Testimony for the Record
Submitted to the
House Committee on Energy & Commerce
Health Subcommittee
For the Hearing
AI in Health Care: Exploring the Opportunities and Challenges

November 29, 2023

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

Chairman Guthrie, Ranking Member Eshoo, and members of the Committee, thank you for the opportunity to address Congress on this critically important topic of artificial intelligence in health care. 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. I have been a faculty member at the Johns Hopkins University School of Medicine for more than two decades, where I am the David Robinson Professor in Vestibular Neurology. I hold joint appointments in Health Sciences Informatics and, at the Bloomberg School of Public Health, in both Epidemiology and Health Policy & Management.

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 artificial intelligence (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 herein are my own and do not necessarily reflect those of The Johns Hopkins University or The Johns Hopkins Health System.

Overview

AI is the branch of computer science concerned with endowing computers with the ability to simulate intelligent human behavior.[1] Like many health care professionals, I see the potential for both promise and peril in the rapidly evolving field of AI for health care. My overarching policy recommendation to the Committee is that oversight for the development, testing, deployment, and ongoing monitoring and evaluation of AI technologies in health care must embrace a public health perspective by considering both the greatest benefits and greatest risks of AI at the population level as well as the potential for unfair distribution of impact across groups of individuals. I believe that the largest potential for both benefit and risk to public health

will come from AI interventions targeting health care sectors facing large gaps in care quality and high volumes: (1) service type—medical diagnosis; (2) care settings—outside the hospital; and (3) patient populations—those who are medically underserved or demographically disadvantaged. This focus, in turn, has important health care policy implications for federal regulatory oversight, clinical payment policy, and research resource allocation, which must each include a specific emphasis on these three dimensions maximally impacting public health.

Potential Benefits of AI in Health Care

Potential benefits of AI in health care include (a) better patient health outcomes, (b) greater access to and efficiency of care delivery, especially for those currently underserved and disadvantaged, and (c) decreased health care workforce burnout. The potential for AI in improving diagnosis is of special interest from a public health perspective. Recent work published by our team suggests that 800,000 Americans die or are permanently disabled each year because of errors in diagnosis among patients with serious, life- or limb-threatening diseases.[2] Thus, AI has tremendous potential for improving public health via that component of health care alone. However, none of these benefits will be realized without tackling foundational data challenges facing AI in health care, which are especially profound for medical diagnosis.

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. For example, with clinical inputs, many blood tests or radiographic imaging studies are routinely obtained with extremely high fidelity, whereas 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) 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] AI systems that learn on faulty data will generally make the same mistakes that humans make.

So, there are three fundamental data challenges for AI in health care: (a) **"garbage in, garbage out"[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).

Put simply, if available electronic health record data sets are used to train AI systems, the best we can hope for is AI systems that replicate and formalize implicit human biases and the worst we can expect is AI systems that are frequently wrong in their recommendations. Such lowering of quality standards in health care is not a desirable outcome of AI-based interventions.

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]

Success Factors for AI in Health Care

In order for AI in health care to maximally benefit the health of all Americans, the following three conditions are essential prerequisites: (1) AI systems must be trained on large, representative, well-curated, properly validated, gold-standard data sets that include complete information on both clinical inputs and care outputs; (2) AI systems must be effectively integrated into clinical workflows, leveraging the strengths of computers and humans together 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 (including diagnostic ones) should be co-deployed so that clinicians and AI systems will continue to "fact check" each other.

The Public Health Imperative for AI in Health Care

The development, testing, deployment, and ongoing monitoring and evaluation of AI technologies in health care should embrace the public health perspective. This means seeking the greatest benefits and minimizing the greatest risks while ensuring health equity. Below I consider three important dimensions of health care: (1) type of health care service (e.g., prognostication for treatment vs. diagnosis); (2) location of service (e.g., inpatient, ambulatory, in-home); and (3) patient population served (e.g., underserved or disadvantaged vs. not). For each, the public health perspective dictates a focus on areas of health care with gaps in care quality and high volumes.

Type of health care service: AI algorithms that assist with disease-specific risk prediction (e.g., risk of a second heart attack after a first or long-term risk of developing dementia) to assign treatments will be easier to develop than those for diagnosis of symptomatic disease. However, as noted earlier, the potential benefit of improving diagnosis is substantial, with an estimated 800,000 deaths and permanent disabilities resulting from diagnostic failures each year in the US.[2] However, because the task for AI systems assisting in diagnosis is more challenging (because of greater degrees of medical uncertainty, lower-quality clinical data relevant to the task, and higher complexity of externally verifying the accuracy of automated diagnosis[10]), the risks of using AI systems for medical diagnosis are also substantially greater.

Location of service: AI systems will be more closely managed and monitored in hospital (inpatient) and emergency department care settings. However, it is likely that the impact of AI systems on the public health will be greater in ambulatory (clinic-based) care or in-home direct-to-patient AI. For example, AI-based systems using large language models might provide what appears to be definitive medical advice for patients about cancer treatment[17]; however, such information may be both definitive-sounding and false, including "hallucinated" treatment

options.[17] There has been a recent proliferation of internet-based symptom checkers that help patients determine whether they should access the health care system. The accuracy of these symptom checkers is low, not improving over time, and generally does not exceed the accuracy of laypersons in determining the presence of an emergency or sufficiency of self-care.[18] Legal disclaimers on such systems generally assert that these tools are *not* providing medical advice (and that users should contact a medical professional if they need actual medical advice), but individuals may treat the feedback as medical advice in a way that places them at risk of delays in care for serious illnesses (false negatives) or psychological and financial harms from over-alerting (false positives).[19] AI-based systems may appear to be even more definitive, causing patients to lend results more credence than merited based on accuracy of recommendations. There is currently little or no oversight of such direct-to-consumer health advice systems.[19]

Patient population served: Privileged patients with greater economic stability, higher health literacy, greater access to health care providers, and therefore greater representation in training data may be better positioned to be modeled accurately in AI, use AI effectively to their benefit, and navigate potential shortcomings of AI-based systems. Those in rural or underserved communities or those with social determinants of health associated with generally worse health outcomes (e.g., economic instability, low health literacy, or marginalized demographic groups) may be least likely to benefit from innovations and most susceptible to suffering adverse consequences of inadequately regulated AI systems (particularly those that involve direct-to-consumer recommendations regarding diagnosis or treatment). We have already begun to see evidence emerging of built-in racial bias of various AI-based commercial tools which create "risk scores" that can dictate access to care management programs[7] or medical treatments.[8]

Policy Implications for AI in Health Care

Adopting a public health lens for policy recommendations means aligning regulatory oversight, clinical payment policy, and research resource allocation with the areas where the greatest positive public health impact can be realized. I believe that the greatest potential for benefit (and risk) to overall public health will come from AI interventions focused on medical diagnosis; those in less intensively monitored care settings (i.e., outside the hospital); and those among patients who are medically underserved or demographically disadvantaged. Without adequate standards, incentives, and monitoring of outcomes, public health benefits will not be realized, and there is a risk of significant negative impact on the health of the American people.

Regulatory oversight and standards: The US Food and Drug Administration (FDA) has a standard for pharmacological or device-based treatments that they must be demonstrated in rigorous research studies to improve patient health outcomes and proven safe for administration prior to being approved for routine clinical use. It is crucial that AI systems be held to a similar standard—they must be demonstrated scientifically to improve care quality over current care (or, at a minimum, to hold care quality constant while improving efficiency/cost or satisfaction). If AI systems are to be deployed in clinical practice prior to such FDA approval, this deployment should only be permitted in the context of research or quality improvement initiatives that incorporate careful monitoring, evaluation, and reporting of important health outcomes.

Oversight should not be limited to hospital settings and should extend to clinic-based care as well as direct-to-patient tools that offer AI-based medical advice. Such oversight should include careful monitoring and evaluation for systematic racial or other demographic biases in AI systems. It must also include longer-term post-marketing surveillance of AI systems that are put in place to identify low-frequency (yet still medically significant) risks, as is done for drug

treatments. The evolutionary nature of "continuously learning" AI systems means that higher levels of long-term vigilance will likely be needed to ensure continuous improvement rather than worsening over time. New policies and procedures will need to be developed that address how evolving systems can be version controlled and what demonstrated benefits should be measured and described in release notes. Standards for how best to balance rigor of oversight against the need for nimble adaptation of AI algorithms have yet to be developed. Methodological and policy research is needed to address these and similar issues in regulatory oversight.

Regulatory oversight should articulate specific standards for AI-based diagnostic tools to demonstrate some combination of increased diagnostic accuracy, reduced diagnostic errors, or improved health outcomes for patients (rather than merely meeting the lesser standard typically applied by FDA to routine diagnostic tests, which need only demonstrate safety and reliable measurement of a specific parameter). Absent such monitoring and evaluation, it would be nearly impossible to detect a substantial reduction in diagnostic accuracy (e.g., from 10% to 20%) following implementation of AI-based diagnostic systems. For diagnostic technologies that may be sensitive to regional variation in disease distribution or population demographics, it is not yet clear how oversight and approval of locally "tailored" versions should be performed. Again, this gap represents an issue in need of further methodological and policy research.

Payment policy and incentives: The Centers for Medicare and Medicaid Services (CMS) has numerous incentives designed to improve care quality while reducing the costs of health care. New incentives will likely be needed around implementation of AI-based systems to ensure that health outcomes are being monitored for improvement and systems are unbiased with respect to patient demographics or social determinants of health. Furthermore, special incentives may be required for AI-based diagnostic tools, since diagnosis is generally unaffected by

(current) disease-based incentive mechanisms (e.g., bundled payments to promote guideline-concordant and resource-efficient treatment for a specific disease, without consideration of whether disease cases were correctly diagnosed or not[20]). For example, it is possible to construct and test a symptom-based payment model ("symptom-related group" [SRG] by analogy to the existing disease-based "diagnostic-related group" [DRG]). In such a model a health care provider organization might be paid a fixed price for diagnostically evaluating (e.g., via AI) a patient with a specific symptom (e.g., dizziness) but not be paid for the consequences of a misdiagnosis (e.g., missed stroke that leads to a hospitalization within 30 days, using an established quality measure[21]). The organization should theoretically then be incentivized (due to shared savings) to offer more efficient diagnostic processes, while increasing the quality of diagnosis (due to penalty costs if harms accrue to patients). Thus, this arrangement ought to incentivize organizations to implement AI-based diagnostic tools only if they help patients.

Research resource allocation: 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,[22] 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. The Agency for Healthcare Research and Quality (AHRQ) has unique strengths relative to systems engineering for health care, so is particularly well suited to receive targeted research funds focused on system-level evaluation of the impact of AI tools on health care quality and safety; furthermore, AHRQ is optimally positioned to monitor specific

impacts of AI on diagnostic safety and quality, given its leading role and expertise on this topic.

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[23]) and adjusted as necessary to match the public health impact. Special attention should be given to "prioritizing awards to improve health care data quality"[24] by deliberately funding programs that support development of large, gold-standard data sets from which high-quality AI systems can be trained; creating resources of this nature, which may require substantial resources and academic-industry partnerships, might be well suited to funding via the Advanced Research Projects Agency for Health (ARPA-H). Also, investment in testing environments that permit safe and well-monitored implementations is key.

Conclusion

In summary, AI has the potential to transform health care for the better by improving health outcomes, increasing access to and efficiency of care delivery, reducing health disparities, and decreasing clinician workforce burnout. However, absent carefully crafted regulations, innovative payment incentives, and new research resources directed to overcome key barriers to successful deployment of high-quality AI systems, risks 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. The guiding principle for policy changes should be public health impact, including an emphasis on the equitable distribution of benefits and risks across the population. AI to improve medical diagnosis poses significant risks but also presents uniquely large opportunities for positive impact.

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. Chen S, Kann BH, Foote MB, Aerts H, Savova GK, Mak RH, Bitterman DS. Use of artificial intelligence chatbots for cancer treatment information. JAMA Oncol. 2023;9(10):1459-62. PMCID: PMC10450584.

18. Schmieding ML, Kopka M, Schmidt K, Schulz-Niethammer S, Balzer F, Feufel MA. Triage accuracy of symptom checker apps: 5-year follow-up evaluation. J Med Internet Res. 2022;24(5):e31810. PMCID: PMC9131144.

19. Muller R, Klemmt M, Ehni HJ, Henking T, Kuhnmunch A, Preiser C, Koch R, Ranisch R. Ethical, legal, and social aspects of symptom checker applications: a scoping review. Med Health Care Philos. 2022;25(4):737-55. PMCID: PMC9613552.

20. Pronovost PJ, Miller J, Newman-Toker DE, Ishii L, Wu AW. We should measure what matters in bundled payment programs. Ann Intern Med. 2018;168(10):735-6.

21. Austin JM, Newman-Toker DE. Avoid Hospitalization After Release with a Misdiagnosis—ED Stroke/Dizziness (Avoid H.A.R.M.—ED Stroke/Dizziness) (3746). Partnership for Quality Measurement: Battelle; 2023. Available from: https://p4qm.org/endorsements/measure/6031#comments.

22. 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.

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

24. 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.