Course Corrections for Clinical AI

Alex J. DeGrave 1,2 Joseph D. Janizek 1,2 and Su-In Lee 1

Introduction
The translation of artificial intelligence (AI) from a promising technology to a routine clinical tool is already well underway, but its benefits to patients and their physicians are not assured, requiring additional guidance from clinicians and industry regulators. In this perspective article, we review the evidence that clinical AI benefits or harms patients and their medical providers. We argue that although clinical AI has already created positive change in medicine, stakeholders such as physicians, hospital administrators, and regulators must demand higher standards of evidence, improved auditing capabilities, and greater attention toward workers' quality of life if we are to maximize this technology's potential.

Recently, a boom in AI in medicine has occurred, driven by modern machine learning (ML) methods, whose ability to automatically learn patterns from large amounts of data has enabled them to widely surpass earlier AI systems that required more manual programming by human experts. Although the idiosyncrasies of ML are occasionally relevant to clinicians, in this article we more broadly define AI to include both ML and other computational approaches that automate intellectual tasks typically reserved for human intelligence. In this way, we consider AI from an operational perspective, namely, in terms of its effect on health care workers and patients. For any type of medical AI system, the salient clinical questions remain the same: how does AI effect patients? How does it affect physicians? And what should we do to make it better?

How Does AI Affect Patients?
AI affects patients by informing diagnosis or treatment decisions through the processing of medical images, physiologic measurements, or health records; these technologies, known as "clinical decision support systems," may decrease cost and improve patient outcomes. Dozens of AI systems have been approved by the United States Food and Drug Administration (FDA), the majority of which analyze radiologic images, whereas the next most common subset monitors cardiac function. Miscellaneous other systems analyze other medical images, monitor other physiologic signals, act as digital assistants, or provide automated therapy (1).

To assess the potential benefits of these systems, we reviewed publicly available FDA application summaries and peer-reviewed publications and found a few devices show benefit in prospective clinical trials. In evaluating these devices, attention to high-quality evidence, namely, prospective data, ideally from controlled trials and concerning clinically meaningful end points, is critical, because the de facto standard for assessment of AI systems in the ML community is evaluation of predictive performance (e.g., accuracy at disease classification) on retrospective data, which may overestimate real-world performance and, more importantly, disregards patient outcomes. In a success case, a prospective clinical trial of intracranial hemorrhage detection software that triages computed tomography scans demonstrated reduced time to diagnosis of intracranial hemorrhage (2), and a number of similar software devices for triage in radiology have since come to market. Similarly, a randomized clinical trial of 68 patients in a single center demonstrated that an AI-based early warning system for hypotension events reduced the frequency and severity of hypotensive episodes during surgery (3). A small prospective trial of three nurses with 30 patients each evidenced that an AI sonography assistant can enable nurses who were not trained in sonography to acquire satisfactory ultrasound images (4), which could perhaps improve access to this imaging modality. Finally, in a randomized clinical trial of 142 patients, an ML-based sepsis prediction tool reduced hospital length of stay and mortality (5); although larger, multicenter trials will be necessary to confirm the generalizability of this result, this study provides rare evidence of meaningful improvements in patient outcomes. Although the ultimate effect of the remainder of these systems on patient outcomes remains to be seen (e.g., does the faster time to stroke diagnosis correspond to lower National Institutes of Health stroke scale scores?), these trials, which examine not just the predictive accuracy of AI systems but rather more meaningful clinical end points, support that AI is already offering improvements to medicine.

However, despite the promise of clinical decision support systems, high-quality empirical evidence for patient benefit is scarce. Most AI systems are approved via Section 510(k) of the Food, Drug and Cosmetic Act (1), which requires only that the device

1 Paul G. Allen School of Computer Science and Engineering, University of Washington, Seattle, Washington
2 Medical Scientist Training Program, University of Washington, Seattle, Washington

Correspondence: Prof. Su-In Lee, Paul G. Allen School of Computer Science and Engineering, University of Washington, 185 E. Stevens Way NE, Room 536, Seattle, WA 98195-2350. Email: suinlee@cs.washington.edu
alerts were associated with increased mortality (16), casting doubt on ML-based alert systems for AKI that are under development (17,18). Although the simple, formula-based alert from this clinical trial lacks the predictive abilities of modern AI systems to anticipate future AKI, plausible mediators of the observed increase in mortality, including alert fatigue or increased prescription of nonindicated interventions (16), may persist regardless of a shift toward more technically complex AI. These counterproductive instances of AI reinforce that health care providers and hospital administrators should consider not purchasing or using a clinical AI system outside clinical trials until high-quality evidence of benefit is available, and likewise they should consider taking part in clinical trials of promising AI to help improve the availability of evidence.

**How Does AI Affect Physicians?**

Predictions of the effect of AI on physicians’ work vary widely, ranging from the assertion that radiologists no longer need to be trained to the more optimistic forecast that AI will make all physicians more efficient and productive, rather than replacing them (19). Although our review of the literature identified no technology sufficiently advanced to supplant any medical provider in the near future, we also remain skeptical that AI’s incorporation will improve physicians’ work experience in the absence of attentive guidance toward this goal.

One area where AI is poised to influence medical providers’ work experiences is the optimization of scheduling. For instance, AI systems better predict the duration of surgical procedures than standard techniques (20), implying they could improve operating room scheduling (21). However, published evaluations of automated scheduling systems frequently retrospectively measure surrogate outcomes, such as the distribution of allocated shifts, rather than prospectively measuring physician satisfaction (22,23). Furthermore, a metareview of interventions to reduce physician burnout demonstrated that schedule adjustments had a mixed effect on physician stress and job satisfaction (24). Given that workload and scheduling are reliable predictors of emotional exhaustion in physicians (25), AI scheduling systems could improve or deteriorate medical providers’ work experience according to the extent they reflect worker preferences.

Clinical AI systems may also indirectly affect scheduling by altering the rates of medical procedures. For instance, an AI system that diagnoses a disease more frequently than physicians (26) may increase the rate of a corresponding intervention, changing the demands on physicians. If not addressed with appropriate staffing changes, such an increase in diagnoses, whether these diagnoses are correct or overdiagnoses, may contribute to physician burnout (27).

Beyond effects on scheduling and procedure rate, which directly modify when and how much medical providers work, details of any clinical AI’s implementation may contribute to improvements or declines in the work experience, as evidenced by existing instances of medical automation. The use of electronic health record systems is associated with lower physician satisfaction and increased burnout (28), whereas improved electronic health record usability mitigates burnout (29), implying the effects of AI on
medical providers’ work experiences may similarly hinge not on the goal of an AI system, but rather its usability. AI voice recognition services may decrease costs relative to traditional dictation services but shift burdens from medical transcriptionists to physicians (30), illustrating the organizational tradeoffs in implementing an AI system may also encumber clinicians. Automated dispensing systems in pharmacies may decrease job stress of pharmacy workers overall but elicit negative responses from pharmacy technicians who work most directly with these systems (31), illustrating how AI systems too could deteriorate the work experience of employees when the burdens of a system are shared unequally. These preliminary studies on the likely effect of AI on physician practice have strengthened the concern that AI systems may increase technological fatigue or inefficiently shift occupational burdens. Examining this issue on a more systematic basis, a recent review of AI systems for medical imaging found that although AI had potential for patient benefit, the vast majority of AI studies (>86%) were likely to increase physician workload due to an increase in both postprocessing and interpretation time (32). Therefore, medical providers should consider an active role in guiding the introduction of these technologies to encourage an improved working environment.

**What Changes Should Be Made To Maximize the Positive Effect of AI?**

Clinical decision support systems that are under development frequently neglect clinical needs, and development of tools to address the most pressing clinical issues may require closer collaboration between medical professionals and AI researchers. For instance, an AI system for detecting skin cancer from photographs of a patient’s skin was hailed by its developers for achieving “dermatologist-level” performance (33), but the absence of a commercial version of this system four years later suggests that accurate predictions may not equate to actionable information when bereft of a well-defined clinical role. Perhaps such an AI system could aid in screening, but crucial questions remain: who should be screened, and when? How should physicians incorporate additional information, such as lesion history and patient risk factors? What specific actions should a physician recommend in response to a positive screen? Ideally, translational questions such as these must be considered from the outset of the design process, when AI experts and clinicians collaborate to find the intersection of the technically feasible and medically useful. In another example, an AI-based screening system for diabetic retinopathy was deployed as part of a clinical trial, but patients opted out of the trial or did not follow up on the AI system’s recommendations, due to lack of transportation (12). Thus, even when a tool’s clinical role is well defined, neglect of patient factors may impair its utility. Given the increasing ease with which highly accurate ML systems are developed, these translational challenges may now be the bottleneck in the utility of a system.

Another opportunity for positive change is the improvement of safeguards that address the frequently “brittle” performance of modern AI that is high in some scenarios but unexpectedly and precipitously lower in others, which may threaten its utility or pose dangers to patients. Much of modern AI relies on ML, in which algorithms automatically infer patterns from large amounts of data, implying the quality of the system depends on the quality of the data. Oftentimes, performance is measured on a “held-out” portion of the training dataset, which exaggerates performance relative to prospectively collected test data, because ML is prone to “overfitting” and memorizing peculiarities of the training data that do not persist in prospective data (34). For instance, multiple studies developing ML-based risk scores for AKI and for CKD progression only examine the score’s performance on held-out data from the same source as the training data (18,35,36), such that we would expect the true performance on data from an external (or prospective) source to fall considerably short of the reported values. Frequently, data may contain spurious associations that enable the AI system to learn “shortcuts” (34): for example, AI may think a patient is sick simply because they come from a hospital where more patients are ill (37,38), or identify a pneumothorax on the basis of the chest tube that was placed to fix it (39). Even if AI learns genuine, rather than spurious, associations, dataset shift between the population of patients used for training the algorithm and the population of patients for whom the tool is used in practice may degrade the model’s performance (40). Training data may also poorly represent some subgroups, leading to inequitable performance. For instance, audits of commercial facial recognition classifiers revealed poor performance in individuals with darker skin (41), and longstanding disparities in melanoma care among patients with darker skin combined with high potential for racially biased training data have heightened concerns that similar biases may arise in clinical AI systems (42). To safeguard against these idiosyncratic failure modes, regulatory bodies might require prospective evaluation or retrospective evaluation on “external” data, and principled subgroup analyses to avert reliance on shortcuts and ensure equitable performance.

Compounding the “brittleness” problem of AI, these systems may be challenging to audit, given their complex “black-box” nature, their frequently proprietary status, and the poor availability of training data. In principle, an understanding of an AI system’s decision process may enable users or auditors to predict failure modes, but many modern AI systems are far too complex to understand without specialized tools. Fortunately, numerous tools to understand complex AI systems, which are known as “explainable AI,” have been developed. These techniques range widely and include highlighting the important parts of an image, identifying which of a patient’s variables contribute most to a prediction, displaying similar cases from the training data, and showing altered, “counterfactual” versions of the input that would have produced different predictions. Explainable AI has been useful for auditing medical ML models, for example revealing that some medical imaging classifiers take the “shortcuts” mentioned above. Our review of FDA approvals identified infrequent use of explainable AI, and more widespread use before deployment may help safeguard against these failures.

Although difficulty understanding the complex decision process of AI systems may be mitigated with explainable AI, other factors that limit auditing of AI elude
Outlook

Because AI methodologies are now maturing sufficiently to translate into the field of medicine, medical providers have a unique opportunity to shape their future. AI is a powerful tool for achieving accurate predictions, but to improve the lives of patients and their providers, clinical expertise is required to guide this technology toward the most exigent applications. Clinicians may involve themselves at each step of the process, by (1) collaborating with AI developers, (2) participating in clinical trials of promising AI, (3) critically evaluating the evidence behind AI devices before use, and (4) advocating for their patients and themselves at the administrative and regulatory levels. With thoughtful guidance, AI may transform medicine for the better.

Disclosures

All authors have nothing to disclose.

Funding

This work was funded by the National Science Foundation (CAREER DBI-1552309 to S.-I. Lee) and the National Institutes of Health (R35 GM 128638 and R01 AG061132 to S.-I. Lee).

Acknowledgments

The content of this article reflects the personal experience and views of the author and should not be considered medical advice or recommendations. The content does not reflect the views or opinions of the American Society of Nephrology (ASN) or Kidney360. Responsibility for the information and views expressed herein lies entirely with the author(s).

Author Contributions

A. DeGrave and J. Janizek conceptualized the study and wrote the original draft; S.-I. Lee was responsible for funding acquisition, project administration, and provided supervision; and all authors reviewed and edited the manuscript.

References

1. Benjamens S, Dhunnoo P, Meskó B: The state of artificial intelligence-based FDA-approved medical devices and algorithms: An online database. Digit Med 3: 118, 2020
2. Arbabsirhani MR, Fornwalt BK, Mongelluzzo GJ, Suever JD, Geise BD, Patel AA, Moore GJ: Advanced machine learning in action: Identification of intracranial hemorrhage on computed tomography scans of the head with clinical workflow integration. Digit Med 1: 9, 2018 https://doi.org/10.1038/s41746-017-0015-z
3. Wijnberge M, Beerts BF, Hol L, Lemmers N, Mulder MP, Berge P, Schenk J, Terwindt LE, Holfmann MW, Vlaar AP, Veelo DP: Effect of a machine learning-derived early warning system for intraoperative hypotension vs standard care on depth and duration of intraoperative hypotension during elective noncardiac surgery. JAMA 323: 1052–1060, 2020 https://doi.org/10.1001/jama.2020.0592
4. US Food and Drug Administration Center for Devices and Radiological Health: De novo classification request for Caption Guidance. Available at: https://www.accessdata.fda.gov/cdrh_docs/reviews/DEN190040.pdf. Accessed April 15, 2021
5. Shimabukuro DW, Barton CW, Feldman MD, Mataraso SJ, Das R: Effect of a machine learning-based severe sepsis prediction algorithm on patient survival and hospital length of stay: A randomised clinical trial. BMJ Open Respiratory Research 4: e000234 2017 https://doi.org/10.1136/bmjresp-2017-000234
6. 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 27: 582–584, 2021 https://doi.org/10.1038/s41591-021-01312-x
7. US Food and Drug Administration Center for Devices and Radiological Health: Section 510(k) premarket notification response for icobrain. Available at: https://www.accessdata.fda.gov/cdrh_docs/pdf18/K181939.pdf. Accessed April 16, 2021
8. Hanley D, Prichep LS, Bazarian J, Huff JS, Nauheim R, Garrett J, Jones EB, Wright DW, O’Neill J, Badjatia N, Gandhi D, Curley KC, Chiaccierrini R, O’Neill B, Hack DC: Emergency department triage of traumatic head injury using a brain electrical activity biomarker: A multisite prospective observational validation trial. Academic Emergency Medicine 24: 617–627, 2017 https://doi.org/10.1111/acee.13175
9. Wong A, O’lles E, Donnelly JP, Krumm A, McCullough J, DeTroyer-Cooley O, Pestruj E, Phillips M, Konye I, Penoza C, Ghous M, Singh K: External validation of a widely implemented proprietary sepsis prediction model in hospitalized patients. JAMA Internal Medicine 181: 1065–1070 2021 https://doi.org/10.1001/jama.2021.2626
10. Nimri R, Battelino T, Lafiel LM, Slover RH, Schatz D, Weinheimer SA, Dove K, Danne T, Phillip M; NextDREAM Consortium: Insulin dose optimization using an automated artificial intelligence-based decision support system in youths with type 1 diabetes. Nat Med 26: 1380–1384, 2020 https://doi.org/10.1038/s41591-020-1045-7
11. Ambramoff MD, Lavin PT, Birch M, Shah N, Folk JC: Pivotal trial of an autonomous AI-based diagnostic system for detection of diabetic retinopathy in primary care offices. Digit Med 1: 39, 2018
12. Beede E, Baylor E, Hersch F, Iurchenko A, Wilcox L, Ruamviboonsuk P, Vardoulakis LM: A human-centered evaluation of a deep learning system deployed in clinics for the detection of diabetic retinopathy. Presented at the CHI ’20: CHI Conference on Human Factors in Computing Systems, Honolulu, HI, April 25–30, 2020
13. Lehman CD, Wellman RD, Buist DS, Kerlikowske K, Tosteson AN, Miglioretti DL; Breast Cancer Surveillance Consortium: Diagnostic accuracy of digital screening mammography with and without computer-aided detection. JAMA Intern Med 175: 1828–1837, 2015 https://doi.org/10.1001/jamainternmed.2015.5231
14. Fenton JL, Taplin SH, Carney PA, Abraham L, Sickles EA, D’Orsi C, Berns EA, Cutter G, Hendrick RE, Barlow WE, Elmore JG: Influence of computer-aided detection on performance of
