Written Testimony for the Record
Submitted to the
House Committee on Veteran's Affairs
Subcommittee on Health
For the Hearing
Artificial Intelligence at VA: Exploring its Current State and Future Possibilities

February 15, 2024

David E. Newman-Toker, MD PhD
Johns Hopkins University School of Medicine

Chairman Miller-Meeks, Ranking Member Brownley, and distinguished members of the Subcommittee, thank you for the opportunity to address Congress on this critically important topic of artificial intelligence (AI) in health care at the VA in support of our veterans. My name is David Newman-Toker, and I am a physician scientist with doctoral-level training in public health and a research focus on improving medical diagnosis, including the development and deployment of novel diagnostic technologies such as AI. I have been a faculty member at the Johns Hopkins University School of Medicine for more than two decades, where I currently serve as the David Robinson Professor of Vestibular Neurology and Director of our AHRQ-funded Center for Diagnostic Excellence. I am also a past President of the Society to Improve Diagnosis in Medicine. My testimony today will focus on opportunities and challenges for AI in health care from a public health perspective, with a special emphasis on AI to improve medical diagnosis. I have tailored my remarks to the VA context as appropriate, but I believe that what I will share here is broadly applicable to healthcare both within and outside the VA system.

I play two primary leadership roles at Johns Hopkins Medicine. First, I lead a clinical neurology unit whose main emphasis is on optimizing diagnosis and management of patients with dizziness and vertigo, including the development of innovative, scalable technology-based diagnostic solutions to improve care for the 18 million Americans seeking treatment for these symptoms each year, nationwide. Second, I serve as Director of the Armstrong Institute Center for Diagnostic Excellence, one of ten federally funded centers around the nation with a central focus on diagnostic safety and quality. Our team uses a mix of research methods and operations improvement techniques to understand, measure, and enhance diagnostic performance in pursuit of a vision of eliminating preventable harms from diagnostic error. Those of us in the field who are working to improve diagnostic accuracy and prevent patient harms appreciate the Members of Congress who, over the past several years, have worked to increase funding in this area.

The teams I lead include those with expertise in development, testing, deployment, and evaluation of technologies and tools under the umbrella of AI or machine learning (ML), such as deep learning and large language models. Our team has led early work on AI-based analysis of eye movements for stroke diagnosis and developed novel approaches to large-scale data mining to monitor harms from medical misdiagnosis. I am not a computer scientist or informatician, so my testimony focuses on opportunities and challenges for AI in health care from a public health perspective. In doing so, I will draw heavily on my training and experience in clinical care, research, and quality improvement focused on medical diagnosis.

I would like to state for the record that the opinions I express in both my written and oral testimony are my own and do not necessarily reflect those of The Johns Hopkins University or Johns Hopkins Medicine.

Overview of AI with an Emphasis on AI for Diagnosis

AI is the branch of computer science concerned with endowing computers with the ability to simulate intelligent human behavior.[1] The most complex cognitive task in medicine is the act of diagnosing the cause of a patient's symptoms. Errors in diagnosis account for an estimated 800,000 deaths or permanent disabilities each year in the US,[2] more than 80% of which are linked to cognitive errors or clinical reasoning failures. This creates a unique quality improvement opportunity for AI-based systems to save American lives at public health scale.

Potential benefits of AI include (1) better health outcomes for patients at lower costs, (2) greater access to and efficiency of care delivery, especially for those currently underserved and disadvantaged, and (3) decreased health care workforce burnout. However, none of these benefits will be realized without tackling foundational data challenges facing AI.

The rate-limiting step for developing and implementing AI systems in healthcare is no longer the technology—it is the sources of data on which the technology must be trained. There are multiple facets of the healthcare data quality problem which I address at greater length below. However, in plain language, they boil down to the problem of "garbage in, garbage out"—if we train AI systems on faulty data, we will get faulty results. AI systems that learn on faulty data will generally make the same mistakes that humans make, or worse. Put simply, if available electronic health record data sets are used to train AI systems, the **best we can hope for** is AI systems which replicate existing safety failures or implicit human biases and the **worst we can expect** is AI systems that are frequently wrong in their recommendations. If AI-based systems are deployed without adequate testing, the quality of health care will drop.

Foundational Data Challenges for AI in Health Care

The quality of AI technologies or tools is constrained by the quality of source data on which ML-based algorithms are trained. This includes data on both clinical inputs (e.g., baseline health state and disease risk factors, details of medical symptoms, relevant clinical examination findings, and laboratory or imaging test results) and care outputs (e.g., correct final diagnoses, treatments administered, patient health outcomes, and costs of care). Without high-quality data on inputs and outputs for training, AI predictions will be inaccurate, unreliable, or biased.

Data quality in health care is far from uniform, and data in support of AI for diagnosis often have a particularly shaky foundation. Although many blood tests or radiographic imaging studies are routinely obtained and recorded with extremely high fidelity, details of patient symptoms or clinical examination findings are often missing or incorrect in the electronic health record (especially for cases with delay or error in diagnosis or other failures in care quality[3,4]). Likewise, with care outputs, there are often high-quality digital data records of resource utilization (e.g., clinical visits, laboratory or imaging tests obtained) and patient deaths, while information about incorrect final diagnoses or disabling outcomes is often lacking or delayed. Final clinical diagnoses are found to be wrong at autopsy in 5-10% of hospital deaths.[5]

So, there are three fundamental data challenges for AI in health care: (a) **erroneous or biased data[6]** (data quality problems with source/training data sets, including false, unreliable, or demographically biased clinical data in electronic health records); (b) **"looking where the light is best"** (training AI systems on data based solely [or largely] on data availability, rather than value or utility to answer clinically relevant questions, and without regard to information bias in data quality or missingness); (c) **lack of routinely gathered health outcomes** (e.g., follow-up to determine accuracy of diagnoses, adverse events, disability, or costs of care).

Potential Risks and Pitfalls for AI in Health Care

Key potential risks and pitfalls include (a) implementation of AI without sufficient evaluation or monitoring (risking worse health outcomes or increased health care costs), (b) dehumanized and demographically biased health care (including racial bias[7,8]), or (c) clinical workforce deskilling,[9] resulting in a progressive decline in health care quality associated with inability of clinicians to practice absent AI systems or to fact-check AI system outputs.[10]

There is precedent for electronic systems to be implemented with the intention of improving quality or workflow efficiency without fully considering or monitoring for unintended consequences.[11,12] For example, "copy and paste" functions in electronic health records have improved workflow efficiency in some aspects of clinical documentation but also often reduce the accuracy and informativeness of such documentation, resulting in potentially serious adverse effects for patients, such as medical misdiagnosis.[13] The risks for AI systems may be worse, since AI systems will copy forward fundamental flaws from their training datasets indefinitely, creating a slightly inferior copy of everything currently wrong with healthcare.

There is a significant risk that apparent workflow efficiencies generated by AI systems will lead to widespread adoption of such systems without appropriate monitoring or evaluation. Imagine a "simple" AI system that uses large language models to automate searching through messages from patients to find those that represent medical emergencies; assume there are 100 messages and 5 represent true medical emergencies. If the AI system is perfect—i.e., identifies all 5 actual emergencies (no false negatives) and does not mislabel any non-emergencies (no false positives), then care quality will increase. However, no systems are perfect. If instead the system identifies all 5 emergencies (no false negatives) and overcalls 45 other patient messages (all false positives), then only 10% of the messages "flagged" by the system will be emergencies. These false alerts will cause alert fatigue, and busy clinicians facing burnout will likely stop paying attention to them. The harm resulting from such alert fatigue is well documented in the context of technology already in use.[14] This "overcalling" will lead developers to refine ("tune") the AI algorithm to produce fewer false positives. An unintended consequence will be that the system will then start to miss true emergencies. Imagine the system now identifies 2 emergencies (3 false negatives) and has just one false alert. Such a system would be readily adopted by overworked clinicians (2 of 3 alerts are "true positives" for medical emergencies) if they remained unaware that more than half (3 of 5) of the true emergencies were missed. Thus, without systematic monitoring and evaluation, workflow efficiency gains may lead to clinical adoption without recognizing that care quality for patients has declined. By way of example, a recent class action lawsuit contends that a low-accuracy AI algorithm deployed by an insurance company was used to systematically (and inappropriately) deny care to elderly patients.[15]

Key Role for the VA in Leading Development & Deployment of Trustworthy AI in Health Care

The VA has already constructed an elegant framework for Trustworthy AI that includes six core principles: (1) purposeful; (2) effective and safe; (3) secure and private; (4) fair and equitable; (5) transparent and explainable; and (6) accountable and monitored. From the vantage point of AI for diagnosis, pillars #2, #4, and #6 are especially mission critical and require special considerations to be executed. More specifically, we do not currently ensure that medical diagnosis (regardless of whether it is delivered by humans or AI) is effective and safe, fair and equitable, and accountable and monitored. Much of this boils down to a lack of data from key process steps: (1) details of the bedside diagnostic encounter, (2) follow-up on diagnostic error/accuracy/outcomes, and (3) feedback, learning and quality improvement mechanisms.

The VA healthcare data environment is better suited than most to delivering high quality data that might train AI. Key attributes include (a) the VA's commitment to healthcare quality and safety; (b) a large, national network of providers; (c) a unified health record offering greater potential for standardizing data capture; (d) independence from financial reimbursement-driven problems in health encounter documentation; and (e) addressing a patient population that tends to stay largely within the VA system, so outcomes can be better tracked over time. These attributes give the VA the opportunity to take a leading role in building high-quality AI systems.

For AI in health care to maximally benefit the health of all Americans, including veterans, the following are essential: (1) AI systems must be trained on gold-standard data sets that are unbiased and include complete information on both clinical inputs and care outputs; (2) AI systems must be effectively integrated into clinical workflows, leveraging the strengths of both computers and humans to produce a better result than could be achieved by either alone[16]; and (3) wherever AI is used, systems to monitor, maintain, and even enhance clinician skills should be co-deployed so that clinicians and AI systems will continue to "fact check" each other.

I have three primary recommendations for the Committee with regard to implementing AI at the VA, with an emphasis on diagnosis: (1) **the next decade must focus on constructing gold-standard data sets for diagnosis**—the promise of AI will not be realized without quantifying bedside evaluations; (2) **AI systems must be held to a high diagnostic standard**—they must be demonstrated scientifically to improve safety and quality over current care and then monitored closely over time; and (3) **the impact of AI on human clinical diagnostic skills must be monitored and managed**—clinical deployment of AI should be explicitly designed to enhance rather than reduce clinician skills by applying educational and human factors science.

Need for Gold-Standard Data Sets and Research to Support High-Quality AI-Based Diagnosis

Developing gold-standard data sets for "visual diagnosis" based solely on medical images (as in radiology, ophthalmology, and dermatology) is already well underway. However, comparable initiatives for the bulk of clinical medicine are either nascent or do not exist at all. Training diagnostically accurate AI systems requires high quality data at both the front end (patient demographics, symptoms, signs, and laboratory/radiographic findings) and back end (accurate final diagnoses, treatment effects, and morbid or mortal outcomes). The front-end inputs must be diagnostically relevant, digitally quantified, reliably captured, complete, and unbiased. This is not a simple task, because much of what we consider clinical medicine (e.g., patient comes for diagnostic evaluation of dizziness, headaches, abdominal pain, or fatigue) involves critical bedside history and physical examinations that are essential to proper diagnosis but meet few if any of the abovementioned criteria. The back-end outputs of diagnostic care including assessments of diagnostic accuracy and health outcomes must generally be disease-specific, meaningfully measured, reliably captured, complete, and unbiased. This is also not a simple task, since such outcomes are often not assessed at all in modern US healthcare.

As a result, significant investments in creating such data sets will be needed, and much of this will need to be undertaken explicitly as part of dedicated research studies to develop these data sets. Multiple federal agencies have begun to support research endeavors related to AI interventions in health care. However, not all areas relevant to the public health are equally addressed. For example, funding for the study of diagnostic errors substantially lags its public health burden,[17] with current funding in the range of $20-30 million per year for an issue that leads to death or permanent disability for an estimated 800,000 Americans annually,[2] translating to just $25-40 per year per serious patient harm and $50-80 per year per death attributable to misdiagnosis; by way of comparison, some diseases receive over $400,000 per year per death.

Key aspects of research resource allocation (e.g., AI, subdivided by diagnosis vs. treatment; AI, subdivided by clinical setting; AI, subdivided by disease) should be routinely tracked (e.g., via categorical spending lists[18]) and adjusted as necessary to match the public health impact. Special attention should be given to "prioritizing awards to improve health care data quality"[19] by deliberately funding programs that support development of large, gold-standard data sets from which high-quality AI systems can be trained.

Conclusion

AI has the potential to transform health care for the better by improving health outcomes, increasing access to and efficiency of care delivery, and reducing health disparities, particularly for the challenging area of medical diagnosis. However, absent dedicated efforts to develop gold-standard data sets to ensure effective and safe diagnostic AI systems combined with dedicated monitoring for diagnostic outcomes, risks of AI for diagnosis will dominate. Such risks include worse health outcomes, concretizing human biases in digital form, and a deskilled clinician workforce unable to know when AI systems are leading them or their patients astray. However, if dedicated resources are applied, the VA is uniquely positioned to help realize the potentially enormous opportunities for positive impact of diagnostic AI for veterans and the general public.

Thank you for this opportunity. I would be pleased to answer any questions you may have.

REFERENCES

1. Shortliffe EH, Cimino JJ. Biomedical informatics : computer applications in health care and biomedicine. 3rd ed. New York, NY: Springer; 2006.

2. Newman-Toker DE, Nassery N, Schaffer AC, Yu-Moe CW, Clemens GD, Wang Z, Zhu Y, Saber Tehrani AS, Fanai M, Hassoon A, Siegal D. Burden of serious harms from diagnostic error in the USA. BMJ Qual Saf. Published Online First: 17 July 2023. doi: 10.1136/bmjqs-2021-014130.

3. Newman-Toker DE. Charted records of dizzy patients suggest emergency physicians emphasize symptom quality in diagnostic assessment. Ann Emerg Med. 2007;50(2):204-5.

4. Schwartz A, Weiner SJ, Weaver F, Yudkowsky R, Sharma G, Binns-Calvey A, Preyss B, Jordan N. Uncharted territory: measuring costs of diagnostic errors outside the medical record. BMJ Qual Saf. 2012;21(11):918-24.

5. Shojania KG, Burton EC, McDonald KM, Goldman L. Changes in rates of autopsy-detected diagnostic errors over time: a systematic review. JAMA. 2003;289(21):2849-56.

6. Teno JM. Garbage in, Garbage out-Words of Caution on Big Data and Machine Learning in Medical Practice. JAMA Health Forum. 2023;4(2):e230397.

7. Obermeyer Z, Powers B, Vogeli C, Mullainathan S. Dissecting racial bias in an algorithm used to manage the health of populations. Science. 2019;366(6464):447-53.

8. Bhagwat AM, Ferryman KS, Gibbons JB. Mitigating algorithmic bias in opioid risk-score modeling to ensure equitable access to pain relief. Nat Med. 2023;29(4):769-70.

9. Aquino YSJ, Rogers WA, Braunack-Mayer A, Frazer H, Win KT, Houssami N, Degeling C, Semsarian C, Carter SM. Utopia versus dystopia: Professional perspectives on the impact of healthcare artificial intelligence on clinical roles and skills. Int J Med Inform. 2023;169:104903.

10. Lyell D, Coiera E. Automation bias and verification complexity: a systematic review. J Am Med Inform Assoc. 2017;24(2):423-31. PMCID: PMC7651899.

11. Powers EM, Shiffman RN, Melnick ER, Hickner A, Sharifi M. Efficacy and unintended consequences of hard-stop alerts in electronic health record systems: a systematic review. J Am Med Inform Assoc. 2018;25(11):1556-66. PMCID: PMC6915824.

12. Colicchio TK, Cimino JJ, Del Fiol G. Unintended consequences of nationwide electronic health record adoption: challenges and opportunities in the post-meaningful use era. J Med Internet Res. 2019;21(6):e13313. PMCID: PMC6682280.

13. Cheng CG, Wu DC, Lu JC, Yu CP, Lin HL, Wang MC, Cheng CA. Restricted use of copy and paste in electronic health records potentially improves healthcare quality. Medicine (Baltimore). 2022;101(4):e28644. PMCID: PMC8797538.

14. Ancker JS, Edwards A, Nosal S, Hauser D, Mauer E, Kaushal R, with the HI. Effects of workload, work complexity, and repeated alerts on alert fatigue in a clinical decision support system. BMC Med Inform Decis Mak. 2019;19(1):227. PMCID: PMC6859609.

15. Rosenblatt C. Risks Of AI In Healthcare Come To Light. Forbes. 2023.

16. Friedman CP. A "fundamental theorem" of biomedical informatics. Journal of the American Medical Informatics Association : JAMIA. 2009;16(2):169-70. PMCID: 2649317.

17. Saltzman AB, Keita M, Saber Tehrani AS, Hassoon A, Hough DE, Newman-Toker DE. US federal research funding on diagnostic error substantially lags its public health burden [abstract]. Diagnostic Error in Medicine 2017; October 8-10, 2017; Boston, MA.

18. Estimates of Funding for Various Research, Condition, and Disease Categories (RCDC). National Institutes of Health. Available from: https://report.nih.gov/funding/categorical-spending#/.

19. Harris LA, Jaikaran C. Highlights of the 2023 Executive Order on Artificial Intelligence for Congress. Congressional Research Service; 2023. Available from: https://crsreports.congress.gov/product/pdf/R/R47843.