T–0713T	ElucidVivo[50]	January 1, 2022
0	Facial phenotype analysis	0731T	Face2Gene[50]	July 1, 2022
0	X-ray bone density	0749T	OsteoApp[50]	January 1, 2023

* A total of 16 medical AI procedures are presented alongside their corresponding CPT codes. Each procedure is associated with an example commercial product that may be reimbursed through the codes. The effective date is the date on which the code was officially recognized by the American Medical Association and can be used for billing and reimbursement purposes. The total claims listed are recent as of June 1, 2023. AI denotes artificial intelligence; CPT, Current Procedural Terminology; CT, computed tomography; ECG, electrocardiogram; MR, magnetic resonance; MRCP, magnetic resonance cholangiopancreatography; and MRI, magnetic resonance imaging.

recently, with 15 of 16 medical AI procedures created since 2021 (Fig. 1). We found that only 4 of 16 have more than 1000 total claims. This is partially because the median age of a medical AI procedure is only about a year (374 days).

GROWTH PATTERNS OF MEDICAL AI DEVICES

We found that the overall utilization of medical AI products is still limited and focused on a few leading procedures. However, utilization has generally increased exponentially for each medical AI procedure (Fig. 2). The procedure with the most AI usage is coronary artery disease (n=67,306; effective January 1, 2018). The associated CPT codes can be used to reimburse products like HeartFlow FFRCT, a medical device that uses computed tomography (CT) scans to create a 3D model of the coronary arteries. The model is then used to calculate the fractional flow reserve (FFR), which is a measure of how well blood flows through the arteries. Among other functions, FFRCT can be used to diagnose coronary artery disease and assess the severity of the disease.[58] HeartFlow FFRCT was given its first FDA approval in 2019 and has had two subsequent updates since.[59] In November 2021, CMS set the national payment rate of the device at $930.34 for an office-based setting.[60] Meanwhile, privately negotiated rates from Anthem in California and New York have a median price of $909.77.

Diabetic retinopathy medical AI has also grown exponentially in usage (n=15,097, effective January 1, 2021). The first FDA approval in this category was given on January 12, 2018, to LumineticsCore,[61] an AI diagnostic system that autonomously diagnoses patients for diabetic retinopathy (including macular edema).[62] It is indicated for use by health care providers to automatically detect more than mild diabetic retinopathy in adults diagnosed with diabetes who have not been previously diagnosed with diabetic retinopathy.[61] The product takes images of the back of the eye, analyzes them, and provides a diagnosis. If more than a mild case is detected, the patient is referred to a specialist.[63] In 2021, the national payment rate set by CMS for CPT code 92229 was $45.36,[49] whereas the median privately negotiated rate was $127.81.

We also found exponential growth at a smaller scale occurring in medical AI for coronary atherosclerosis and liver MR. Cleerly's Coronary Computer Tomography Angiography (CCTA) algorithm (n=4459, effective January 1, 2021) received its first FDA approval on October 9, 2019, and aims to identify atherosclerosis, the plaque buildup in the arteries of the heart, as well as vascular morphology features for all identified arteries in the CCTA data.[64] Although pricing for this code is not given through CMS, we found it has a median private negotiated

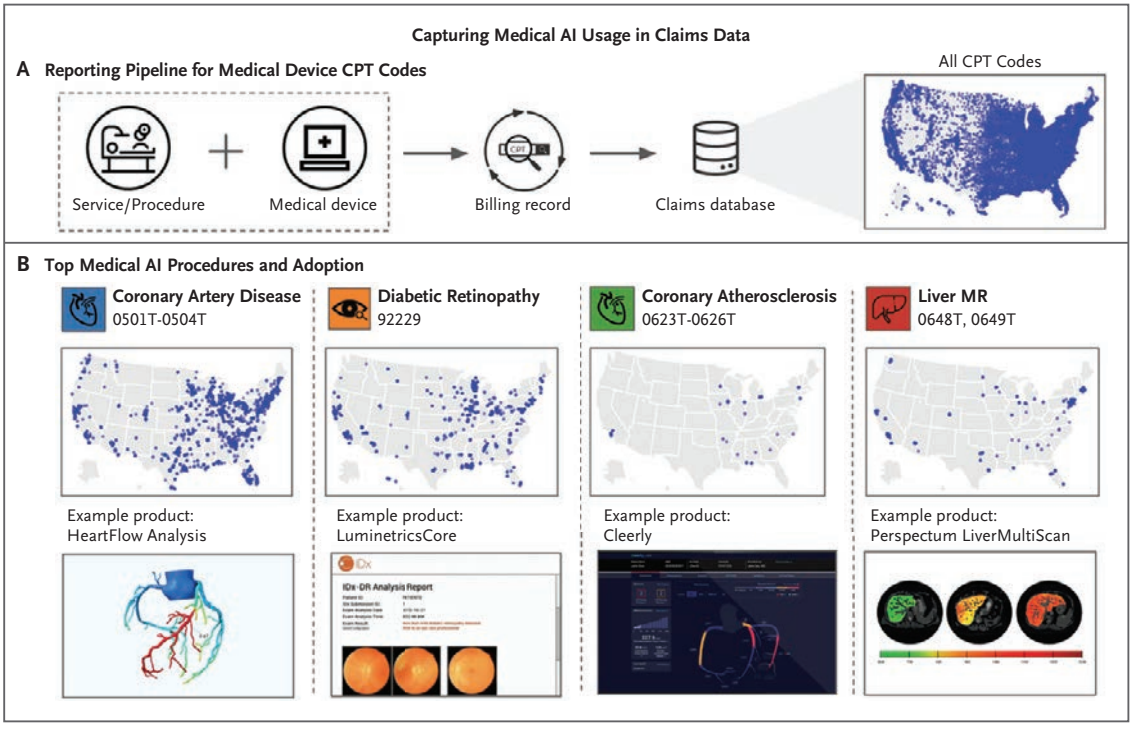


Figure 1. Capturing Medical AI Usage in Claims Data.

Panel A shows the reporting pipeline for CPT codes. Medical device usage is captured through billing records and aggregated in our claims database. Each service/procedure that uses a medical device is associated with a CPT code that hospitals and medical practices report for billing purposes. On the right, we provide a map of the geographic distribution of zip codes for a random sample of 1 million claims (out of 11 billion in our dataset from January 1, 2018, to June 1, 2023) for comparison with AI CPT codes. Each blue dot represents a single unique zip code where the procedure was billed. Panel B provides details on the top four billable medical AI procedures through CPT codes. Under the procedure name, we list the CPT codes and a map of the geographic distribution of zip codes where they have been billed (cumulative over time). In the bottom row, we provide examples of billable products under each code and a product description image taken from marketing materials from the respective companies. AI denotes artificial intelligence; CPT, Current Procedural Terminology; and MR, magnetic resonance.

rate of $692.91. Perspectum's LiverMultiScan (n=2428, effective January 1, 2021) is a noninvasive diagnostic technology for evaluating liver diseases present in multiparametric MRI by quantifying liver tissue.[65] Receiving its FDA approval on September 6, 2017, it provides a number of quantification tools, such as region-of-interest placements, to be used for the assessment of regions of an image to aid in the diagnosis of liver disorders.[59] The associated CPT code, 0648T, does not have a national payment rate through CMS but has a median privately negotiated rate of $371.55. We include a full table of available pricing information in Table S2.

Finally, we also observed that several procedures had only nominal or zero usage. CT Vertebral Fracture Assessment and Noninvasive Arterial Plaque Analysis had only a single occurrence in our CPT database since January 1, 2022, and procedures (Facial Phenotype Analysis and X-Ray Bone Density) did not have any occurrences in our database.

CHARACTERISTICS OF DEPLOYED ZIP CODES

To better understand the drivers of medical AI device adoption, we represented each zip code by three features: whether it had a high median income level (median annual household income greater than $100,000), whether it was metropolitan (classification by the U.S. Department of Agriculture), and whether it had at least one academic hospital (determined by the Association of American Medical Colleges). We performed logistic regression on the outcome variable of AI adoption within a

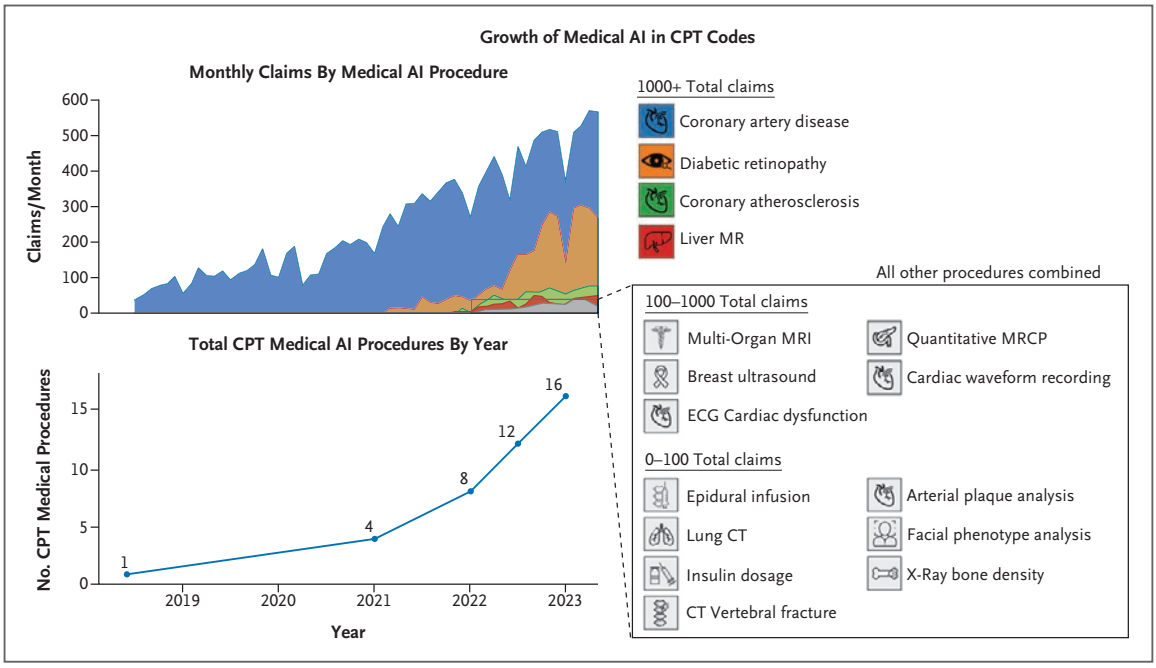


Figure 2. Growth of Medical AI in CPT Codes.

Panel A presents the number of claims per month for each medical AI procedure between January 1, 2018, and June 1, 2023. The top four procedures by total claims are presented in colors, whereas the remaining 12 are grouped and added together into an "Other" group in gray. On the right-hand side, we provide a legend for each of the medical AI procedures. These procedures are further grouped by their usage tiers (0 to 100, 100 to 1000, and ≥1000 total claims). All procedures in the "Other" category are contained in the callout box on the bottom right. Panel B presents the cumulative number of CPT AI medical procedures available each year from 2018 to 2023. AI denotes artificial intelligence; CPT, Current Procedural Terminology; CT, computed tomography; ECG, electrocardiogram; MR, magnetic resonance; MRCP, magnetic resonance cholangiopancreatography; and MRI, magnetic resonance imaging.

zip code (defined as at least one occurrence of billing of an AI CPT code) (Table 2). Only zip codes with at least one institutional NPI (National Provider Identifier) were included in our analysis. In total, we included 22,704 zip codes, of which 2182 had at least one medical AI billing. All three variables were statistically significant (P<0.001), whereas the presence of an academic hospital had the largest effect on the likelihood of AI adoption (17 times more likely). Metropolitan zip codes had the second largest effect (5.25 times more likely), whereas high-income zip codes had a 1.45 times likelihood of AI adoption. Of all zip codes with an academic hospital, 71% had at least one medical AI billing. In contrast, only 9% of zip codes without an academic hospital had at least one medical AI billing. We also found a difference between high income and low income (18% vs. 9%) and metropolitan versus nonmetropolitan (14% vs. 3%) zip codes in whether the area had at least one medical AI billing.

Table 2. A Multivariate Logistic Regression on Whether a Zip Code Has at Least One Documented Billing of a Medical AI CPT Code.*

Zip Code Characteristic	Log-Odds Coefficient
High income	0.373†
Metropolitan	1.65†
Has academic hospital	2.85†

* Only zip codes with at least one institutional NPI (National Provider Identifier) are included in the analysis (n=22,704). AI denotes artificial intelligence, and CPT, Current Procedural Terminology.

† P<0.001.

Consistent with the results from our regression model, 32% of zip codes where AI devices are deployed are high income, which is significantly higher than non-AI claims (17%, P<0.001) as well as the U.S. general population average (10%, P<0.001). An average of 89% of the zip codes for AI are metropolitan, which is much higher than the U.S. average (41%, P<0.001) and marginally higher than the value for the random sample of non-AI claims (87%, P=0.002) (Fig. S2). Additionally, we created a map of the geographic distribution of claims for the top four

medical AI procedures for each year of their availability and found that usage is generally well distributed across coasts and regions of the United States (Fig. S1). We also list the top 10 zip codes and city/state for each of the top four medical AI procedures in our analysis (Table S1).

Discussion

Our study found that the commercialization of FDA-approved AI products is still nascent but growing, with over 50% of CPT codes effective since 2022. However, only a handful of these devices have reached substantial market adoption, suggesting that the medical AI landscape is still in its early stages. Such usage patterns underscore key themes regarding the deployment of AI in medicine, including clinical implementation challenges, payment, and equal access.

Successful clinical adoption of medical AI involves overcoming key implementation barriers. First, the addition of AI may require significant changes to the clinical workflow. For example, studies have detailed how the success of a diabetic retinopathy detection algorithm is mediated by deployment factors like patient consent, Internet speed and connectivity, and poor lighting conditions.[66,67] Another study found that the added benefit of an AI algorithm in pathology depends on the pathologist's interaction with the algorithm's outputs.[68] Moreover, the value of an AI algorithm to clinical practices is a function of its health care setting.[69] For instance, researchers have argued that clinics that use diabetic retinopathy algorithms may operate at a deficit for every patient evaluated and propose modifications to the existing payment structure to encourage adoption.[15] However, patients may be incentivized to visit practices that provide state-of-the-art technologies. Medical AI devices need to have a clear value proposition to health care providers to achieve widespread adoption, but the value of AI is multifaceted and context dependent.[70,71]

In particular, Medicare pricing for medical AI can provide insight into how AI is currently valued. The reimbursement amounts for CPT codes are determined based on three factors: physician work, practice expense, and malpractice cost.[72] For a given code, each factor is associated with a relative value unit (RVU) that is adjusted to account for differences between procedures. For example, a higher RVU for physician work means that the procedure involves more physician time and/or expertise. A key value proposition of medical AI devices is their ability to reduce or remove the work burden of physicians. We find this reflected in the pricing for CPT code 92229 (diabetic retinopathy) in the CMS fee schedule. Despite having a relative value of 0 for physician work, the practice expense relative value (peRVU) for this code is 1.34, which is higher than that of its non-AI counterpart (CPT code 92228, peRVU=0.53).[73] This difference illustrates how the pricing of AI devices shifts some of the value typically assigned to physicians toward the costs of purchasing and operating the device itself.

Interestingly, the privately negotiated rate for diabetic retinopathy is substantially higher than the CMS rate ($127.81 vs. $45.36). Whereas CMS rates are designed with the aim of cost containment for taxpayer-funded programs, private insurers may negotiate rates that better reflect the actual cost or perceived value of services in a specific market. Currently, because of their Category III status, there is no Medicare pricing available for the majority of AI CPT codes. However, in private insurance data, we observed pricing for several devices. For example, AI interpretation of breast ultrasound (CPT codes 0689T–0690T) has a median negotiated reimbursement rate of $371.55, which is comparable to the national average cost of a traditional (non-AI) breast ultrasound of $360.[74] However, AI analysis of cardiac CT for atherosclerosis has a median negotiated rate of $692.91, which is higher than the average cost range of a cardiac CT of $100 to $400.[75] As insurance companies consider reimbursements for emerging AI technologies, determining appropriate pricing remains an important step in wide AI adoption.

The payment mechanism for medical AI has implications for how it will be used and adopted. Although CPT and other procedure-based billing methods like the NTAP method are done on a per-use basis, other payment schemes may adjust for value or outcomes. For example, a recent study of reimbursement strategies has proposed forgoing separate reimbursements altogether, because the near-zero marginal costs of AI may lead to its overuse.[7] Alternatives include a fixed cost with discounts if certain clinical or economic outcomes are not met and a revenue-sharing deal between the AI developers and health care systems.[7] Other outcome-based schemes involve higher reimbursements if certain positive outcomes are demonstrated in a postmarketing study, with early examples in Europe and the United Kingdom.[76] Researchers have also proposed factoring in the proportion of eligible patients who receive a given service in an "access-maximizing" model.[8] Recently,

RadNet, a diagnostic imaging services company, rolled out a program in which patients can opt in for AI interpretation of their mammograms for a $60 out-of-pocket fee.[77]

We observed that the presence of academic medical centers is a significant factor in the adoption of medical AI, as reflected in the fact that over 70% of zip codes with academic centers have at least one medical AI billing, compared with 9% in zip codes without such centers. Furthermore, metropolitan and higher-income zip codes are among those that have a higher likelihood of having AI adoption. These findings are reflective of broader adoption trends within digital health care[78] and are also observed in other emerging technologies like electric cars,[79] because areas with greater resources and infrastructure are better positioned to take on the subsequent risks and rewards. Although such differences in adoption do not necessarily imply disparities in health care outcomes,[80] regulators and stakeholders should consider potential obstacles to equitable access that may be in place as AI becomes a more permanent fixture of health care.

Our analysis of medical AI usage has several limitations. First, although our dataset of 16 billion claims (IQVIA PharMetrics® Plus) is representative of the U.S. patient population less than 65 years of age, it does not capture all medical claims. As such, the number of claims reported in our work only represents a fraction of total usage and should mainly be interpreted through its relative magnitude over time. Second, our analysis focuses specifically on CPT codes, which do not capture all potential types of AI usage. For example, products such as Viz.ai's large vessel occlusion detection algorithm are reimbursable under Medicare's NTAP program, but we did not capture such usage in our study. Additionally, medical AI usage in clinical pilot studies that are not reimbursed will not appear in large national databases. Furthermore, AI software included as part of a hardware system often does not include separate CPT billing. For example, GE's Edison Digital Health Platform and Critical Care Suite are included as part of their AMX and Definium x-ray systems but are not available as a separate software offering. Our analysis also does not capture the usage of medical AI devices that are billed under non–AI-specific CPT codes. For example, CPT code 77066 is used for CAD for mammograms but does not differentiate the usage of current deep learning–based approaches from older traditional models from the 1990s. As such, although new models for mammography are developed and approved by the FDA, their usage cannot be cleanly identified in claims data.

Our analysis focuses specifically on AI Software as a Medical Device, which is a subset of all medical AI. For example, proprietary laboratory analyses can often involve ML algorithms that analyze the collected data. Products like KidneyIntelX and RenalytixAI use AI in diabetes clinical care, whereas PreciseDx provides a breast cancer test. Although such products are billable under CPT codes, they are not regulated through the FDA. Another example is AI in practice management software, which is often implemented through electronic health record vendor software. For example, Epic has been reported to have about 20 predictive algorithms.[81] These applications are also not regulated through the FDA and are primarily paid for as part of a larger software subscription. Finally, with recent innovations in large language models, applications to areas like question-answering and clinical note-taking have emerged. However, such products are still yet to be clinically validated and regulated.[82]

As CPT codes are developed by the AMA for use within the United States, our analysis of AI adoption does not provide direct insight into other countries. However, broad trends in the clinical adoption of AI can be shared across countries because of the similarities of the underlying AI technology and the incentives of health care providers. For example, a recent survey by researchers in the Netherlands found that clinical use of AI is much greater in academic hospitals (57%) than in general hospitals (14%), which is reflected in our study as well.[83] At the same time, several factors make the United States a distinct marketplace for AI. For example, in contrast to single-payer systems, the United States has a mixture of private and public payers, in which CMS establishes a payment policy and private payers follow later.[84] Differences across regulatory agencies can also affect the degree of trust providers have regarding AI. An FDA approval, for instance, always requires a full clinical trial, whereas a CE mark (European Union equivalent) accepts a review of published data from existing devices.[85] Finally, the market for AI health care is significantly larger in the United States, with North America accounting for nearly 50% of the global market in 2022.[86] Such factors affect the entire AI adoption pipeline, from product investment and development to reimbursement and usage.

The small percentage that AI usage takes up relative to total billing highlights the inherent barriers to uptake and the current clinical usefulness and necessity of AI. Although AI CPT codes represent a frontier in terms of the maturity of AI, they are just one of many steps required

for the wide adoption of medical AI.[87,88] For example, a survey of health care providers regarding CE-marked AI devices in radiology found the main obstacles to uptake involved budgeting and information technology integration — issues that are beyond the scope of clinical validation alone.[83] As such, our findings reflect the fact that successful uptake of AI requires understanding the entire translational pipeline of AI technology. The usage and adoption of medical AI are the product of a complex ecosystem involving AI developers, health care providers, payers, and patients. Although the last few years have seen rapid growth in the capabilities of AI, careful consideration of forces beyond algorithmic development is required for AI models to have a meaningful clinical impact. As such, monitoring the usage and clinical adoption of medical AI is key to ensuring that these new technologies fulfill the promise of improving the quality of health care for broad patient populations.

Disclosures

Author disclosures and other supplementary materials are available at ai.nejm.org.

Supported by funding from the Chan-Zuckerberg Initiative.

The IQVIA PharMetrics Plus claims dataset utilized in this research is available for licensing or for research through IQVIA. The insurance pricing data employed in this study are publicly accessible as part of the Transparency in Coverage regulation and can be obtained directly through the Centers for Medicare & Medicaid Services at https://www.cms.gov/healthplan-price-transparency.

Author Affiliations

[1] Department of Biomedical Data Science, Stanford University, Stanford, CA

[2] Department of Electrical Engineering, Stanford University, Stanford, CA

[3] Department of Computer Science, University of Illinois Urbana-Champaign, Urbana, IL

[4] Department of Computer Science, Stanford University, Stanford, CA

[5] IQVIA, Durham, NC

[6] Carle Illinois College of Medicine, University of Illinois Urbana-Champaign, Urbana, IL

References

1. U.S. Food and Drug Administration. Artificial intelligence and machine learning (AI/ML)-enabled medical devices. October 5, 2022 (https://www.fda.gov/medical-devices/software-medical-device-samd/artificial-intelligence-and-machine-learning-software-medical-device).
2. Wu E, Wu K, Daneshjou R, Ouyang D, Ho DE, Zou J. How medical AI devices are evaluated: limitations and recommendations from an analysis of FDA approvals. Nat Med 2021;27:582-584. DOI: 10.1038/s41591-021-01312-x.
3. Wong A, Otles E, Donnelly JP, et al. External validation of a widely implemented proprietary sepsis prediction model in hospitalized patients. JAMA Intern Med 2021;181:1065-1070. DOI: 10.1001/jamainternmed.2021.2626.
4. Wu, E., Wu, K. & Zou, J. Explaining medical AI performance disparities across sites with confounder Shapley value analysis. Cornell University. November 12, 2021 (https://arxiv.org/abs/2111.08168).
5. Lehman CD, Wellman RD, Buist DS, Kerlikowske K, Tosteson AN, Miglioretti DL. Diagnostic accuracy of digital screening mammography with and without computer-aided detection. JAMA Intern Med 2015;175:1828-1837. DOI: 10.1001/jamainternmed.2015.5231.
6. Fenton JJ. Is it time to stop paying for computer-aided mammography? JAMA Intern Med 2015;175:1837-1838. DOI: 10.1001/jamainternmed.2015.5319.
7. Parikh RB, Helmchen LA. Paying for artificial intelligence in medicine. NPJ Digit Med 2022;5:63. DOI: 10.1038/s41746-022-00609-6.
8. Abràmoff MD, Roehrenbeck C, Trujillo S, et al. A reimbursement framework for artificial intelligence in healthcare. NPJ Digit Med 2022;5:72. DOI: 10.1038/s41746-022-00621-w.
9. American Medical Association. CPT® overview and code approval. (https://www.ama-assn.org/practice-management/cpt/cpt-overview-and-code-approval).
10. Hirsch JA, Leslie-Mazwi TM, Nicola GN, et al. Current procedural terminology; a primer. J Neurointerv Surg 2015;7:309-312. DOI: 10.1136/neurintsurg-2014-011156.
11. Hassan AE. New Technology Add-On Payment (NTAP) for Viz LVO: a win for stroke care. J Neurointerv Surg 2021;13:406-408. DOI: 10.1136/neurintsurg-2020-016897.
12. Wiens J, Saria S, Sendak M, et al. Do no harm: a roadmap for responsible machine learning for health care [published correction appears at Nat Med 2019;25:1627]. Nat Med 2019;25:1337-1340. DOI: 10.1038/s41591-019-0548-6.
13. Kanagasingam Y, Xiao D, Vignarajan J, Preetham A, Tay-Kearney ML, Mehrotra A. Evaluation of artificial intelligence-based grading of diabetic retinopathy in primary care. JAMA Netw Open 2018;1:e182665. DOI: 10.1001/jamanetworkopen.2018.2665.
14. National Heart, Lung, and Blood Institute, National Institutes of Health. Widely used sepsis prediction tool is less effective than Michigan doctors thought. June 29, 2021 (https://www.nhlbi.nih.gov/news/2021/widely-used-sepsis-prediction-tool-less-effective-michigan-doctors-thought).
15. Chen EM, Chen D, Chilakamarri P, Lopez R, Parikh R. Economic challenges of artificial intelligence adoption for diabetic retinopathy. Ophthalmology 2021;128:475-477. DOI: 10.1016/j.ophtha.2020.07.043.
16. Kaushal A, Altman R, Langlotz C. Geographic distribution of US cohorts used to train deep learning algorithms. JAMA 2020;324:1212-1213. DOI: 10.1001/jama.2020.12067.
17. Wen D, Khan SM, Ji Xu A, et al. Characteristics of publicly available skin cancer image datasets: a systematic review. Lancet Digit Health 2022;4:e64-e74. DOI: 10.1016/S2589-7500(21)00252-1.

Back to Table of Contents

18. Kvedar JC, Mittermaier M, Pritzker J. The industry impact of the American Medical Association's Digital Medicine Payment Advisory Group (DMPAG). NPJ Digit Med 2022;5:193. DOI: 10.1038/s41746-022-00743-1.

19. American Medical Association. CPT Appendix S: AI taxonomy for medical services & procedures. August 12, 2022 (https://www.ama-assn.org/practice-management/cpt/cpt-appendix-s-ai-taxonomy-medical-services-procedures).

20. American Medical Association. Category III codes. June 30, 2023 (https://www.ama-assn.org/practice-management/cpt/category-iii-codes).

21. Thorwarth WT Jr. CPT®: an open system that describes all that you do. J Am Coll Radiol 2008;5:555-560. DOI: 10.1016/j.jacr.2007.10.004.

22. AAPC. Codify by AAPC features. 2022 (https://www.aapc.com/codify/features.aspx).

23. Premera. Medical policies. 2023 (https://www.premera.com/wa/provider/reference/medical-policies-search/?q=artificial%20intelligence&af1eq=&hpp=20&p=PBC-MedicalPolicy&).

24. Amerigroup. Medical policies and clinical UM guidelines. 2021 (https://medpol.providers.amerigroup.com/green-provider/medical-policies-and-clinical-guidelines).

25. Blue Cross of Idaho. Search results. (https://providers.bcidaho.com/policy-search-results.page).

26. Gao Y, Geras KJ, Lewin AA, Moy L. New frontiers: an update on computer-aided diagnosis for breast imaging in the age of artificial intelligence. AJR Am J Roentgenol 2019;212:300-307. DOI: 10.2214/AJR.18.20392.

27. Benjamens S, Dhunnoo P, Meskó B. The state of artificial intelligence-based FDA-approved medical devices and algorithms: an online database. NPJ Digit Med 2020;3:118. DOI: 10.1038/s41746-020-00324-0.

28. U.S. Food and Drug Administration. Laboratory developed tests. 2023 (https://www.fda.gov/medical-devices/in-vitro-diagnostics/laboratory-developed-tests).

29. IQVIA. IQVIA PharMetrics® Plus. 2023 (https://www.iqvia.com/locations/united-states/library/fact-sheets/iqvia-pharmetrics-plus).

30. Snowflake. IQVIA PharMetrics® Plus claims data. 2023 (https://app.snowflake.com/marketplace/listing/GZSOZ5C6AR/iqvia-iqvia-pharmetrics%C2%AE-plus-claims-data).

31. Feuerstadt P, Stong L, Dahdal DN, Sacks N, Lang K, Nelson WW. Healthcare resource utilization and direct medical costs associated with index and recurrent *Clostridioides difficile* infection: a real-world data analysis. J Med Econ 2020;23:603-609. DOI: 10.1080/13696998.2020.1724117.

32. Chen J, Ferre C, Ouyang L, Mohamoud Y, Barfield W, Cox S. Changes and geographic variation in rates of preterm birth and stillbirth during the prepandemic period and COVID-19 pandemic, according to health insurance claims in the United States, April–June 2019 and April–June 2020. Am J Obstet Gynecol MFM 2022;4:100508. DOI: 10.1016/j.ajogmf.2021.100508.

33. Chua K-P, Conti RM, Becker NV. US insurer spending on ivermectin prescriptions for COVID-19. JAMA 2022;327:584-587. DOI: 10.1001/jama.2021.24352.

34. Barrett CE, Koyama AK, Alvarez P, et al. Risk for newly diagnosed diabetes >30 days after SARS-CoV-2 infection among persons aged <18 years — United States, March 1, 2020–June 28, 2021. MMWR Morb Mortal Wkly Rep 2022;71:59-65. DOI: 10.15585/mmwr.mm7102e2.

35. Jolles S, Smith BD, Vinh DC, et al. Risk factors for severe infections in secondary immunodeficiency: a retrospective US administrative claims study in patients with hematological malignancies. Leuk Lymphoma 2022;63:64-73. DOI: 10.1080/10428194.2021.1992761.

36. Berger JS, Laliberté F, Kharat A, et al. Real-world effectiveness and safety of rivaroxaban versus warfarin among non-valvular atrial fibrillation patients with obesity in a US population. Curr Med Res Opin 2021;37:881-890. DOI: 10.1080/03007995.2021.1901223.

37. Murage MJ, Anderson A, Casso D, et al. Treatment patterns, adherence, and persistence among psoriasis patients treated with biologics in a real-world setting, overall and by disease severity. J Dermatolog Treat 2019;30:141-149. DOI: 10.1080/09546634.2018.1479725.

38. Most JF, Ambrose CS, Chung Y, et al. Real-world assessment of asthma specialist visits among U.S. patients with severe asthma. J Allergy Clin Immunol Pract 2021;9:3662-3671.e1. DOI: 10.1016/j.jaip.2021.05.003.

39. Cohen JT, Lin PJ, Sheinson DM, et al. Are National Comprehensive Cancer Network evidence block affordability ratings representative of real-world costs? An evaluation of advanced non-small-cell lung cancer. J Oncol Pract 2019;15:e948-e956. DOI: 10.1200/JOP.19.00241.

40. dosReis S, Saini J, Hong K, Reeves G, Spence OM. Trends in antipsychotic use for youth with attention-deficit/hyperactivity disorder and disruptive behavior disorders. Pharmacoepidemiol Drug Saf 2022;31:810-814. DOI: 10.1002/pds.5445.

41. Kalilani L, Faught E, Kim H, et al. Assessment and effect of a gap between new-onset epilepsy diagnosis and treatment in the US. Neurology 2019;92:e2197-e2208. DOI: 10.1212/WNL.0000000000007448.

42. Royston M, Kielhorn A, Weycker D, et al. Neuromyelitis optica spectrum disorder: clinical burden and cost of relapses and disease-related care in US clinical practice. Neurol Ther 2021;10:767-783. DOI: 10.1007/s40120-021-00253-4.

43. CMS.gov. HCPCS - general information. August 21, 2023 (https://www.cms.gov/medicare/coding/medhcpcsgeninfo).

44. IRS. IRS increases visits to high-income taxpayers who haven't filed tax returns. August 1, 2023 (https://www.irs.gov/newsroom/irs-increases-visits-to-high-income-taxpayers-who-havent-filed-tax-returns).

45. USDA, Economic Research Service. Rural-urban commuting area codes. September 25, 2023 (https://www.ers.usda.gov/data-products/rural-urban-commuting-area-codes/).

46. CMS.gov. Physician fee schedule look-up tool. September 27, 2023 (https://www.cms.gov/medicare/medicare-fee-for-service-payment/pfslookup).

Back to Table of Contents

47. CMS.gov. Transparency in Coverage final rule fact sheet (CMS-9915-F). October 29, 2020 (https://www.cms.gov/newsroom/fact-sheets/transparency-coverage-final-rule-fact-sheet-cms-9915-f).

48. HeartFlow. Our technology core. 2023 (https://www.heartflow.com/heartflow-ffrct-analysis/article/our-technology-core/).

49. Larson PM. Medicare: what's new for 2022. Review of Ophthalmology. January 10, 2022 (https://www.reviewofophthalmology.com/article/medicare-whats-new-for-2022).

50. Premera Blue Cross. Medical Policy — 10.01.533 Non-covered experimental/investigational services. 2023 (https://www.premera.com/medicalpolicies/10.01.533.pdf).

51. Perspectum. Perspectum announces American Medical Association issues unique category III CPT® codes for LiverMultiScan quantitative multiparametric MR. (https://www.perspectum.com/our-company/news/perspectum-announces-american-medical-association-issues-unique-category-iii-cpt-codes-for-livermultiscan-quantitative-multiparametric-mr/).

52. Perspectum. FDA grants clearance to Perspectum's CoverScan — a new platform-based tool to assess multiple organs in one MRI scan. May 22, 2022 (https://www.perspectum.com/our-company/news/fda-grants-clearance-to-perspectum-s-coverscan-a-new-platform-based-tool-to-assess-multiple-organs-in-one-mri-scan/).

53. AuntMinnie.com. Koios touts CPT codes for ultrasound AI decision-support software. November 3, 2022 (https://www.auntminnie.com/index.aspx?sec=log&itemID=138483).

54. Perspectum. MRCP+. 2023 (https://www.perspectum.com/our-products/mrcpplus/).

55. Milestone Scientific. American Medical Association issues technology-specific CPT® code for Milestone Scientific's CompuFlo Epidural System. July 6, 2022 (https://www.milestonescientific.com/american-medical-association-issues-technology-specific-cpt-code-for-milestone-scientifics-compuflo-epidural-system).

56. Lassiter R. Press Release: CMS assigns new technology payment classification for Optellum's Lung Cancer Prediction score. Optellum. June 28, 2022 (https://optellum.com/2022/06/press-release-cms-assigns-new-technology-payment-classification-for-optellums-lung-cancer-prediction-score/).

57. Business Wire. Hygieia announces new CPT codes for autonomous insulin dose titration. July 26, 2022 (https://www.businesswire.com/news/home/20220726005310/en/Hygieia-announces-new-CPT-codes-for-autonomous-insulin-dose-titration).

58. Heartflow. Category III CPT® Codes for FFRCT. 2017 (https://cdn-corpweb.heartflow.com/assets/docs/Category_III_CPT_Codes_for_FFRCT.pdf).

59. Brindza LJ. What is a premarket notification 510(k)? Clin Microbiol Newsl 1980;2:4-5. DOI: 10.1016/S0196-4399(80)80161-9.

60. Walter M. CMS sets payment rate for HeartFlow's FFR-CT solution in 2022 Medicare Physician Fee Schedule. Cardiovascular Business. November 3, 2021 (https://cardiovascularbusiness.com/topics/cardiac-imaging/cms-heartflow-2022-medicare-physician-fee-schedule).

61. U.S. Food and Drug Administration. FDA permits marketing of artificial intelligence-based device to detect certain diabetes-related eye problems. April 12, 2018 (https://www.fda.gov/news-events/press-announcements/fda-permits-marketing-artificial-intelligence-based-device-detect-certain-diabetes-related-eye).

62. Digital Diagnostics. IDx-DR. 2022 (https://www.digitaldiagnostics.com/products/eye-disease/idx-dr-eu/).

63. Savoy M. IDx-DR for diabetic retinopathy screening. Am Fam Physician 2020;101:307-308. https://www.aafp.org/pubs/afp/issues/2020/0301/p307.html.

64. American Academy of Ophthalmology. Medicare carrier underprices new AI screening code, breaking from peers. March 25, 2021 (https://www.aao.org/eye-on-advocacy-article/medicare-carrier-underprices-new-ai-screening-code).

65. LiverMultiScan. (https://perspectum-diagnostics.euwest01.umbraco.io/products/livermultiscan).

66. Beede E, Baylor E, Hersch F, et al. A human-centered evaluation of a deep learning system deployed in clinics for the detection of diabetic retinopathy. Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems. 2020:1-12. DOI: 10.1145/3313831.3376718.

67. Widner K, Virmani S, Krause J, et al. Lessons learned from translating AI from development to deployment in healthcare. Nat Med 2023;29:1304-1306. DOI: 10.1038/s41591-023-02293-9.

68. Kiani A, Uyumazturk B, Rajpurkar P, et al. Impact of a deep learning assistant on the histopathologic classification of liver cancer. NPJ Digit Med 2020;3:23. DOI: 10.1038/s41746-020-0232-8.

69. Hendrix N, Veenstra DL, Cheng M, Anderson NC, Verguet S. Assessing the economic value of clinical artificial intelligence: challenges and opportunities. Value Health 2022;25:331-339. DOI: 10.1016/j.jval.2021.08.015.

70. Ruamviboonsuk P, Chantra S, Seresirikachorn K, Ruamviboonsuk V, Sangroongruangsri S. Economic evaluations of artificial intelligence in ophthalmology. Asia Pac J Ophthalmol (Phila) 2021;10:307-316. DOI: 10.1097/APO.0000000000000403.

71. Pietris J, Lam A, Bacchi S, Gupta AK, Kovoor JG, Chan WO. Health economic implications of artificial intelligence implementation for ophthalmology in Australia: a systematic review. Asia Pac J Ophthalmol (Phila) 2022;11:554-562. DOI: 10.1097/APO.0000000000000565.

72. Seidenwurm DJ, Burleson JH. The medicare conversion factor. AJNR Am J Neuroradiol 2014;35:242-243. DOI: 10.3174/ajnr.A3674.

73. CMS.gov. PFS relative value files. September 6, 2023 (https://www.cms.gov/Medicare/Medicare-Fee-for-Service-Payment/PhysicianFeeSched/PFS-Relative-Value-Files).

74. New Choice Health. Breast ultrasound cost and flat cost comparison. (http://www.newchoicehealth.com/Directory/Procedure/58/Breast%20Ultrasound).

75. Yetman D. How much does a coronary calcium scan cost? Healthline. November 9, 2022 (https://www.healthline.com/health/heart/coronary-calcium-scan-cost).

76. Koehring M. An introduction to value-based healthcare in Europe. Economist. April 25, 2015 (https://impact.economist.com/perspectives/health/introduction-value-based-healthcare-europe).

Back to Table of Contents

77. EBCD. Add the power of artificial intelligence to your mammogram: enhanced breast cancer detection. (https://myebcdmammo.com/).

78. Suh J, Horvitz E, White RW, Althoff T. Disparate impacts on online information access during the COVID-19 pandemic. April 22, 2022 (https://www.medrxiv.org/content/10.1101/2021.09.14.21263545v2). Preprint.

79. Reese P. How income disparities in California are affecting the adoption of electric cars. The Sacramento Bee, October 21, 2022 (https://www.sacbee.com/news/local/article266878311.html).

80. Chohlas-Wood A, Coots M, Zhu H, Brunskill E, Goel S. Learning to be fair: a consequentialist approach to equitable decision-making. Cornell University. February 1, 2023 (https://arxiv.org/abs/2109.08792).

81. Ross C. Epic's AI algorithms, shielded from scrutiny by a corporate firewall, are delivering inaccurate information on seriously ill patients. Stat. July 6, 2021 (https://www.statnews.com/2021/07/26/epic-hospital-algorithms-sepsis-investigation/).

82. Meskó B, Topol EJ. The imperative for regulatory oversight of large language models (or generative AI) in healthcare. NPJ Digit Med 2023;6:120. DOI: 10.1038/s41746-023-00873-0.

83. van Leeuwen KG, de Rooij M, Schalekamp S, van Ginneken B, Rutten MJCM. Clinical use of artificial intelligence products for radiology in the Netherlands between 2020 and 2022. Eur Radiol 2023 July 29 (Epub ahead of print). DOI: 10.1007/s00330-023-09991-5.

84. Chen MM, Golding LP, Nicola GN. Who will pay for AI? Radiol Artif Intell 2021;3:e210030. DOI: 10.1148/ryai.2021210030.

85. Reinstein DZ, Kanellopoulos JA. CE Mark versus FDA approval: which system has it right? CRST Global, Europe Edition. February 2015 (https://crstodayeurope.com/articles/2015-feb/ce-mark-versus-fda-approval-which-system-has-it-right).

86. MarketsandMarkets. Artificial intelligence (AI) in healthcare market by offering (hardware, software, services), technology (machine learning, NLP, context-aware computing, computer vision), application, end user and region — global forecast to 2028. January 2023 (https://www.marketsandmarkets.com/Market-Reports/artificial-intelligence-healthcare-market-54679303.html).

87. Varghese J. Artificial intelligence in medicine: chances and challenges for wide clinical adoption. Visc Med 2020;36:443-449. DOI: 10.1159/000511930.

88. Singh RP, Hom GL, Abramoff MD, Campbell JP, Chiang MF. Current challenges and barriers to real-world artificial intelligence adoption for the healthcare system, provider, and the patient. Transl Vis Sci Technol 2020;9:45. DOI: 10.1167/tvst.9.2.45.

Back to Table of Contents

NEJM AI

NEJM AI 2023; 1 (1)
DOI: 10.1056/AIpc2300038

POLICY CORNER

Development Pipeline and Geographic Representation of Trials for Artificial Intelligence/Machine Learning–Enabled Medical Devices (2010 to 2023)

Miquel Serra-Burriel, Ph.D.,[1,2] Luca Locher, B.Sc.,[1] and Kerstin N. Vokinger, M.D., J.D., Ph.D.[1]

Received: July 13, 2023; Revised: September 19, 2023; Accepted: September 29, 2023; Published: November 9, 2023

Abstract

A high number of artificial intelligence/machine learning (AI/ML)-enabled medical devices are currently in development. To understand the development pipeline and worldwide geographic distribution of clinical trials for AI/ML-enabled medical devices that may enter the market in the upcoming years, we analyzed the trends in registration of clinical trials for AI/ML-enabled medical devices between 2010 and 2023 as well as their geographic distribution. We aggregated all registered trials initiated between January 1, 2010, and August 31, 2023, through the World Health Organization's International Clinical Trials Registry Platform and included all clinical studies for AI/ML-enabled medical devices in our study cohort. Among the 710,800 registered clinical trials in this time period, 2669 clinical trials for AI/ML-enabled medical devices were identified and included in our study cohort. Of these, 2517 clinical trials provided information on the locations where the trial was conducted. Most of the trials were conducted for the medical specialties of radiology, general hospital, gastroenterology, and urology. Almost all were national trials; 1095 were conducted in China, followed by the United States (196), Japan (162), India (139), and Korea (118). The countries with the most enrolled patients in clinical trials per 100,000 inhabitants were mainly smaller countries in Asia and Europe. More international trials should be encouraged — including the involvement of low- and middle-income countries — to improve equality and ensure that the algorithms perform well across populations. (Funded by the Swiss National Science Foundation.)

Introduction

The number of approved artificial intelligence/machine learning (AI/ML)-enabled medical devices and those in development have increased in recent years.[1] This trend is expected to continue.[2]

The author affiliations are listed at the end of the article.

Dr. Vokinger can be contacted at kvokinger@llm16.law.harvard.edu or at Academic Chair for Regulation in Law, Medicine, and Technology, Faculty of Law, University of Zurich, Ramistrasse 74, Zurich, Switzerland 8001.

Back to Table of Contents

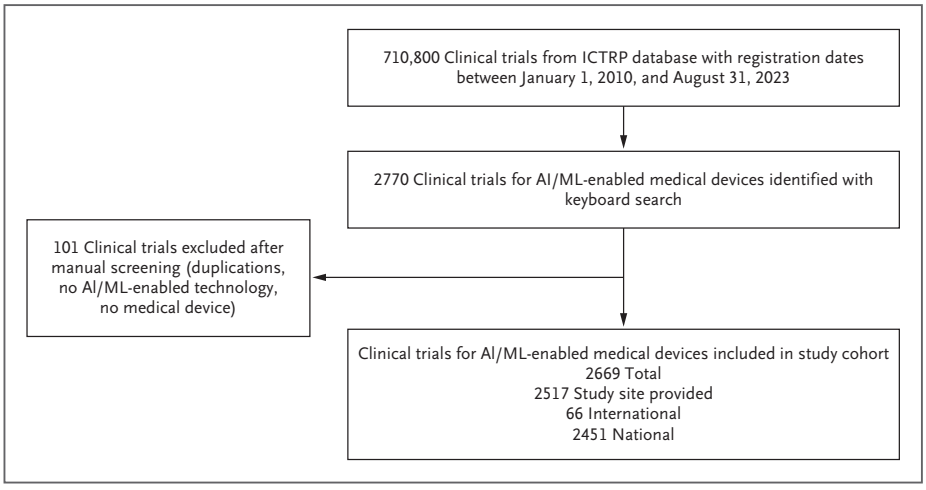


Figure 1. Flow Chart.

AI/ML denotes artificial intelligence/machine learning; and ICTRP, International Clinical Trials Registry Platform.

Previous studies have highlighted the importance of a diverse population representation in clinical studies in order for AI/ML-enabled medical devices to be applicable broadly to patients across health systems.[3-7]

To understand the development pipeline and worldwide geographic distribution of clinical trials for AI/ML-enabled medical devices that may enter the market in the upcoming years, we analyzed the trends in clinical trials for AI/ML-enabled medical devices registered between 2010 and 2023.

Methods

We aggregated all registered trials initiated between January 1, 2010, and August 31, 2023, through the World Health Organization (WHO) International Clinical Trials Registry Platform (ICTRP).[8]

To identify all AI/ML-related trials, we searched the database ICTRP for trials with these keywords: *artificial intelligence OR machine learning OR deep learning OR artificial neural network OR neural network model OR convolutional neural network OR recurrent neural network OR supervised learning OR unsupervised learning OR natural language processing OR generative model OR generative ai OR conversational ai OR large language model OR generative pre trained transformer OR chatgpt*.

We extracted all trials that included at least one of the keywords in the trial's title or description. For each identified clinical trial we then extracted the following information, if available: title of the clinical trial, description of the clinical trial, study sites, and number of patients enrolled. We manually screened all the identified clinical trials and included all clinical trials in our study cohort that included an AI/ML technology and featured a medical device. We determined the medical specialty for each clinical trial using the Food and Drug Administration (FDA) classification as a guidance.[9]

To compare trials for AI/ML-enabled medical devices with non–AI/ML-enabled medical devices, we extracted all clinical trials for medical devices from the database ICTRP.

Descriptive statistics were performed using R, version 4.2.2 (R Foundation for Statistical Computing, Vienna).

Results

OVERVIEW

Among the 710,800 clinical trials registered between January 1, 2010, and August 31, 2023, 2669 clinical trials for AI/ML-enabled medical devices were included in our

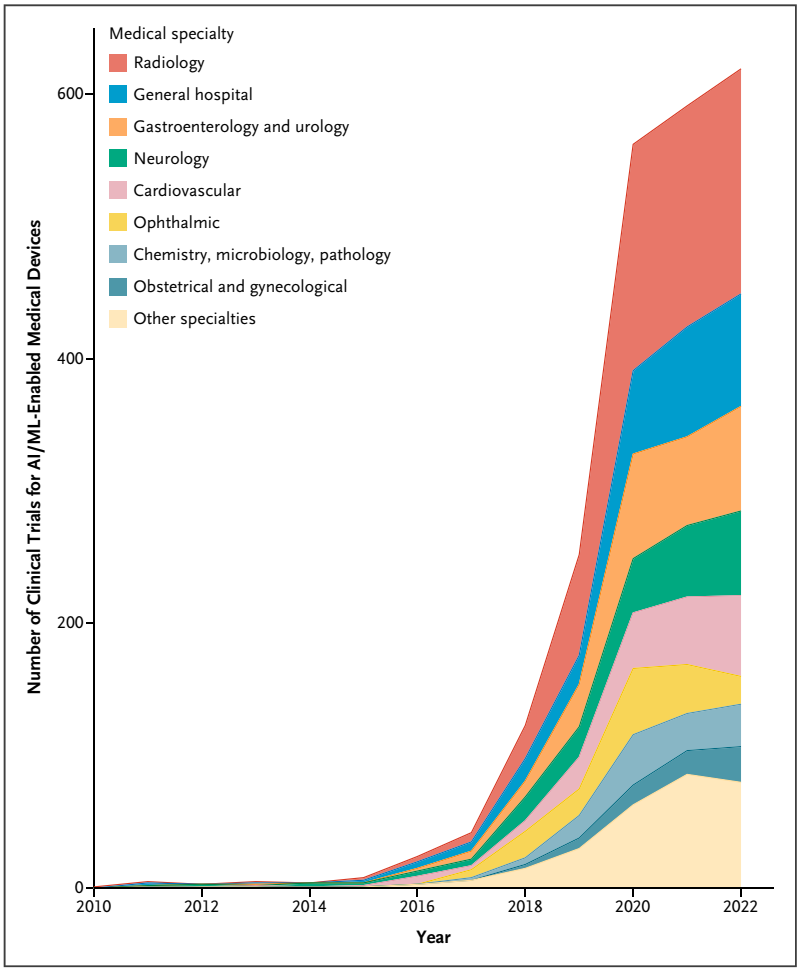


Figure 2. Temporal Trends for Clinical Trials by Medical Specialty.

AI/ML denotes artificial intelligence/machine learning.

study cohort. Of these, 2517 clinical trials provided information on the location(s) where the trial was conducted: 2451 (97%) were conducted in a single country and 66 (3%) were international collaborations (Fig. 1).

The number of clinical trials for AI/ML-enabled medical devices increased from 1 in 2010 to 619 in 2022.

MEDICAL SPECIALTIES

Among the 2669 included clinical trials, most targeted the medical specialty radiology (724, 27%), followed by general hospital (341, 13%), gastroenterology and urology (331, 12%), neurology (264, 10%), and cardiology (228, 9%) (Fig. 2).

However, differences were observed between countries and their focus on the medical specialties in their clinical trials. For example, most of the clinical trials for AI/ML-enabled medical devices conducted in China focused on radiology (382 clinical trials), followed by gastroenterology and urology (157 clinical trials), whereas in the United States, most clinical trials were attributable to general hospital (52 clinical trials), followed by neurology (37 clinical trials) and both radiology (31 clinical trials) and cardiology (31 clinical trials) (Fig. 3).

GEOGRAPHIC DISTRIBUTION

When analyzing the geographic distribution of the 2451 clinical trials for AI/ML-enabled medical devices conducted in single countries, most of them were conducted

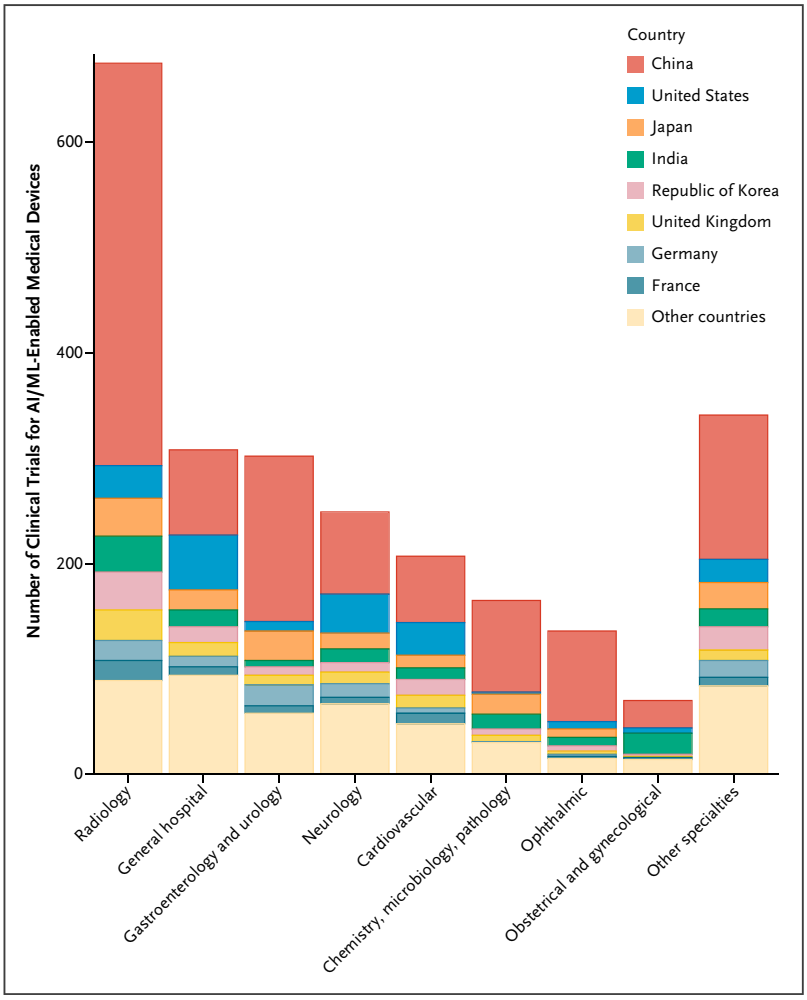


Figure 3. Distribution of Trial Locations by Medical Specialty.

AI/ML denotes artificial intelligence/machine learning.

in China (1095 clinical trials), followed by the United States (196 clinical trials), Japan (162 clinical trials), India (139 clinical trials), and Republic of Korea (118 clinical trials) (Fig. 4A).

This order changed when analyzing the number of all national clinical trials for AI/ML-enabled and non–AI/ML-enabled medical devices (113,815 clinical trials) between January 1, 2010, and August 31, 2023. Most of these clinical trials were conducted in the United States (21,323), followed by China (9809) (Fig. 4B).

China had a total enrollment of approximately 12 million patients in the AI/ML-related trials, followed by Germany (5.5 million), the United Kingdom (3.7 million), the Republic of Korea (1.3 million), Japan (1.1 million), Australia (1.03 million), the United States (441,282), New Zealand (402,534), India (369,323), and Taiwan (265,756) (Fig. 4C). New Zealand, followed by Germany, the United Kingdom, Australia, the Republic of Korea, Sweden, Hong Kong, Denmark, Taiwan, and Switzerland, was the country with the highest patient enrollment in clinical trials per 100,000 inhabitants (Fig. 4D).

On the continental level, Asia and Europe had the largest increase in the number of trials for AI/ML-enabled medical devices between 2010 and 2022. In total, Asia accounted for 68.8% of all clinical trials, Europe for 17.8%, North America for 9.4%, Australia and Oceania for 2.3%, South America for 0.9%, and Africa for 0.8% of the

Back to Table of Contents

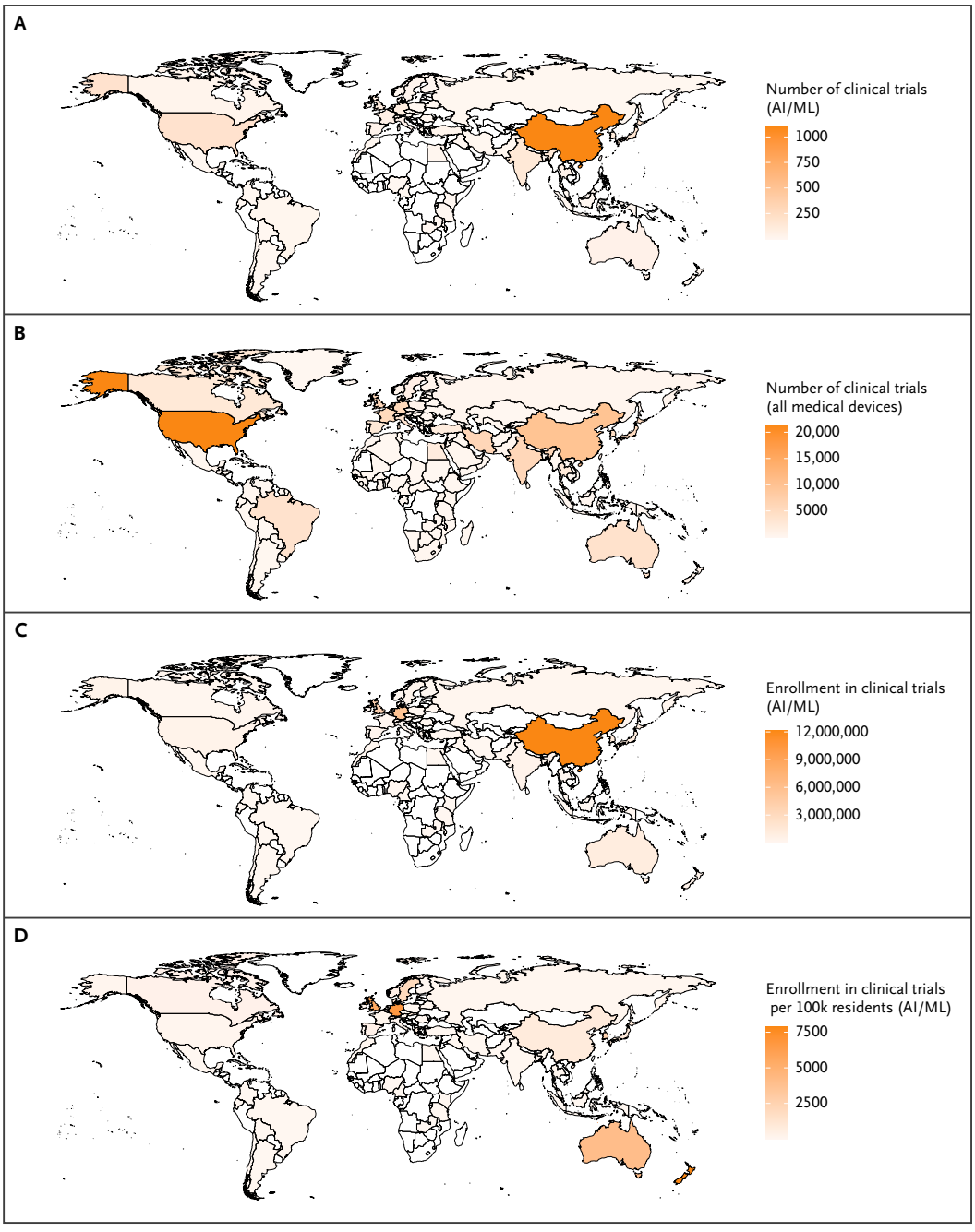


Figure 4. Geographic Distribution of Clinical Trials for AI/ML-Enabled Medical Devices.

Panel A depicts the number of clinical trials for AI/ML-enabled medical devices for each country. Panel B depicts the number of clinical trials for all medical devices. Panel C depicts the sum of enrolled patients in clinical trials for AI/ML-enabled medical devices for each country. Panel D depicts the sum of the enrolled patients in clinical trials for AI/ML-based medical devices per 100,000 inhabitants per country. AI/ML denotes artificial intelligence/machine learning.

Back to Table of Contents

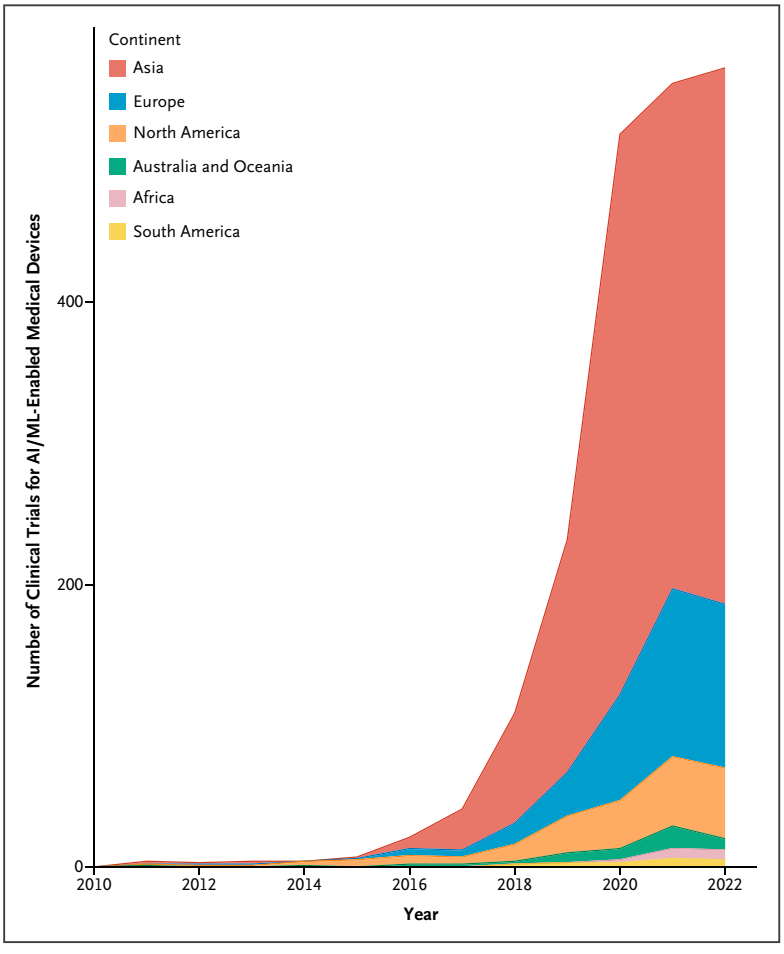


Figure 5. Temporal Trends for Clinical Trials for AI/ML-Enabled Medical Devices by Continent.
AI/ML denotes artificial intelligence/machine learning.

clinical trials registered between January 1, 2010, and December 31, 2022 (Fig. 5).

Discussion

The substantial increase in registered clinical trials for AI/ML-enabled medical devices over the past years indicates that the number of such devices that will be approved and applied in the clinical setting will further increase in the upcoming years. The increase may be even higher because medical devices are often approved without clinical trials.[10,11] Whereas radiology may keep its leading position across medical specialties, other medical specialties[1] such as gastroenterology or urology could introduce more AI/ML-enabled medical devices in the clinical setting in the near future. As of August 31, 2023, only four AI/ML-enabled medical devices for the medical specialty gastroenterology and urology have been cleared in the United States.[2] One example for such a device in the development pipeline is a clinical trial with 200,000 enrolled participants with the goal of evaluating the effect of AI systems used during colonoscopy for the detection of precancerous polyps in the colon (NCT05888623).

The current dominance of national trials instead of international collaborations indicates that the results of the clinical trials may lack external validity. Clinical practice standards are known to differ internationally, and trial and target populations will differ in regions where the trial was conducted.[4-6] As previous studies and the FDA have pointed out, it is crucial that approval agencies, physicians, patients, and the public be informed about the selected data with regard to patient demographic and baseline characteristics, such as number of patients, distribution of

patients' ages, and representation of race and ethnicity as well as gender.[2,4,12-15] This information helps to understand whether a specific AI/ML-enabled medical device is appropriate for the diagnosis or therapy of a patient in the relevant clinical setting. Furthermore, these findings indicate that a stronger collaboration between countries on clinical trials for AI/ML-enabled medical devices is desirable and should be more strongly encouraged to ensure that the algorithms perform well across representative populations.

Reports highlight the global competition between countries on the successful development of AI technologies has emerged over the past decade, with medicine at the forefront of interest.[16-18] Findings in our study show that China, followed by the United States, dominated in terms of absolute number of trials. Smaller countries, mainly in Asia (Korea, Taiwan, and Hong Kong) and Europe (Germany, the United Kingdom, Denmark, and Switzerland), and New Zealand had the most enrolled patients per 100,000 inhabitants. With the exception of China and India, low- and middle-income countries were underrepresented. The involvement of low- and middle-income countries is challenging for different reasons, including the lack of expertise, time, and financial resources such AI/ML-enabled medical devices may require.[19] Nonetheless, it is crucial that these countries also be involved in clinical trials to overcome the selection bias because the demographics of high-income countries may not match those of other countries.[5,20] Moreover, some of these devices — for example, AI/ML-enabled optimal antibiotic treatment strategies for severe bacterial infections (NCT01338116) — would help patients suffering from these diseases with high prevalence and burden in low- and middle-income countries and thus could help to improve the health status of patients and potentially save labor and financial costs in the longer run.

LIMITATIONS

Our study has limitations. It was not always possible to determine whether a clinical trial for a medical device had an AI/ML component. Additionally, we included clinical trials for AI/ML-enabled medical devices only if they were registered at the WHO's ICTRP, which is not comprehensive for all trial registries. However, it can be assumed that the WHO's ICTRP contains the globally major and largest clinical trial registries. Furthermore, not all included medical devices in our study will be cleared and used for patients in the clinical setting. Moreover, a majority of AI/ML-enabled medical devices are cleared without a clinical trial.[10] Thus, our findings are representative not of all AI/ML-enabled medical devices but rather of those for which clinical trials were conducted.

Disclosures

Author disclosures are available at ai.nejm.org.

This study was funded by the Swiss National Science Foundation (grant number 407740_197485).

Author Affiliations

[1] Academic Chair for Regulation in Law, Medicine, and Technology, Faculty of Law, University of Zurich, Zurich, Switzerland

[2] Epidemiology, Biostatistics and Prevention Institute, University of Zurich, Zurich, Switzerland

References

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-e203. DOI: 10.1016/S2589-7500(20)30292-2.

2. U.S. Food and Drug Administration. Artificial intelligence and machine learning (AI/ML)-enabled medical devices. October 5, 2022 (https://www.fda.gov/medical-devices/software-medical-device-samd/artificial-intelligence-and-machine-learning-aiml-enabled-medical-devices).

3. Brajer N, Cozzi B, Gao M, et al. Prospective and external evaluation of a machine learning model to predict in-hospital mortality of adults at time of admission. JAMA Netw Open 2020;3:e1920733. DOI: 10.1001/jamanetworkopen.2019.20733.

4. Kaushal A, Altman R, Langlotz C. Geographic distribution of US cohorts used to train deep learning algorithms. JAMA 2020;324:1212-1213. DOI: 10.1001/jama.2020.12067.

5. Ricci Lara MA, Echeveste R, Ferrante E. Addressing fairness in artificial intelligence for medical imaging. Nat Commun 2022;13:4581. DOI: 10.1038/s41467-022-32186-3.

6. Obermeyer Z, Powers B, Vogeli C, Mullainathan S. Dissecting racial bias in an algorithm used to manage the health of populations. Science 2019;366:447-453. DOI: 10.1126/science.aax2342.

7. Abbasi-Sureshjani S, Raumanns R, Michels BEJ, Schouten G, Cheplygina V. Risk of training diagnostic algorithms on data with demographic bias. In: Cardoso, J, Van Nguyen, H, Heller, N, et al., eds. Interpretable and annotation-efficient learning for medical image computing. Cham, Switzerland: Springer International Publishing, 2020:183–92. DOI: 10.1007/978-3-030-61166-8_20.

8. World Health Organization. International Clinical Trials Registry Platform (https://trialsearch.who.int/).

9. U.S. Food and Drug Administration. Device classification panels. August 31, 2018 (https://www.fda.gov/medical-devices/classify-your-medical-device/device-classification-panels).

10. Wu E, Wu K, Daneshjou R, Ouyang D, Ho DE, Zou J. How medical AI devices are evaluated: limitations and recommendations from an

analysis of FDA approvals. Nat Med 2021;27:582-84. DOI: 10.1038/s41591-021-01312-x

11. Marcus HJ, Payne CJ, Hughes-Hallett A, et al. Regulatory approval of new medical devices: cross sectional study. BMJ 2016;353:i2587. DOI: 10.1136/bmj.i2587.

12. Liu X, Cruz Rivera S, Moher D, Calvert MJ, Denniston AK. Reporting guidelines for clinical trial reports for interventions involving artificial intelligence: the CONSORT-AI extension. Nat Med 2020;26:1364-1374. DOI: 10.1038/s41591-020-1034-x.

13. Vokinger KN, Feuerriegel S, Kesselheim AS. Mitigating bias in machine learning for medicine. Commun Med (Lond) 2021;1:25. DOI: 10.1038/s43856-021-00028-w.

14. U.S. Food and Drug Administration. Artificial intelligence/machine learning (AI/ML)-based software as a medical device (SaMD). January 2021 (https://www.fda.gov/media/145022/download).

15. Cruz Rivera S, Liu X, Chan A-W, Denniston AK, Calvert MJ. Guidelines for clinical trial protocols for interventions involving artificial intelligence: the SPIRIT-AI extension. Nat Med 2020;26:1351-1363. DOI: 10.1038/s41591-020-1037-7.

16. Stanford University. Human-centered artificial intelligence. Artificial Intelligence Index Report 2022 (https://aiindex.stanford.edu/wp-content/uploads/2022/03/2022-AI-Index-Report_Master.pdf).

17. Savage N. The race to the top among the world's leaders in artificial intelligence. Nature 2020;588:S102-S104. DOI: 10.1038/d41586-020-03409-8.

18. Stanford University. Human-centered artificial intelligence. Artificial Intelligence Index Report 2023 (https://aiindex.stanford.edu/wp-content/uploads/2023/04/HAI_AI-Index-Report_2023.pdf).

19. Compton B, Barash DM, Farrington J, et al. Access to medical devices in low-income countries: addressing sustainability challenges in medical device donations. NAM Perspect 2018;8. DOI: 10.31478/201807a.

20. Howard FM, Dolezal J, Kochanny S, et al. The impact of site-specific digital histology signatures on deep learning model accuracy and bias. Nat Commun 2021;12:4423. DOI: 10.1038/s41467-021-24698-1.

Back to Table of Contents

NEJM AI

NEJM AI 2023; 1 (1)
DOI: 10.1056/AIcs2300066

CASE STUDY

High-Impact Medical Journals Reflect Negative Sentiment Toward Psychiatry

Roy H. Perlis, M.D., M.Sc.,[1,2] and David S. Jones, M.D., Ph.D.[2,3]

Received: August 3, 2023; Revised: September 29, 2023; Accepted: October 3, 2023; Published: November 9, 2023

Abstract

Psychiatry as a medical specialty has historically been considered less scientifically grounded than other disciplines. Despite progress in understanding the biological basis of psychiatric disease and efforts to diminish stigma associated with mental illness, a negative bias against psychiatry may persist in the medical community. Large language models provide an opportunity to investigate this hypothesis at scale. Our objective was to characterize the extent to which articles published in high-impact medical journals may reflect more negative sentiment about psychiatry compared with those from other medical specialties. We analyzed Entrez/PubMed entries published between 2017 and 2022 relating to psychiatry, neurology, oncology, and cardiology in four high-impact medical journals: *British Medical Journal*, *Journal of the American Medical Association*, *Lancet*, and *New England Journal of Medicine*. We used the large language model GPT-4 to score each article's title and abstract in terms of valence — that is, whether the abstract and title were likely to increase or decrease optimism about progress in a given medical specialty. Overall, and in each of the four journals, publications in psychiatry were significantly more likely to be negatively valenced than those of other specialties ($P<0.001$ for all omnibus χ^2 and post hoc pairwise contrasts). Negative-valence scores were found for 19.5% of publications relating to psychiatry, compared with 6.1% for cardiology, 6.1% for oncology, and 10.7% for neurology. Results were similar in analyses restricted to publications with abstracts, those reporting original research, those published before the Covid-19 pandemic, or those published in 2020 or later. In logistic regression models adjusting for journal, publication year, article type, and presence or absence of abstract, psychiatric publications were significantly more likely to be negatively valenced than all other specialties. Permuting the article specialty did not meaningfully change estimates of valence, indicating that the results were not attributable to psychiatry-specific terms. Published psychiatry articles were on average more likely than other specialty articles to reflect negative valence about the specialty. Whether this difference reflects volume and type of submission, or bias among editors or reviewers, merits further study. Regardless of the

The author affiliations are listed at the end of the article.

Dr. Perlis can be contacted at rperlis@mgh.harvard.edu or at Massachusetts General Hospital, 185 Cambridge Street, 6th Floor, Boston, MA 02114.

Back to Table of Contents

mechanism, the potential contribution of these articles to perpetuating negative attitudes toward psychiatry is also worthy of further investigation.

Introduction

Psychiatry has long faced a status problem in American medicine. Since the late 19th century, it lagged behind other areas of medicine in knowledge of the biological bases of disease, reflecting in part the challenges in studying an organ that is difficult to access during life and does not show gross disease-associated pathology after death.[1] These scientific challenges, coupled with negative portrayals in popular media, may have contributed to longstanding underinvestment in psychiatric care despite the high prevalence of mental illnesses and their substantial contributions to disability and mortality.[2]

The low prestige of psychiatry has been apparent in both lay and academic discourse. For example, survey studies of medical students decades ago reflected negative attitudes toward psychiatry.[3-6] Despite efforts to improve perceptions of psychiatry and individuals with mental illness, recent investigations suggest that some trainees still see psychiatry less positively than other areas of medicine.[7] A study published in 1981 suggested that such negative views reflect exposures during training.[6] Because many of the physicians who trained during this period now hold mentorship and leadership positions in Anglo-American health care institutions, medical schools, and medical journals, they are likely to continue to shape perceptions of the field.

We hypothesized that persistent bias against psychiatry exists in high-impact medical journals, which may help to perpetuate negative sentiments toward psychiatry within the medical community. To examine this potential bias, we drew on the ability of large language models to capture sentiment in more nuanced ways than older natural language processing tools, particularly in few-shot learning contexts,[8] and at a greater scale than human raters. We examined the extent to which articles published about psychiatry in high-impact journals might be more negative in overall presentation than those examining other areas of medicine.

Methods

We queried Entrez/PubMed for all entries that had publication dates between 2017 and 2022 in four high-impact medical journals (*British Medical Journal*, *Lancet*, *Journal of the American Medical Association*, and *New England Journal of Medicine*) and that included the MeSH terms Mental Disorders, Nervous System Diseases, Cardiovascular Disease, or Neoplasms or the words Psychiatry, Neurology, Cardiology, or Oncology. The three comparator specialties were selected a priori, the first reflecting the same primary site of illness as psychiatry, the second focusing on another primary site of illness, and the third unlinked to any individual site.

We used the GPT-4 large language model[9] (gpt-4-0613; GPT refers to generative pretrained transformer) to classify each Entrez/PubMed entry, i.e., the publication's abstract and title, according to whether a reader would feel more positive or more negative about a particular field after seeing that entry. Specifically, we prompted with the string "Score each paper in terms of how positive or negative a reader would feel about {specialty} as a field after reading the abstract or title. That is, how good or bad would they feel about progress in {specialty}. Rate on a 1–10 scale, where 1 is the most negative and 10 is the most positive." (Varying the wording of this prompt did not meaningfully change results.) To aid in interpreting these results, Supplement A in the Supplementary Appendix includes examples of each level of sentiment generated by GPT-4 with the prompt "Now, using this scale, generate an example of a title and abstract that would be scored a {number}." As an application of zero-shot learning in the absence of a gold standard, the model was not further trained to approximate a human rater; however, intraclass correlation coefficients between model scores and those from a blinded human rater (R.H.P.) for 100 publications in each specialty from 2019 (i.e., before the Covid-19 pandemic) were also calculated for descriptive purposes.

The primary analysis examined the proportion of entries reflecting negative sentiment (scores of 3 or less) within a given specialty in a given journal. Articles identified as reflecting multiple specialties were randomly assigned to one specialty for subsequent analysis; including these multispecialty articles in each category or excluding them altogether did not yield meaningfully different results. Sensitivity analyses limited assessments to articles with

abstracts (i.e., research articles and reviews), articles labeled as original research (either by Entrez/PubMed keyword or with the term "trial "or "study" indicated in the title or abstract), and articles published either before 2020 or in 2020 through 2022. Chi-square tests followed by post hoc contrasts were used to identify specialties differing significantly from others in frequency of negatively valenced articles. We then applied logistic regression models to examine the association between specialty and likelihood of negative valence, adjusted for journal, article type, presence or absence of an abstract, and year of publication. (In these regression models, sensitivity analyses with inclusion of clustering by journal did not meaningfully change results.) For purposes of comparison, the analyses were then repeated for those articles that were scored as positively valenced (scores of 7 or greater).

Finally, to confirm that results were not driven by negatively valenced terms, such as "depression" and "anxiety," we permuted titles and abstracts via GPT-4 with the prompt "Alter this {title/abstract} to refer to {new_specialty} instead of {old_specialty}," followed by the title or abstract, thus altering abstracts from psychiatry to neurology, cardiology, or oncology, and vice versa. We compared scores from the original abstracts to those from the permuted abstracts.

All analyses were completed between July 17 and August 1, 2023, using R version 4.3.[10]

Results

Characteristics of publications between 2017 and 2022 in the four selected journals are summarized in Table 1, including those addressing cardiology (n=3527), psychiatry (n=1843), oncology (n=2953), and neurology (n=2013). The intraclass correlation coefficient (a measure of consistency between two raters, in this case, the model output and the blinded rater) was 0.74 (95% confidence interval [CI], 0.69 to 0.78).

Overall, 360 (19.5%) of the publications relating to psychiatry were scored as negatively valenced, compared with 216 (6.1%) for cardiology, 181 (6.1%) for oncology, and 216 (10.7%) for neurology (omnibus χ^2 P<0.001). Distribution

Table 1. Characteristics of Publications Across Four Subject Areas in Four High-Impact Medical Journals, 2017 to 2022.

Publications	Cardiology (n=3527)	Neurology (n=2013)	Oncology (n=2953)	Psychiatry (n=1843)	Total (n=10,336)	P value
Article type						<0.001
Editorial or letter	1226 (34.8%)	554 (27.5%)	1032 (34.9%)	565 (30.7%)	3377 (32.7%)	
Original research	1003 (28.4%)	421 (20.9%)	654 (22.1%)	278 (15.1%)	2356 (22.8%)	
Other	1169 (33.1%)	970 (48.2%)	1179 (39.9%)	926 (50.2%)	4244 (41.1%)	
Review	129 (3.7%)	68 (3.4%)	88 (3.0%)	74 (4.0%)	359 (3.5%)	
Journal						<0.001
BMJ	443 (12.6%)	385 (19.1%)	422 (14.3%)	504 (27.3%)	1754 (17.0%)	
JAMA	915 (25.9%)	557 (27.7%)	629 (21.3%)	574 (31.1%)	2675 (25.9%)	
Lancet	764 (21.7%)	423 (21.0%)	573 (19.4%)	417 (22.6%)	2177 (21.1%)	
N Engl J Med	1405 (39.8%)	648 (32.2%)	1329 (45.0%)	348 (18.9%)	3730 (36.1%)	
Abstract present	1098 (31.1%)	539 (26.8%)	798 (27.0%)	425 (23.1%)	2860 (27.7%)	<0.001
Year						<0.001
2017	583 (16.5%)	387 (19.2%)	536 (18.2%)	420 (22.8%)	1926 (18.6%)	
2018	609 (17.3%)	363 (18.0%)	568 (19.2%)	318 (17.3%)	1858 (18.0%)	
2019	619 (17.6%)	351 (17.4%)	508 (17.2%)	341 (18.5%)	1819 (17.6%)	
2020	612 (17.4%)	288 (14.3%)	490 (16.6%)	255 (13.8%)	1645 (15.9%)	
2021	593 (16.8%)	294 (14.6%)	402 (13.6%)	244 (13.2%)	1533 (14.8%)	
2022	511 (14.5%)	330 (16.4%)	449 (15.2%)	265 (14.4%)	1555 (15.0%)	
Score, mean (SD)	5.6 (1.4)	5.4 (1.5)	5.8 (1.5)	5.1 (1.7)	5.5 (1.5)	<0.001
Negative valence*	216 (6.1%)	216 (10.7%)	181 (6.1%)	360 (19.5%)	973 (9.4%)	<0.001
Positive valence	986 (28.0%)	539 (26.8%)	1021 (34.6%)	414 (22.5%)	2960 (28.6%)	<0.001

* Negative valence, score of 3 or less; positive valence, score of 7 or greater. SD denotes standard deviation.

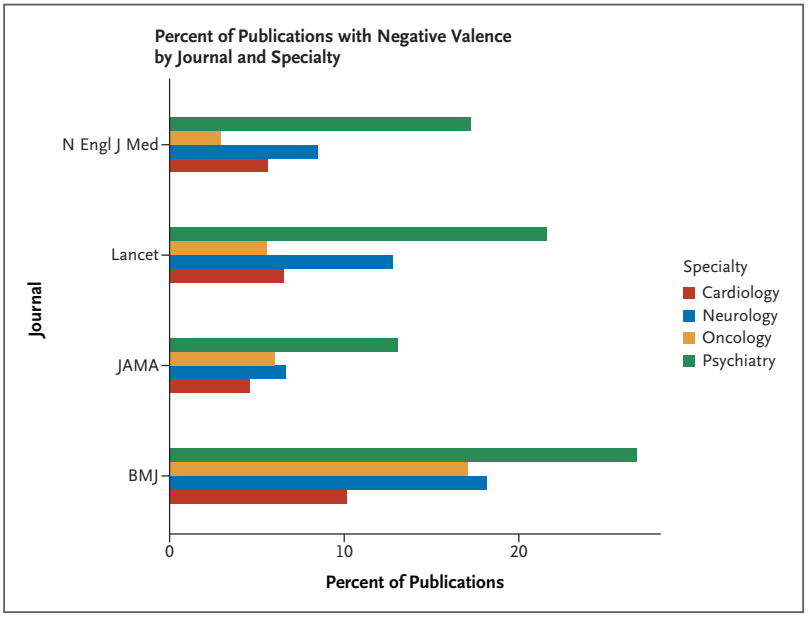


Figure 1. Percentage of Publications Reflecting Negative Valence, by Journal and Specialty, 2017 to 2022.

of scores by journal is illustrated in Figure S1. For 200 articles randomly chosen to be scored twice, the intraclass correlation coefficient was 0.85 (95% CI, 0.80 to 0.89).

In each of the four journals, the proportion of negatively valenced papers was significantly greater for psychiatry than for the other three specialties (Fig. 1; P<0.001 for all pairwise contrasts). Results in sensitivity analyses that limited publications to those with abstracts (Fig. S1A) or to those reporting original research (Fig. S1B) followed a similar pattern. Likewise, the proportion of negatively valenced papers relating to psychiatry was significantly greater than those of other specialties, whether the papers were published before or during the Covid-19 pandemic (Fig. S1C and D).

To examine the extent to which effects might arise from other differences between publications, we fit a logistic regression model to examine the association between specialty and negative valence, adjusting for journal, publication type, presence or absence of abstract, and publication year (Fig. 2). In adjusted models, articles from psychiatry were substantially and statistically significantly more likely to be negatively valenced than articles from cardiology (odds ratio, 2.94; 95% CI, 2.44 to 3.57), oncology (odds ratio, 3.13; 95% CI, 2.63 to 3.85), and neurology (odds ratio, 1.75; 95% CI, 1.45 to 2.13).

We then examined articles with positive valences (Fig. 3). Overall, psychiatric publications were statistically significantly less likely to be positively valenced than those from other specialties (Table 1; omnibus χ^2 P<0.001). However, in a logistic regression model adjusting for journal, year, article type, and abstract (Fig. S2), only oncology was statistically significantly different from psychiatry (odds ratio, 1.67; 95% CI, 1.46 to 1.93), whereas cardiology (odds ratio, 1.10; 95% CI, 0.96 to 1.27) and neurology (odds ratio, 1.14; 95% CI, 0.98 to 1.33) were not.

Finally, sensitivity analyses examined whether permuting titles and abstracts so that they referred to a different specialty affected the valence scores. Permuting psychiatry abstracts to any of the three other specialties did not yield significantly different valence scores (P=0.45; analysis in Supplement B). Similarly, permuting any of the other three specialties to psychiatry did not significantly alter valence scores (P=0.14, 0.64, and 0.25 for neurology, oncology, and cardiology, respectively).

Discussion

In this study of more than 12,000 publications in four high-impact medical journals between 2017 and 2022, we found that papers rated by a large language model to

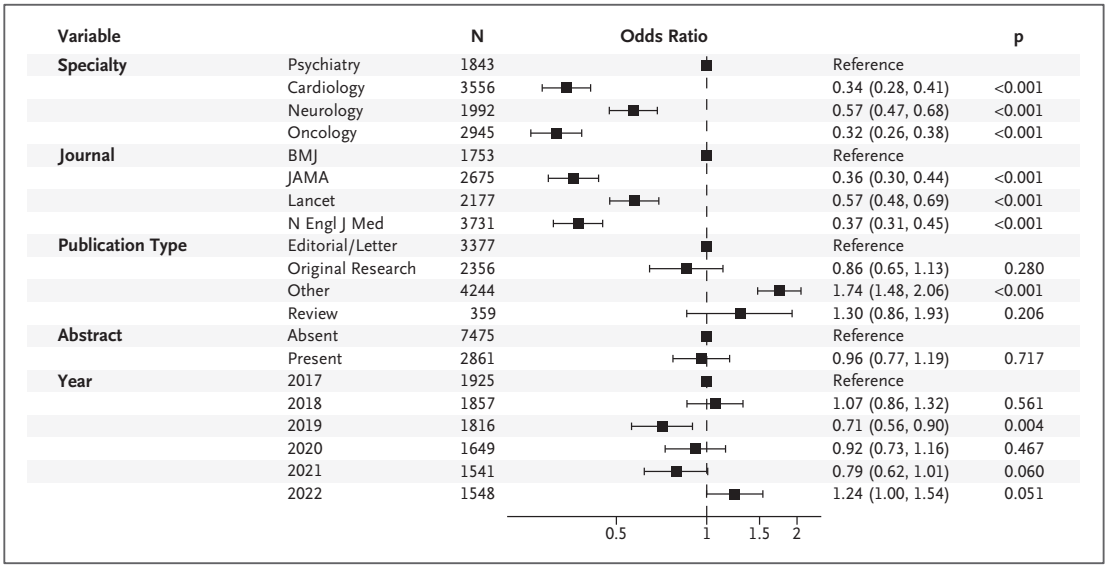


Figure 2. Logistic Regression Model of Negative Valence in Individual Publications Incorporating Terms for Specialty, Journal, Publication Type, Presence or Absence of Abstract, and Publication Year.

be negatively valenced (that is, reflecting sentiment of low optimism about the specialty) were overrepresented in psychiatry compared with three selected specialties. These differences were similar when analyses were limited to original research, papers with abstracts, or those published either before or during the Covid-19 pandemic period.

Although our results are difficult to compare with those from prior studies, our work is broadly similar in reflecting

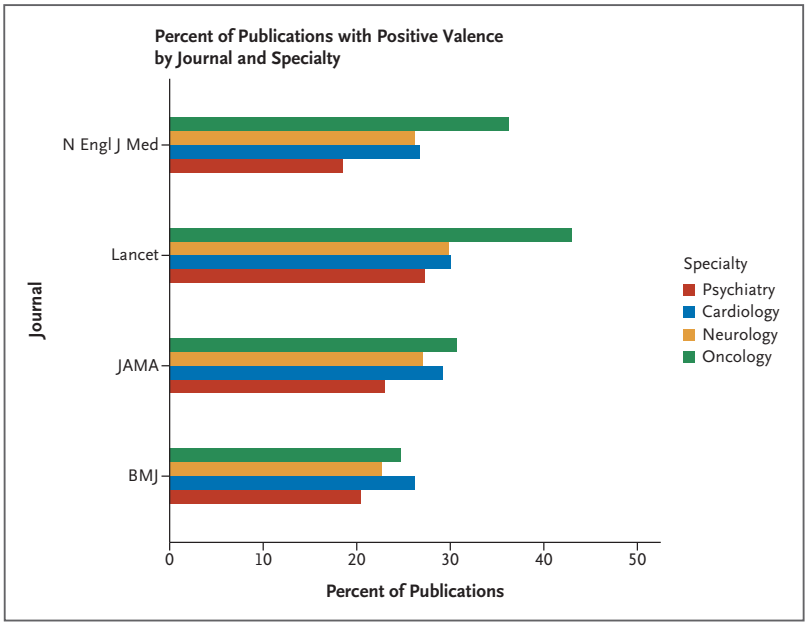


Figure 3. Percentage of Publications Reflecting Positive Valence, by Journal and Specialty, 2017 to 2022.

negative attitudes toward psychiatry. Such attitudes are apparent in surveys of medical trainees,[3-6] with even recent surveys reflecting negativity about psychiatry compared with other medical specialties.[7]

The low status of psychiatry has deep historical roots. When psychiatry emerged as a distinct practice in the 19th century, it took shape as a practice of asylum superintendents: physicians who worked in institutions on the outskirts of society, caring for people marginalized or ostracized by society. Moreover, as psychiatry and neurology divided the realm of mental disorders, psychiatrists were left with the diseases in which no lesion could be found,[11] a division that left psychiatric science facing difficult obstacles in the 20th century. Psychiatrists did not always help their cause: In an effort to develop treatments for mental illness with little understanding of the mechanism, they embarked on numerous therapeutic misadventures.[12,13] Psychiatrists sought to upgrade their status in the early 20th century by offering service to outpatients, a development that let them engage with all sectors of society[14] and move beyond asylums. Psychiatrists redoubled efforts to rehabilitate psychiatry's status after World War II by committing to basic and clinical research to upgrade the substance — and standing — of psychiatric science.[15] These efforts achieved considerable success, with both psychotherapy and psychopharmacology becoming widely accepted.[16] But psychiatry has continued to struggle to gain respect from other physicians, from patients, and within the public culture. *One Flew Over the Cuckoo's Nest* is the best-known critique, but it is just one of countless attacks on psychiatry that gained popularity in the 1960s and 1970s.[17] A vigorous antipsychiatry movement has persisted for decades[18]; at least before the Covid-19 pandemic, no other area of medicine had elicited a similar attack.

Given this complex history, it is likely that psychiatry's ongoing challenges are multifactorial. Therefore, our results do not necessarily reflect bias by journal editors or reviewers per se. For example, an alternative explanation could be that submissions of psychiatric manuscripts also reflect greater negativity — that is, authors are simply more likely to focus on negative aspects of psychiatric research in submissions to these journals. Regardless of mechanism, however, the results suggest that an individual reading high-impact medical journals and, possibly, subsequent media coverage of articles published in these journals would be left with a more negative impression of psychiatry than of the other three examined specialties. As such, the distribution of these articles may help to perpetuate negative attitudes toward psychiatry by medical practitioners and possibly in the broader community that may disproportionately receive health news reflecting articles in high-impact journals.

LIMITATIONS

We note several important limitations in interpreting this work. First, the quantification of sentiment relies on a large language model where the appearance of comprehension may be misleading.[19] The examples in Supplement A suggest that the model is applying human-interpretable standards insofar as it maintains explainability — that is, the scores appear to make sense. They also correlate with scores from a human rater. In sensitivity analyses, we show via permutation that the effects observed in psychiatry are not simply attributable to the presence of specialty terms associated with negative valence, such as depression or anxiety. However, in the absence of a true gold standard for this form of sentiment, which goes beyond the identification of positive or negative terms to predictions about the impact of a text on a reader, these scores must be interpreted cautiously.

Second, the costs associated with scoring algorithms via GPT-4 limited the number of specialties and journals examined in this pilot investigation. Future work on a larger scale will be valuable for more precisely estimating the effects we observed and determining their specificity.

Conclusions

Taken together, these findings support the hypothesis that high-impact medical journals continue to reflect a negative bias toward psychiatry as a medical specialty. In addition to ensuring better medical education about psychiatry and greater public efforts to diminish stigma, medical publications may represent an opportunity to achieve more balanced attitudes toward psychiatry, as psychiatrists seek to overcome a long and complicated history.

Disclosures

Author disclosures and other supplementary materials are available at ai.nejm.org.

Author Affiliations

[1] Massachusetts General Hospital, Boston
[2] Harvard Medical School, Boston
[3] Harvard University, Cambridge, MA

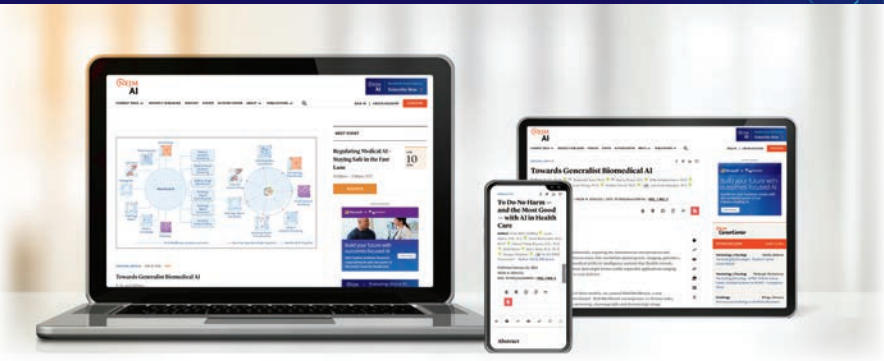
References

1. Harrison PJ. The neuropathology of schizophrenia. A critical review of the data and their interpretation. Brain 1999;122:593-624. DOI: 10.1093/brain/122.4.593.
2. GBD 2019 Mental Disorders Collaborators. Global, regional, and national burden of 12 mental disorders in 204 countries and territories, 1990–2019: a systematic analysis for the Global Burden of Disease Study 2019. Lancet Psychiatry 2022;9:137-150. DOI: 10.1016/S2215-0366(21)00395-3.
3. Wilkinson DG, Greer S, Toone BK. Medical students' attitudes to psychiatry. Psychol Med 1983;13:185-192. DOI: 10.1017/S0033291700050194.
4. Wiriyacosol P, Oon-Arom A, Suradom C, Wongpakaran N, Wongpakaran T. Medical students' attitude towards psychiatry: a comparison of past and present. Sci Rep 2023;13:8714. DOI: 10.1038/s41598-023-35797-y.
5. Rajagopal S, Rehill KS, Godfrey E. Psychiatry as a career choice compared with other specialties: a survey of medical students. Psychiatr Bull 2004;28:444-446. DOI: 10.1192/pb.28.12.444.
6. Crowder MK, Hollender MH. The medical student's choice of psychiatry as a career: a survey of one graduating class. Am J Psychiatry 1981;138:505-508. DOI: 10.1176/ajp.138.4.505.
7. Pascucci M, Stella E, La Montagna M, et al. Attitudes toward psychiatry and psychiatric patients in medical students: can real-world experiences reduce stigma? Eur Psychiatry 2016;33(Suppl 1):S218. DOI: 10.1016/j.eurpsy.2016.01.531.
8. Zhang W, Deng Y, Liu B, Pan SJ, Bing L. Sentiment analysis in the era of large language models: a reality check. May 24, 2023 (http://arxiv.org/abs/2305.15005). Preprint.
9. Bommineni VL, Bhagwagar S, Balcarcel D, Bommineni V, Davazitkos C, Boyer D. Performance of ChatGPT on the MCAT: the road to per-sonalized and equitable premedical learning. March 7, 2023 (https://www.medrxiv.org/content/10.1101/2023.03.05.23286533v1). Preprint.
10. R Core Team. R: a language and environment for statistical computing (https://www.r-project.org/).
11. Goetz CG, Bonduelle M, Gelfand G. Charcot: constructing neurology. Oxford: Oxford University Press, 1995.
12. Braslow J. Mental ills and bodily cures: psychiatric treatment in the first half of the twentieth century. Berkeley, CA: University of California Press, 1997.
13. Scull A. Madhouse: a tragic tale of megalomania and modern medicine. New Haven, CT: Yale University Press, 2005.
14. Lunbeck E. The psychiatric persuasion: knowledge, gender, and power in modern America. Princeton, NJ: Princeton University Press, 1994.
15. Pressman J. Last resort: psychosurgery and the limits of medicine. Cambridge, UK: Cambridge University Press, 1998.
16. Harrington A. Mind fixers: psychiatry's troubled search for the biology of mental illness. New York: W.W. Norton and Company, 2019.
17. Szasz T. The myth of mental illness: foundations of a theory of personal conduct. New York: Harper & Row, 1961.
18. Whitley R. The antipsychiatry movement: dead, diminishing, or developing? Psychiatr Serv 2012;63:1039-1041. DOI: 10.1176/appi.ps.201100484.
19. Birhane A, Kasirzadeh A, Leslie D, Wachter S. Science in the age of large language models. Nat Rev Phys 2023;5:277-280. DOI: 10.1038/s42254-023-00581-4.

Back to Table of Contents

NEJM AI

Deepen your understanding of how AI impacts your practice.



Subscribe Now to *NEJM AI*!

NEJM AI, the new digital journal from NEJM Group, empowers you with the evidence you need to stay ahead of AI.

Subscribe today for one year of unlimited access to 12 monthly digital issues packed with original research and insights into the deep issues at the intersection of AI and medicine, including:

- Original Research reports on clinical trials of AI or AI-assisted trial design, diagnosis, patient communications, and breakthrough medical AI applications.
- Review Articles that provide context for both clinical and technical readers.
- Dataset, Benchmarks, and Protocols: a common set of resources and tools to build and test new algorithms
- Diverse Perspectives from leading experts
- Plus a weekly newsletter, monthly AI Grand Rounds podcasts, and free virtual events twice a year

"AI is already changing medicine at an incredible pace. Now is the time for us to help people understand where that field is and where it's going.

No other clinical journal has made this their central focus."

– Isaac S. Kohane, MD, PhD, Editor-in-Chief, *NEJM AI*

Get the evidence you need to stay ahead of the fast-changing world of medical AI.

Subscribe today at **ai.nejm.org**

MONTHLY AND ANNUAL OPTIONS AVAILABLE

NEJM AI

NEJM Group appreciates the foresight and support of the exclusive sponsors of *NEJM AI:*