


ScienceDirect

Contents lists available at **sciencedirect.com**
Journal homepage: **www.elsevier.com/locate/jval**

Themed Section: Artificial Intelligence

Assessing the Economic Value of Clinical Artificial Intelligence: Challenges and Opportunities

Nathaniel Hendrix, PharmD, PhD, David L. Veenstra, PharmD, PhD, Mindy Cheng, PhD, Nicholas C. Anderson, MA, MBA, Stéphane Verguet, PhD

ABSTRACT

Objectives: Clinical artificial intelligence (AI) is a novel technology, and few economic evaluations have focused on it to date. Before its wider implementation, it is important to highlight the aspects of AI that challenge traditional health technology assessment methods.

Methods: We used an existing broad value framework to assess potential ways AI can provide good value for money. We also developed a rubric of how economic evaluations of AI should vary depending on the case of its use.

Results: We found that the measurement of core elements of value—health outcomes and cost—are complicated by AI because its generalizability across different populations is often unclear and because its use may necessitate reconfigured clinical processes. Clinicians' productivity may improve when AI is used. If poorly implemented though, AI may also cause clinicians' workload to increase. Some AI has been found to exacerbate health disparities. Nevertheless, AI may promote equity by expanding access to medical care and, when properly trained, providing unbiased diagnoses and prognoses. The approach to assessment of AI should vary based on its use case: AI that creates new clinical possibilities can improve outcomes, but regulation and evidence collection may be difficult; AI that extends clinical expertise can reduce disparities and lower costs but may result in overuse; and AI that automates clinicians' work can improve productivity but may reduce skills.

Conclusions: The potential uses of clinical AI create challenges for health technology assessment methods originally developed for pharmaceuticals and medical devices. Health economists should be prepared to examine data collection and methods used to train AI, as these may impact its future value.

Keywords: artificial intelligence, cost-effectiveness analysis, health technology assessment, value of health care.

VALUE HEALTH. 2022; 25(3):331–339

Introduction

Artificial intelligence (AI) has been defined in several different ways over its approximately 70-year history, but at present, the term generally refers to methods that predict and potentially interact with the world through rules that the machine itself creates.[1,2] The application of AI to medicine has been discussed since the field's inception, though specific AI technologies have only recently been approved for clinical use.[3,4] Researchers have hoped that AI can improve decision-making and allow for constant monitoring of patient health as it can draw from a larger body of information than clinicians can and can continuously function without rest.[5–7] Lowering healthcare expenditures is another potential advantage of clinical AI, for which the marginal cost of producing additional outputs approaches 0.[8]

A published taxonomy has identified 3 main ways that clinical AI can be used: to enhance clinical possibilities, extend clinician expertise, and automate clinician work.[9] First, it can do things that human clinicians cannot do, either because of cognitive constraints or competing demands on their time. Existing examples include interpretation of chest x-rays with purported higher accuracy than radiologists and detection of atrial fibrillation from constant monitoring of smartwatch data.[10,11] Next, AI has the potential to improve access to healthcare, particularly access to specialty services, which are often relatively scarce. The US Food and Drug Administration has approved one such technology that allows primary care providers to screen for diabetic retinopathy, thereby reducing the number of ophthalmologist referrals.[12] Potential future examples include using AI-enabled applications on smartphones to allow patients to screen themselves for skin cancer.[13,14] Finally, AI may be able to automate some repetitive clinical tasks entirely. Examples of this use include acting as a scribe to reduce the need for clinician input of documentation or interpolating scans from medical imaging to lower the amount of time that each patient uses the imaging equipment.[15,16] This taxonomy was one of the more parsimonious ones in terms of usage among several available taxonomies of clinical AI.[17–19]

Despite AI's potential advantages, many uncertainties remain on how these technologies can be used to create rather than destroy value. Health economists will likely play an important role in resolving these debates. Economic evaluation methods are well-suited to modeling health impacts from partial data and identifying promising areas for future research.[20] Long-standing conversations in the field about the nature of value also allow health economists to contribute to ethical and logistical debates about how AI can best be used.[21–23]

In this article, we explain emerging issues in AI to health economists, with a focus on identifying opportunities for future work on clinical AI and discussing how AI requires a rethinking of the traditional health technology assessment (HTA) methodology. We use the International Society for Pharmacoeconomics and Outcomes Research value framework developed by Lakdawalla et al[24] (Fig. 1) as a lens through which to categorize and discuss the ways that clinical AI can create or destroy value. This framework defines 12 ways that health technologies can contribute value and categorizes them by how frequently they are applied in contemporary HTAs. Throughout this discussion of AI, we italicize certain terms to emphasize their connection to this framework for value assessment.

Evidence Creation and Generalizability

Trials of clinical AI have, to date, consisted largely of retrospective, observational trials focused on the accuracy of AI-based predictions.[25] One major reason for this is that regulatory agencies have struggled to balance evidence requirements with an approach that is flexible enough to handle AI-powered software and devices.[26] They have adopted a process that demands increasing evidence based on the risk and the user base (ie, clinician or consumer), with requirements ranging from avoidance of harms in cases of malfunction to internal review of retrospective data—requirements that have been criticized as inadequate.[27–29]

The main outcome of HTA is a cost-effectiveness ratio that consists of summed incremental health outcomes (eg, morbidity and mortality reduction outcomes, and associated constructed measures like quality-adjusted life-years) divided by the incremental costs associated with using the intervention under consideration. Clinical trials often supply crucial health outcome data such as mortality or quality of life outcomes, though researchers may use simulation models to estimate the intervention's effects over a longer period or among a different population.

Figure 1. Visualization of the elements of value in the International Society for Pharmacoeconomics and Outcomes Research value framework. Blue lines indicate elements of value used in analyses from payer and health plan perspectives; dark blue circles, potentially novel elements of value identified in the framework; green circles, commonly used elements of value; light blue circles, common but inconsistently used elements; red lines, elements that may be used in analyses from a societal perspective. Reprinted with permissions from Elsevier.

Acquiring data on health outcomes associated with AI is challenging, as the observational studies used to evaluate its accuracy generally include only clinical validity outcomes such as an algorithm's diagnostic sensitivity and specificity. Researchers must include many more assumptions in models that attempt to estimate outcomes from these clinical validity data, such as the behavioral response to novel technology, the intervention's generalizability, and so forth—none of which are frequently included in observational studies of AI. Key metrics for assessing the net costs associated with using AI, such as the time needed to complete a task, are also rarely included in trials.[25] Many trials of clinical AI should, therefore, be interpreted as indicating whether AI could potentially work for an intended task rather than providing an estimate of clinical impact.[30] In the absence of evidence supporting value-based adoption decisions, health systems may currently choose to use AI-based on a desire to prepare for the future or to establish themselves as leaders in the adoption of novel technologies.[31,32]

A major consequence of these difficulties in evidence generation is uncertainty around how generalizable AI performance might be across different settings. When AI lacks generalizability, it is often because it is "overfitted" to a certain population. In extreme cases, AI can rely on data artifacts such as image compression settings from high prevalence settings rather than physiological features for its assessments.[33] Overfitting and poor generalizability can be hard to detect because data and code are frequently not accessible for reproducing these studies.[34]

A lack of generalizability can also occur when developers train AI on data collected under ideal conditions (eg, optimal lighting) and with imaging errors and artifact examples removed.[35] This means that, when an algorithm is translated into real clinical settings, its performance drops dramatically. For low- and middle-income countries where chronic healthcare worker shortages make AI especially appealing, this is a notable concern. In one example, an AI-based screening tool for diabetic retinopathy could not be used with over 20% of trial participants in a middle-income country because room lighting could not be dimmed enough to capture a clear image.[36]

In its most pernicious form, a lack of generalizability in AI may mean that it perpetuates disparities in the healthcare system. For example, a widely-used algorithm allowing payers to identify high-acuity patients for extra care after discharge was found to recommend less care for black patients than white patients with illnesses of comparable severity.[37] This likely occurred because the AI treated costs as a proxy for disease severity without accounting for the lower access to healthcare that many black patients experience. Such health equity gaps are concerning and must be critically addressed when expanding AI applications. At the same time, AI has been used in research settings to correct for racial bias in the clinician assessment of breast cancer risk and knee osteoarthritis severity.[38,39]

The difficulties surrounding evidence creation and the generalizability of AI present an important opportunity for health economists, many of whom are trained in the use of simulation-based methods that allow for the prediction of outcomes based on changes to inputs (Table 1). Disease modeling can be used to estimate how changes in predictive performance can impact patient outcomes. Studies evaluating AI can best prove their generalizability by real-world testing, especially in the context of clinical trials.[40] In the absence of this, studies may report results by location of data acquisition (eg, facility or country), technology for data acquisition (eg, camera type), and operator type, as appropriate.[41,42] As the evidence for generalizability weakens, the lower bound of the confidence interval around the AI's effectiveness should be decreased in the sensitivity analysis.

Table 1. Selected challenges and opportunities for health economists when evaluating clinical AI.

Challenge	Opportunity
Most evidence on clinical AI performance comes from retrospective studies	Outcome modeling provides methods for assessing clinical impact from limited evidence such as surrogate outcomes
AI may perpetuate inequities in healthcare	A critical perspective on data and outcomes used and subgroup analysis can reveal the potential group harms of AI
AI performance is often compared with clinician performance in unrealistic ways	Simulation methods can estimate outcomes in different settings and identify the types of collaboration between clinicians and AI for further research
A single AI algorithm can be used at many different thresholds of sensitivity and specificity	Cost-effectiveness can be used to select thresholds that optimize value
AI impacts on clinician productivity are uncertain	Sensitivity analysis can be used to explore how productivity impacts would affect the choice of how to implement AI
Technological improvements and access to more data mean that AI will likely improve over time	Dynamic models, perhaps similar to those used for infectious disease modeling, can be expanded to simulate improved performance over time
Coverage for AI is still very unclear	Cost-effectiveness modeling can be used to test different ways that coverage decisions can be used to incentivize appropriate and equitable use of AI

AI indicates artificial intelligence.

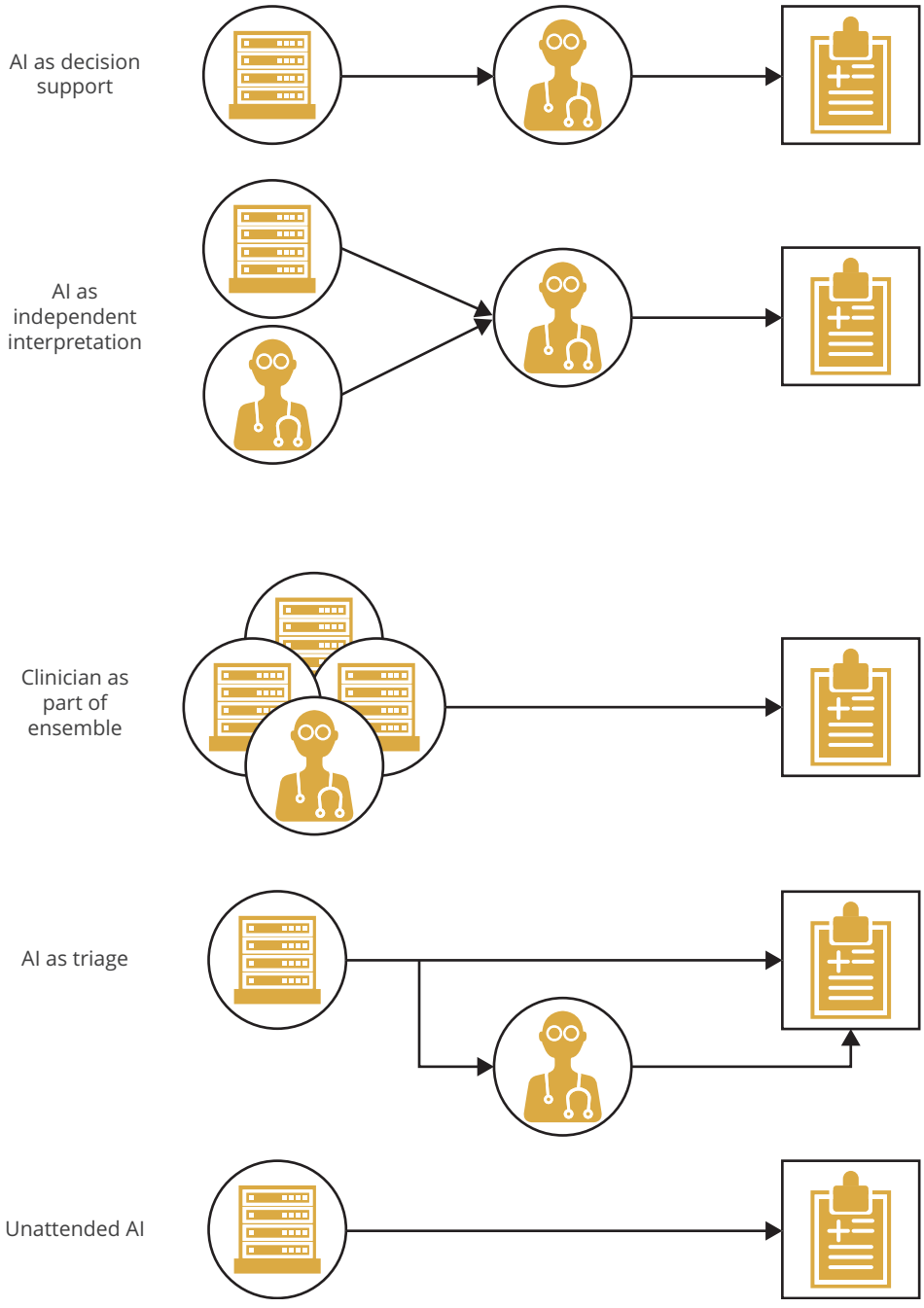
Assessments of the provenance of training data should inform HTA conducted on AI-based tools. Researchers should examine the circumstances under which data were captured and any subsequent data cleaning operations to understand how the training data might compare to real-world data. If data are captured under conditions that are not representative of the clinical settings in which the AI will be used, wider uncertainty ranges around performance should be used to account for the possibility that performance will worsen outside the study setting. Researchers should also seek to determine whether diverse individuals are represented in the training data, which may impact the AI's ability to perform equally well on all patients. This may include oversampling some relevant subgroups (with, potentially, a health equity focus) to ensure statistical power, rather than aiming for distribution in the data that represents the overall population.[43]

Finally, and perhaps most importantly, researchers should explicitly model AI's equity impacts. In part, this means critically examining the endpoints used in training AI to determine whether these endpoints incorporate failures of human judgment and our healthcare system. It also means conducting subgroup analysis, when feasible, to ensure that AI either ameliorates or does not contribute to health disparities. The quality-adjusted life-year alone is unlikely to sufficiently capture equity impacts, and economic evaluation methodologies specifically tailored to AI applications must be developed to assess how AI impacts all patients, especially the poorest and most marginalized.[32]

Integration of AI Into Clinician Work

Trials of AI frequently compare its performance to clinicians when realistic applications of AI technology would deploy it alongside clinicians.[35] There are many potential ways for AI and clinicians to work together (Fig. 2). At one extreme, AI could serve as a decision-support for a clinician, providing timely information or an independent judgment for the clinician to weigh as they make their decision. It is also possible for AI to serve as a second reader for some visual diagnostic tasks. This represents a modification of the double-reader system commonly used for breast cancer screening in Europe, in which a second clinician reviews the patients' images and the first clinician's interpretation, then resolves any discrepancies in interpretation through discussion.[44]

Figure 2. Potential ways for AI and clinicians to work together to produce clinical decisions. As decision support, AI would supply additional information or predictions to a clinician who makes the ultimate decision. As an independent interpreter, AI would reach a conclusion on its own, and its decision may be considered as a second opinion by clinicians. In an ensemble comprising both AI and clinicians, the final decision would be reached by weighting each participant's decision according to prespecified values that optimize outcomes. When AI acts as triage, it arrives at some decisions without clinician oversight and defers others to clinicians. Finally, AI may operate in some instances without routine oversight by clinicians.



AI indicates artificial intelligence.

Because a clinician cannot discuss discrepancies with an AI, a second clinician may stand in to provide another opinion.

Clinicians' judgments may also be entered into an ensemble, including different AI-based tools whose assessments are weighted according to a prespecified algorithm before recording the diagnosis. This method of collaboration is currently theoretical but has emerged as an area for future research after multiple studies found improved performance over either AI or clinicians alone.[45,46] AI can also triage decisions, where simple or low-risk cases are handled entirely by AI and higher-risk or less certain decisions are referred to a clinician. Finally, AI may work in some roles with only periodic oversight, such as when it controls the distribution of insulin in an artificial pancreas for patients with diabetes, interpolates medical images to improve resolution or clarity, or tracks changes in dysplastic nevi over time ("mole mapping").[16,47,48]

This flexibility of many AI-based tools to be used in a multitude of ways makes their assessment a challenge. Each configuration of the AI-clinician collaboration would likely produce different outcomes. Most AI-based tools also produce a continuous probability of the target outcome rather than a binary determination. Researchers must subsequently decide what probability should be used to distinguish between negatives and positives, which is sometimes referred to as the operating point.[49,50] Because the operating point corresponds to the sensitivity and specificity of the AI-based tool, optimal outcomes may be produced with different operating points in different configurations of the AI-clinician collaboration. For example, if AI were used to triage low-risk cases, it may be best to select an operating point with very high sensitivity and low specificity so that only a few positive cases fail to be referred to the clinician.[51] Other applications may require a different balance of sensitivity and specificity. The potential combinations of operating points and configurations of AI-clinician work may lead to a very large number of comparators.

Depending on how AI is used, it may increase or decrease clinician productivity. One economic argument for AI is that it would allow repetitive, simple, or low-acuity cases to be handled by AI while the clinician focuses on cases where their expertise is most valuable.[52] In other words, if AI is selectively used for tasks where the opportunity cost of the clinician's time is high, clinician productivity will increase. There are some indications that using AI in a decision-support role may also reduce heterogeneity around clinician assessments, as in one recent study using AI to assess breast density in screening mammograms.[53] This not only improves productivity but reduces uncertainty caused by clinicians with different performance levels.

In contrast, it is well known that AI can reduce productivity as well. If users have high confidence in an imperfect AI, they may become complacent and fail to adequately monitor its performance.[54] Complacency can lead to a feeling of passivity, which transforms into burnout over time.[55] If this situation continues, the user may also lose the skills that allow them to perform independently, a process called "deskilling."[56,57] When AI does not perform well, though, it can also increase the user's workload by forcing them to dedicate substantial time and energy to monitoring the AI.[54]

When assessing an AI's impact on clinician productivity, it may be important to evaluate how similar its decisions are to a clinician's. If the AI, for example, detects the same cases as a clinician, it is unlikely to improve the clinician's productivity and instead may best be used in an unattended role. Productivity is instead optimized when AI and clinicians make complementary decisions, such as might be quantified using inter-rater reliability statistics such as Cohen's kappa.[58] Nevertheless, when AI and clinicians make different assessments on a given case, assigning value to the marginal gains or losses may be complicated. In particular, it is important to assess the role of "loss aversion," which refers to the concept that individuals place greater value on losses than they do on equivalent gains.[59] An AI-based tool may be undesirable if it makes mistakes that no clinician would, even if its overall performance is better than the clinicians.

The complexities and uncertainties around how AI will be used present challenges to conventional HTA methodology. If HTA is being used to decide how an AI-based tool is implemented, health economists should consider including as comparators (ie, counterfactuals) all the potential ways that an AI could be used alongside clinicians. If work is reconfigured around AI, it may be appropriate to include the reconfigured workflow without AI as a comparator. Among these comparators should be included a reasonable range of potential operating points, meaning that a very large number of comparators may be considered. Novel methods may need to be developed for comprehensively assessing and visualizing value when the number of comparators substantially exceeds the reduced number used in most analyses.

Health economists should also incorporate AI's productivity impacts on clinicians. Nevertheless, it cannot be assumed that AI will increase productivity in all circumstances. It may also be appropriate in some cases to model clinician deskilling over time as a consequence of using AI, despite little being known about this process in most applications at present.

Finally, researchers should, when possible, incorporate assessments of the overlap between AI and clinician performance into HTAs, although not all trials of AI-based tools provide the necessary information to identify these overlaps. These overlaps can indicate the maximum potential sensitivity of an AI-clinician collaboration. In scenarios wherein AI is granted some autonomy (such as triage or unattended uses), loss aversion can be included in models to differently weigh the utility of newly-missed versus newly-identified cases.

Speed and Real-Time Technological Innovation

Depending on how AI is integrated into clinics, it may produce radically faster decisions than the status quo. Rather than sending data to a specialist for interpretation, which may take days or even weeks, AI can often produce similar results in minutes. This can affect clinician productivity as outlined above, but can also improve patient adherence.[60] Although the reasons for lack of follow-up are multifactorial, AI's ability to provide near-instant predictions and interpretations may reduce the amount of delay in or loss to follow-up that often occurs between encounters with the healthcare system. To the extent that disadvantaged patients experience barriers to access such as transportation, scheduling difficulties, or significant opportunity costs associated with time losses, AI's speed may have some equity impacts as well.[61,62]

A novel advantage of AI is that it can learn from data to improve its performance over time. This clear path to improvement may offer patients the value of hope in AI's ability to optimize decisions and in the state of healthcare technology overall. The algorithms underlying AI can sometimes be reused across applications or settings. Thus, AI's improvement over time may offer scientific spillover if an improved algorithm for 1 application leads to performance improvements in another application as well.[63,64]

Nevertheless, this ability to change over time also introduces some challenges to the modeling of these technologies. Many specific subjects of traditional HTA, such as pharmaceuticals and, to a lesser extent, diagnostics, might not change much in effectiveness over time, even as healthcare practice—for example, the

development of new guidelines—and the availability of other interventions changes around them. The changes to practice that contribute to the appropriate use of medications and conventional devices are frequently included as a sensitivity or scenario analysis but are much more central to the potential value of AI. This means that time-varying, dynamic methods may be useful in accounting for changes in effectiveness over time and determining when performing audits and subsequent HTA should be conducted. Methods used to evaluate vaccines may provide a basis for modeling these attributes of AI, given that health outcomes change as herd immunity is achieved and health system factors leading to successful adoption may change over time. The improvement in AI's performance is also often because of improved access to data. As such, the rate of improvement may be related to the product's uptake.[65,66]

There are several ways that health economists can modify their HTA models to include the impacts of AI's speed and its ability to improve over time. First, it will be important to identify when AI reduces the need for patients to attend additional visits or to wait for results. In these cases, models can include patients who are lost to follow-up and subsequently experience uncontrolled disease or disease progression. Next, models will need to account for AI's changing effectiveness over time. For AI that is likely to be used only once per patient, such as detection of large vessel occlusion stroke in an emergency setting, this can mean modeling a cohort of patients exposed to the AI as it operates at introduction; then, subsequently, sensitivity analyses can be used to estimate value over time for future cohorts.[67] When AI is used repeatedly, as, in a screening setting, the same patients are exposed to AI as it improves over time. This improvement will need to be captured explicitly in the model. Finally, uptake can be modeled to estimate the pace of AI's performance improvements in association with the amount of data available.

Costs

The marginal cost of producing decisions with AI is almost 0.[8] Despite this, AI's impacts on net costs are unclear. Even if AI directly substitutes for clinician labor, the prices for access to these technologies are uncertain and there are many potential indirect costs associated with the use of AI that should be accounted for on top of the cost of AI itself.

The implementation of AI in a clinic setting requires the work of many individuals in addition to clinicians. Information technology professionals will be needed to maintain the software and also any required hardware associated with the AI. Trainers may be needed to introduce novel technologies and to provide a setting in which clinicians can learn to use AI without fearing patient harm. Greater uptake of AI by clinicians does not necessarily mean that more of these labor inputs would be required. Rather, these labor costs and the cost of AI itself are likely to be fixed, and economies of scale may mean that they contribute a smaller share of the net costs when uptake is high.

Most developers emphasize the ease of using their AI products. Lowering the time and economic barriers to care may result in increased overall resource use. For example, if an AI reduces the amount of time needed to collect a given type of image, clinicians and patients may choose to collect these images in more circumstances. Moreover, reducing the time needed to collect these images would create an incentive to extract more profit from the imaging device—likely a large capital expenditure—by using it more often. When AI is used more, the resource use associated with exploring a greater number of false positives, incidental findings, and overdiagnoses should also be included in the analysis, even if the AI exceeds human performance overall.

The regulatory environment around clinical AI may also impose costs. The US Food and Drug Administration has indicated that it will ask AI vendors to collect data on their algorithms' performance and, potentially, impact on patient outcomes in the real world.[68] Whereas much remains undecided on how clinical AI will be regulated, there is a substantial possibility that clinics will need to maintain records of how AI is used, which may impose data storage and processing costs.[27] Another source of potential regulatory costs for clinics is that liability laws for clinicians are still nascent and, at present, expose clinicians to more liability when they use AI.[69] Protections from this liability are likely to be necessary before clinicians feel empowered to use AI.

Yet another issue is uncertainty around financing—that is, how AI will be covered by governments and insurance plans. When AI is used as a diagnostic, there is a clearer pathway for coverage and reimbursement, although there is less clarity when it is used as decision support.[70] Because of AI's flexibility, however, a given algorithm may be covered in different ways depending on how it is used. For example, an AI that independently assesses low-risk screening mammograms for breast cancer while referring suspicious images to radiologists would act as both a diagnostic and decision support. This would incentivize a higher threshold for referral to radiologists and, thus, more independent use of AI. The complexities and uncertainties around covering AI (eg, seeking reimbursement, publicly financing) may also mean that clinics pay for AI themselves and seek to recoup its costs through improvements in accuracy or efficiency.

Given the uncertainties around regulation, financing, coverage, and reimbursement, larger health systems may also develop their own AI in-house, in part, ensuring that their patient population is adequately and equitably served by the technology. Nevertheless, the development costs would be challenging. All but the largest and most advanced health systems would likely have difficulty acquiring adequate training data, meaning that their performance may lag behind commercial entities or that they would need to purchase data from other sources. Furthermore, the health system would be entirely liable for any harms resulting from unattended AI, meaning that using AI as clinical decision support (n which clinicians have the ultimate liability) may be the only feasible option.[71,72] All of these factors—development costs, the potential for lower performance, data acquisition costs, liability, and equity—should be accounted for when considering the development of clinical AI in a HTA.

When modeling the costs associated with AI, its price is only the starting place. It is also important to estimate the additional resources use that may emerge when the decisions, predictions, and procedures enabled by AI become more convenient than they were before. This naturally means that, in the absence of strict usage guidelines, low-value care may increase and partially offset any cost savings associated with the use of AI. It will also be important to model costs associated with data collection and with liability protection, which are likely to vary with how AI is used. Finally, when using a health system perspective, adequate uncertainty should be included around financing arrangements (eg, levels of public finance, coverage, and reimbursement rates) such that it allows for the possibility that AI may more suitably be treated as a capital expenditure, especially when its use is novel.

Conclusions

With the beginning of AI's deployment into clinical settings, a multitude of questions remains on how these technologies can be used to create value. The health economists' experience with simulation methods, comparative effectiveness, outcomes

Table 2. Varying elements of value and issues of special importance among different use cases for AI.

Use case	Explanation	Examples	Potential sources of value	Special considerations
Enhance clinical possibilities	Performs tasks that are not possible for human clinicians or performs them with greater speed and accuracy than humans can	• Inpatient sepsis risk detection • Artificial pancreas for people with diabetes	• Better performance may lead to improved outcomes • Scientific spillover: a given algorithm may be effective at improving performance at similar tasks • Value of hope: technological advance of medicine may seem greater	• Regulatory compliance: Unclear how to ensure performance of and protect from liability for system that outperforms humans • Improvement over time: As more data are collected, may improve more • Loss aversion: Even if it outperforms clinicians, AI will still miss cases
Extend clinician expertise	Replicates the work of human clinicians to improve their accessibility, as in performing specialist tasks in a primary care office or clinical tasks outside of clinic, e.g., via smartphone	• Diabetic retinopathy screening in primary care • Dermatology diagnosis by smartphone	• Equity: "democratizes" access to high-quality medical care • Lowers net costs: fewer specialist visits • Shorter time to initiation of treatment may lead to QALY gains • Reduces uncertainty by reducing heterogeneity and inconsistency around clinician findings	• Comparators: Difficult to assess if substantial heterogeneity in specialist performance • Increased use: if more convenient, may lead to unintended consequences • Patterns of clinician time may change, requiring advanced costing techniques
Automate clinician work	Replaces human tasks in order to reduce clinician burden or enhance efficiency	• Image synthesis for MRI with fewer images captured • Disease progression monitoring	• Increase productivity of providers by reducing workload • Increases clinician job satisfaction • May improve doctor-patient relationship by allowing more time face to face	• Deskilling: Performance may suffer when AI cannot be used • Performance must be measured as integrated into workflow

AI indicates artificial intelligence; MRI, magnetic resonance imaging; QALY, quality-adjusted life-year.

measurement, and identification of elements of value position them to play an active role in answering these questions. That said, the flexibility of AI to be used in many ways, the difficulty of gathering evidence about its impacts, and uncertainties about stakeholders' responses mean that traditional HTA methods will need to be expanded to fit these technologies.

In this article, we highlight in depth some challenges of modeling AI, as well as highlighting potentially novel sources of value that it may provide. We emphasize that no single approach will be suitable to modeling all AI products and that the challenges and potential value of implementation differ by its use—that is, whether AI enhances clinical possibilities, extends clinician expertise, or automates clinician work (Table 2). Understanding what considerations are necessary for a specific case in its use may also require studying the training data and methods for that algorithm, in addition to clinical, regulatory, equity, and legal considerations around its deployment.

There are many opportunities for future work in the economic evaluation of clinical AI. Among these are how models can capture behavioral responses to AI (including productivity and changes in resource use), the potential for synergy between clinicians and AI-based tools that produce complementary judgments, and AI's dynamic performance improvements in response to more data. Models of AI-based tools may also include radically greater numbers of comparators than usual HTAs. As such, methods for assessing and visualizing the results of these studies with many comparators will need development. Although these challenges are daunting, the enhancement of HTA methods to fit these technologies will help ensure that their translation into clinical settings is value-based and effective at improving patients' lives.

Article and Author Information

Accepted for Publication: August 17, 2021

Published Online: October 8, 2021

doi: https://doi.org/10.1016/j.jval.2021.08.015

Author Affiliations: Department of Global Health and Population, Harvard T.H. Chan School of Public Health, Boston, MA, USA (Hendrix, Verguet); The Comparative Health Outcomes, Policy, and Economics (CHOICE) Institute, University of Washington, Seattle, WA, USA (Veenstra); Global Access and Health Economics, Roche Molecular Systems, Inc, Pleasanton, CA, USA (Cheng); N C Anderson Consulting, Highland, UT, USA (Anderson).

Correspondence: Nathaniel Hendrix, PharmD, PhD, Department of Global Health and Population, Harvard T.H. Chan School of Public Health, 665 Huntington Ave 1-1104, Boston, MA 02115, USA. Email: nhendrix@hsph.harvard.edu

Author Contributions: *Concept and design:* Hendrix, Veenstra, Cheng, Anderson, Verguet

Analysis and interpretation of data: Hendrix, Veenstra, Cheng, Anderson, Verguet
Drafting of the manuscript: Hendrix
Critical revision of the paper for important intellectual content: Hendrix, Veenstra, Cheng, Anderson, Verguet
Supervision: Verguet

Conflict of Interest Disclosures: Dr Veenstra is an editor for *Value in Health* and had no role in the peer-review process of this article. Dr Cheng is an employee of Roche Molecular Systems Inc and this collaboration was done in her personal capacity. Roche did not fund or provide any financial support for this study and the views and opinions expressed herein are the authors' own and do not represent those of Roche Molecular Systems Inc. No other disclosures were reported.

Funding/Support: The authors received no financial support for this research.

Acknowledgment: The authors thank the 3 anonymous reviewers for their thoughtful and constructive comments on our work.

REFERENCES

1. Garvey C. Interview with Colin Garvey, Rensselaer Polytechnic Institute. Artificial intelligence and systems medicine convergence. *Omics*. 2018;22(2):130–132.
2. Agrawal A, Gans J, Goldfarb A. Introduction. In: Agrawal A, Gans J, Goldfarb A, eds. *The Economics of Artificial Intelligence*. Chicago, IL: University of Chicago Press; 2019.
3. Miller RA. Computer-assisted diagnostic decision support: history, challenges, and possible paths forward. *Adv Health Sci Educ Theory Pract*. 2009;14(suppl 1):89–106.
4. Topol EJ. High-performance medicine: the convergence of human and artificial intelligence. *Nat Med*. 2019;25(1):44–56.
5. Athey S. Beyond prediction: using big data for policy problems. *Science*. 2017;355(6324):483–485.
6. Stinton C, Jenkinson D, Adekanmbi V, Clarke A, Taylor-Phillips S. Does time of day influence cancer detection and recall rates in mammography? In: Proceeding from the SPIE 10136, Medical Imaging 2017: Image Perception, Observer Performance, and Technology Assessment; March 10, 2017; Orlando, FL. Abstract 10136B.
7. Berrouiguet S, Barrigón ML, Castroman JL, Courtet P, Artés-Rodríguez A, Baca-García E. Combining mobile-health (mHealth) and artificial intelligence (AI) methods to avoid suicide attempts: the Smartcrises study protocol. *BMC Psychiatry*. 2019;19(1):277.
8. Beam AL, Kohane IS. Translating artificial intelligence into clinical care. *JAMA*. 2016;316(22):2368–2369.
9. Price WN II. Artificial intelligence in the medical system: four roles for potential transformation. Preprint. Posted online March 18, 2019. University of Michigan Public Law Research Paper No. 631. https://ssrn.com/abstract=3341692. Accessed December 11, 2020.
10. Rajpurkar P, Irvin J, Zhu K, et al. CheXNet: radiologist-level pneumonia detection on chest X-rays with deep learning. Preprint. Posted online November 14, 2017. arXiv:1711.05225. http://arxiv.org/abs/1711.05225. Accessed December 11, 2020.
11. Perez MV, Mahaffey KW, Hedlin H, et al. Large-scale assessment of a smartwatch to identify atrial fibrillation. *N Engl J Med*. 2019;381(20):1909–1917.
12. van der Heijden AA, Abramoff MD, Verbraak F, van Hecke MV, Liem A, Nijpels G. Validation of automated screening for referable diabetic retinopathy with the IDx-DR device in the Hoorn Diabetes Care System. *Acta Ophthalmol*. 2018;96(1):63–68.
13. Zakhem GA, Motosko CC, Ho RS. How should artificial intelligence screen for skin cancer and deliver diagnostic predictions to patients? *JAMA Dermatol*. 2018;154(12):1383–1384.
14. Hekler A, Utikal JS, Enk AH, et al. Superior skin cancer classification by the combination of human and artificial intelligence. *Eur J Cancer*. 2019;120:114–121.
15. Lin SY, Shanafelt TD, Asch SM. Reimagining clinical documentation with artificial intelligence. *Mayo Clin Proc*. 2018;93(5):563–565.
16. Chaudhari AS, Sandino CM, Cole EK, et al. Prospective deployment of deep learning in MRI: a framework for important considerations, challenges, and recommendations for best practices. *J Magn Reson Imaging*. 2021;54(2):357–371.
17. Dixon BE, Zafar A, McGowan JJ. Development of a taxonomy for health information technology. *Stud Health Technol Inform*. 2007;129(Pt 1):616–620.
18. Jiang F, Jiang Y, Zhi H, et al. Artificial intelligence in healthcare: past, present and future. *Stroke Vasc Neurol*. 2017;2(4):230–243.
19. Combi C, Pozzi G. Clinical information systems and artificial intelligence: recent research trends. *Yearb Med Inform*. 2019;28(1):83–94.
20. Habbema JD, Wilt TJ, Etzioni R. Models in the development of clinical practice guidelines. *Ann Intern Med*. 2015;162(7):530–531.
21. Morley J, Machado CCV, Burr C, et al. The ethics of AI in health care: a mapping review. *Soc Sci Med*. 2020;260:113172.
22. Ostherr K. Artificial intelligence and medical humanities [published online July 11, 2020]. *J Med Humanit*. https://doi.org/10.1007/s10912-020-09636-4.
23. Morley J, Floridi L. An ethically mindful approach to AI for health care. *Lancet*. 2020;395(10220):254–255.
24. Lakdawalla DN, Doshi JA, Garrison LP, Phelps CE, Basu A, Danzon PM. Defining elements of value in health care—a health economics approach: an ISPOR Special Task Force Report:[3]. *Value Health*. 2018;21(2):131–139.
25. Nagendran M, Chen Y, Lovejoy CA, et al. Artificial intelligence versus clinicians: systematic review of design, reporting standards, and claims of deep learning studies. *BMJ*. 2020;368:m689.
26. General wellness: policy for low risk devices; guidance for Industry and Food and Drug Administration staff; availability. *Fed Regist*. 2016;81:49993–49995.
27. Hwang TJ, Kesselheim AS, Vokinger KN. Lifecycle regulation of artificial intelligence–and machine learning–based software devices in medicine. *JAMA*. 2019;322(23):2285–2286.
28. Babic B, Gerke S, Evgeniou T, Glenn Cohen I. Direct-to-consumer medical machine learning and artificial intelligence applications. *Nat Mach Intell*. 2021;3(4):283–287.
29. 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(4):582–584.
30. Cabitza F, Zeitoun JD. The proof of the pudding: in praise of a culture of real-world validation for medical artificial intelligence. *Ann Transl Med*. 2019;7(8):161.
31. Fan W, Liu J, Zhu S, Pardalos PM. Investigating the impacting factors for the healthcare professionals to adopt artificial intelligence-based medical diagnosis support system (AIMDSS). *Ann Oper Res*. 2020;294(1):567–592.
32. Obermeyer Z, Weinstein JN. Adoption of artificial intelligence and machine learning is increasing, but irrational exuberance remains. *NEJM Catal*. 2020;1(1):2–18.
33. Zech JR, Badgeley MA, Liu M, Costa AB, Titano JJ, Oermann EK. Variable generalization performance of a deep learning model to detect pneumonia in chest radiographs: a cross-sectional study. *PLoS Med*. 2018;15(11):e1002683.
34. Schwalbe N, Wahl B. Artificial intelligence and the future of global health. *Lancet*. 2020;395(10236):1579–1586.
35. Topol EJ. Welcoming new guidelines for AI clinical research. *Nat Med*. 2020;26(9):1318–1320.
36. 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. In: Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems; Honolulu HI; April 21, 2020; 1-12.
37. 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):447–453.
38. Yala A, Lehman C, Schuster T, Portnoi T, Barzilay R. A deep learning mammography-based model for improved breast cancer risk prediction. *Radiology*. 2019;292(1):60–66.
39. Pierson E, Cutler DM, Leskovec J, Mullainathan S, Obermeyer Z. An algorithmic approach to reducing unexplained pain disparities in underserved populations. *Nat Med*. 2021;27(1):136–140.
40. Park SH, Choi J, Byeon JS. Key principles of clinical validation, device approval, and insurance coverage decisions of artificial intelligence. *Korean J Radiol*. 2021;22(3):442–453.
41. Park SH, Han K. Methodologic guide for evaluating clinical performance and effect of artificial intelligence technology for medical diagnosis and prediction. *Radiology*. 2018;286(3):800–809.
42. Rose S. Machine learning for prediction in electronic health data. *JAMA Netw Open*. 2018;1(4):e181404.
43. Parikh RB, Teeple S, Navathe AS. Addressing bias in artificial intelligence in health care. *JAMA*. 2019;322(24):2377–2378.
44. Salim M, Wåhlin E, Dembrower K, et al. External evaluation of 3 commercial artificial intelligence algorithms for independent assessment of screening mammograms. *JAMA Oncol*. 2020;6(10):1581–1588.
45. Schaffter T, Buist DSM, Lee CI, et al. Evaluation of combined artificial intelligence and radiologist assessment to interpret screening mammograms [published correction appears in *JAMA Netw Open*. 2020;3(3):e204429]. *JAMA Netw Open*. 2020;3(3):e200265.
46. Wu N, Phang J, Park J, et al. Deep neural networks improve radiologists' performance in breast cancer screening. *IEEE Trans Med Imaging*. 2020;39(4):1184–1194.
47. Kesavadev J, Saboo B, Krishna MB, Krishnan G. Evolution of insulin delivery devices: from syringes, pens, and pumps to DIY artificial pancreas. *Diabetes Ther*. 2020;11(6):1251–1269.
48. Sondermann W, Utikal JS, Enk AH, et al. Prediction of melanoma evolution in melanocytic nevi via artificial intelligence: a call for prospective data [published correction appears in *Eur J Cancer*. 2019;123:171]. *Eur J Cancer*. 2019;119:30–34.
49. van Erkel AR, Pattynama PM. Receiver operating characteristic (ROC) analysis: basic principles and applications in radiology. *Eur J Radiol*. 1998;27(2):88–94.
50. Flach PA. ROC analysis. In: *Encyclopedia of Machine Learning and Data Mining*. Boston, MA: Springer; 2016:1–8.

51. McKinney SM, Sieniek M, Godbole V, et al. International evaluation of an AI system for breast cancer screening. *Nature*. 2020;577(7788):89–94.
52. Acemoglu D, Restrepo P. Artificial intelligence, automation and work. National Bureau of Economic Research. https://www.nber.org/papers/w24196. Accessed January 17, 2021.
53. Lehman CD, Yala A, Schuster T, et al. Mammographic breast density assessment using deep learning: clinical implementation. *Radiology*. 2019;290(1):52–58.
54. Parasuraman R, Sheridan TB, Wickens CD. A model for types and levels of human interaction with automation. *IEEE Trans Syst Man Cybern A Syst Hum*. 2000;30(3):286–297.
55. Yu KH, Kohane IS. Framing the challenges of artificial intelligence in medicine. *BMJ Qual Saf*. 2019;28(3):238–241.
56. Miller DD, Brown EW. Artificial intelligence in medical practice: the question to the answer? *Am J Med*. 2018;131(2):129–133.
57. Levy J, Jotkowitz A, Chowers I. Deskilling in ophthalmology is the inevitable controllable? *Eye (Lond)*. 2019;33(3):347–348.
58. Dzindolet MT, Peterson SA, Pomranky RA, Pierce LG, Beck HP. The role of trust in automation reliance. *Int J Hum Comput Stud*. 2003;58(6):697–718.
59. Kahneman D, Knetsch JL, Thaler RH. Anomalies: the endowment effect, loss aversion, and status quo bias. *J Econ Perspect*. 1991;5(1):193–206.
60. Schapira MM, Barlow WE, Conant EF, et al. Communication practices of mammography facilities and timely follow-up of a screening mammogram with a BI-RADS 0 assessment. *Acad Radiol*. 2018;25(9):1118–1127.
61. Jones BA, Dailey A, Calvocoressi L, et al. Inadequate follow-up of abnormal screening mammograms: findings from the race differences in screening mammography process study (United States). *Cancer Causes Control*. 2005;16(7):809–821.
62. Lake M, Shusted CS, Juon H-S, et al. Black patients referred to a lung cancer screening program experience lower rates of screening and longer time to follow-up. *BMC Cancer*. 2020;20(1):561.
63. Wiens J, Guttag J, Horvitz E. A study in transfer learning: leveraging data from multiple hospitals to enhance hospital-specific predictions. *J Am Med Inform Assoc*. 2014;21(4):699–706.
64. Gupta P, Malhotra P, Narwariya J, Vig L, Shroff G. Transfer learning for clinical time series analysis using deep neural networks. *J Healthc Inform Res*. 2020;4(2):112–137.
65. Mukherjee S, Tamayo P, Rogers S, et al. Estimating dataset size requirements for classifying DNA microarray data. *J Comput Biol*. 2003;10(2):119–142.
66. Figueroa RL, Zeng-Treitler Q, Kandula S, Ngo LH. Predicting sample size required for classification performance. *BMC Med Inform Decis Mak*. 2012;12(1):8.
67. Morey JR, Fiano E, Yaeger KA, Zhang X, Fifi JT. Impact of viz LVO on time-to-treatment and clinical outcomes in large vessel occlusion stroke patients presenting to primary stroke centers. Preprint. Posted online July 5, 2020. medRxiv 2020.07.02.20143834. https://doi.org/10.1101/2020.07.02.20143834.
68. Developing software precertification program: a working model. U.S. Food &Drug Administration. https://www.fda.gov/media/119722/download. Accessed December 11, 2020.
69. Price WN II, Gerke S, Cohen IG. Potential liability for physicians using artificial intelligence. *JAMA*. 2019;322(18):1765–1766.
70. Daniel G, Silcox C, Sharma I, Wright M. Current state and near-term priorities for AI-enabled diagnostic support software in health care. Duke Margolis Center for Health Policy. https://healthpolicy.duke.edu/publications/current-state-and-near-term-priorities-ai-enabled-diagnostic-support-software-health. Accessed January 19, 2021.
71. Sutton RT, Pincock D, Baumgart DC, Sadowski DC, Fedorak RN, Kroeker KI. An overview of clinical decision support systems: benefits, risks, and strategies for success. *NPJ Digit Med*. 2020;3(1):17.
72. Maliha G, Gerke S, Cohen IG, Parikh RB. Artificial intelligence and liability in medicine: balancing safety and innovation. *Milbank Q*. 2021;99(3):629–647.