MAYO CLINIC

MAYO CLINIC PROCEEDINGS:
DIGITAL HEALTH

REVIEW

Machine Learning Operations in Health Care: A Scoping Review

Anjali Rajagopal, MBBS; Shant Ayanian, MD, MS; Alexander J. Ryu, MD; Ray Qian, MD; Sean R. Legler, MD; Eric A. Peeler, MD; Meltiady Issa, MD, MBA; Trevor J. Coons, MHA; and Kensaku Kawamoto, MD, PhD, MHS

Abstract

The use of machine learning tools in health care is rapidly expanding. However, the processes that support these tools in deployment, that is, machine learning operations, are still emerging. The purpose of this work was not only to provide a comprehensive synthesis of existing literature in the field but also to identify gaps and offer insights for adoption in clinical practice. A scoping review was conducted using the MEDLINE, PubMed, Google Scholar, Embase, and Scopus databases. We used MeSH and non-MeSH search terms to identify pertinent articles, with the authors performing 2 screening phases and assigning relevance scores: 148 English language articles most salient to the review were eligible for inclusion; 98 offered the most unique information and these were supplemented by 50 additional sources, yielding 148 references. From the 148 references, we distilled 7 key topic areas, based on a synthesis of the available literature and how that aligned with practitioner needs. The 7 topic areas were machine learning model monitoring; automated retraining systems; ethics, equity, and bias; clinical workflow integration; infrastructure, human resources, and technology stack; regulatory considerations; and financial considerations. This review provides an overview of best practices and knowledge gaps of this domain in health care and identifies the strengths and weaknesses of the literature, which may be useful to health care machine learning practitioners and consumers.

■ Mayo Clin Proc Digital Health 2024;2(3):421-437

The adoption of machine learning (ML) tools in health care has accelerated sharply in recent years. However, the science and processes that support and monitor these tools in live use, which we collectively refer to as machine learning operations (MLOps),[1,2] are still nascent. The rapid adoption of ML tools, combined with even faster advancement in ML technologies (eg, large language models), has introduced numerous risks to health care workflows, which have the potential to endanger patients. Specifically, errors in operationalizing ML models can lead to direct patient harm,[3] bias against underprivileged groups,[4] mistrust of artificial intelligence (AI), inefficiency for providers, and poor allocation of limited ML resources.

In many cases, an established framework to ensure effective deployment of ML tools is lacking. Therefore, it is critical that MLOps practices evolve in tandem to monitor and mitigate these risks.

Software development also began with a freeform approach. However, as the complexity of managing increasingly sophisticated live applications grew, there was a need for greater standardization. This need arose to reduce real-world errors and confusion among developers. Consequently, standardized processes were developed, eventually maturing into the established field of DevOps (software development operations).[5,6]

DevOps codifies a practice used for iteratively developing, deploying, and maintaining reliable software. Machine learning operations has similarly developed into a nascent field with dedicated scientists and engineers. Now more than ever, MLOps advances and shortcomings warrant urgent dissemination and evaluation to prevent harm and maximize the effectiveness of powerful new ML tools.

In this scoping review, we aimed to synthesize and categorize existing literature around key MLOps topic areas distilled from

From the Department of Medicine, Artificial Intelligence and Innovation, Mayo Clinic Rochester, MN (A.R.); Division of Hospital Internal Medicine, Department of Medicine, Mayo Clinic, Rochester, MN (S.A., A.J.R., R.Q., S.R.L., E.A.P., M.I.); Heart, Vascular and Thoracic Institute, Cleveland Clinic Abu Dhabi, United Arab Emirates (T.J.C.); and Department of Biomedical Informatics, University of Utah, Salt Lake City, UT (K.K.).

ARTICLE HIGHLIGHTS

- Given the proliferation of artificial intelligence tools available for clinical use, there is an urgent need to define health care machine learning operations (MLOps) best practices, which can promote safe, consistent, and equitable use of artificial intelligence in health care.
- Our scoping review of health care MLOps identified 148 relevant articles; most of these were published in 2022 or after and were led by North American or European investigators.
- We identified 7 key topic areas covered by existing literature: (1) machine learning model monitoring, (2) automated retraining systems, (3) ethics, equity, and bias, (4) clinical workflow integration, (5) infrastructure, human resources, and technology stack, (6) regulatory considerations, and (7) financial considerations.
- There remains a need for studies that rigorously evaluate MLOps practices in terms of prospective impacts on patients and the health care system. Many studies in our review were focused on statistical assessments of machine learning model performance, retrospective MLOps evaluations or MLOps simulations.

our literature review. We also aimed to highlight strengths in the literature as well as areas meriting further investigation.

MATERIALS AND METHODS

To compile our review, we searched the MEDLINE, PubMed, Embase, Scopus, and Google Scholar databases. We used MeSH and non-MeSH search terms to identify articles of interest. Search terms were chosen to find articles that pertained directly to MLOps considerations in health care, specifically, postdeployment monitoring, maintenance, fairness, and cost considerations. For example, search terms included the following: *health care ML monitoring, health care ML implementation, and health care ML drift*. Search queries were repeated substituting the terms AI, ML, and model in each query, with queries applied to all article fields. Specific queries are enumerated in Supplemental Appendix—Search Queries (available online at https://www.mcpdigitalhealth.org/). For Google Scholar, the first 3, and up to 10, pages of results were reviewed, if relevant results persisted beyond the initial 3 pages.

Only peer-reviewed, English language articles were included. Research manuscripts, review manuscripts, and commentary pieces were considered for inclusion, whereas student theses, letters to the editor, book chapters, and abstracts were not. Articles published any time up until October 15, 2023, were included. The initial review was completed on March 23, 2023; one subsequent update was completed on October 15, 2023. For Embase and Scopus results, articles were additionally filtered to those containing *healthcare* or *artificial intelligence* or *machine learning* as keywords, to ensure that results were substantially focused on MLOps in health care. Results from Scopus were also limited to those not including *internet of things* as a keyword because this otherwise yielded a large number of theoretical internet of things investigations with minimal MLOps relevance.

Data were extracted to Excel using built-in export tools for Embase and Scopus, whereas a librarian provided an exported spreadsheet for MEDLINE. Results were exported manually to Excel from Google Scholar.

Two authors (A.J.R., S.A.) performed an initial screening of articles by title and abstract to exclude those clearly unrelated to health care MLOps. Specifically, articles that did not have any focus on postdeployment issues were excluded: for example, studies that did not pertain to AI/ML, retrospective ML model validation studies, and articles that focused on reviewing the hypothetical impact of AI on various health care domains. Articles without direct relevance to health care were also excluded, such as articles primarily discussing ML in another practical domain, with only passing mention of health care applications. The remaining articles were then reviewed by 1 of the coauthors and scored 0-3 for relevance as follows: 0, no relevance to MLOps; 1, brief or tangential mention of MLOps; 2, MLOps commentary or editorial; and 3, MLOps research study or review. Disagreements on relevance scores were adjudicated by A.J.R. or S.A.

Eight authors (A.R., S.A., A.J.R., R.Q., S.L., E.P., M.I., and T.C.) then reviewed articles scoring 2 or 3 for relevance to develop a consensus on the main topic areas represented. Title, journal, year of publication, authors, and the lead author's country affiliation were also abstracted into Excel. In addition to our search results, we also included a limited number of additional, relevant peer-reviewed articles, guideline articles, and government white papers that the coauthors felt provided importance evidence or context. A limited number of non—peer-reviewed sources were also included to provide supplementary data or perspectives where no peer-reviewed sources could be found.

Our overall review process is summarized in Figure 1, a Consolidated Standards of Reporting Trials (CONSORT) diagram. The included articles were distilled into key topic areas based on the authors' assessment of the literature and how that information might be most usefully presented to other practitioners. These topic areas served as the foundation of the review and split among the coauthors for drafting. Strengths and weaknesses of the main topic areas were summarized.

RESULTS

In total, our initial search and update yielded 2271 abstracts for review, after duplicates were removed. Title and abstract review narrowed this to 446 articles of interest, on which relevance scoring was completed; 148 articles scored 2 or 3 and were thus eligible for inclusion in the review. Ninety-eight were ultimately included as they contained the most relevant and unique information. These 98 articles were supplemented by 9 guideline or whitepaper articles, 29 additional peer-reviewed journal articles, and 12 non peer-reviewed articles, for a total of 148 articles that were included in the review. These 148 articles are listed in Supplemental Table (available online at https://www.mcpdigitalhealth.org/). For the 177 peer-reviewed articles (including those that were not ultimately referenced), a bar graph of year of publication is included in Figure 2, and a heat map of lead author country affiliation is shown in Figure 3.

The 7 key topic areas identified were as follows: (1) ML model monitoring, (2) automated retraining systems, (3) ethics, equity, and bias, (4) clinical workflow integration, (5) infrastructure, human resources and technology stack, (6) regulatory considerations, and (7) financial considerations (Figure 4). Our review of these 7 topic areas as detailed further.

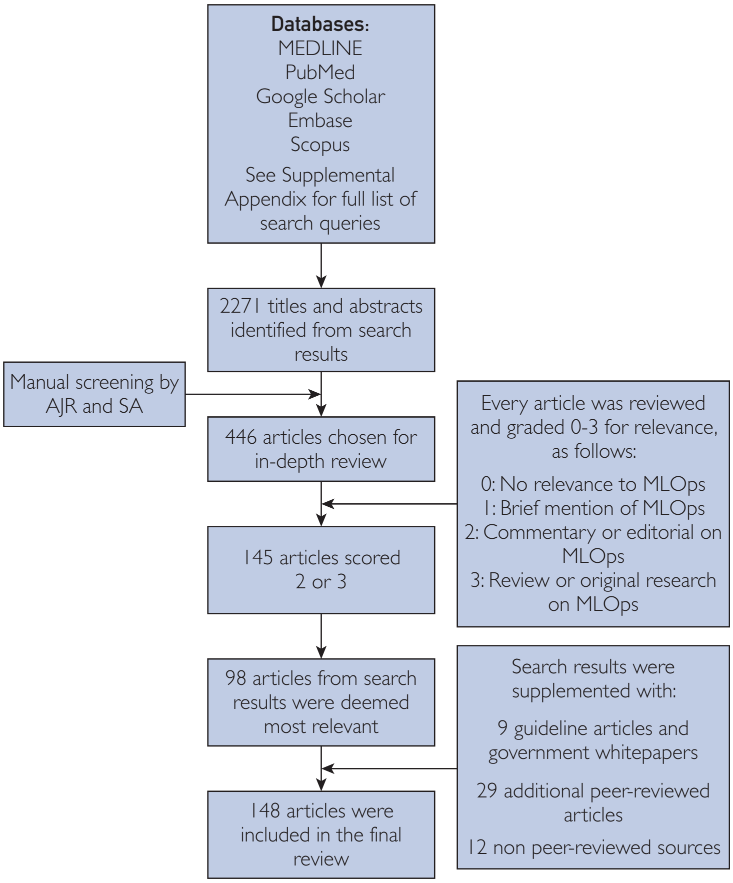
Databases:
MEDLINE
PubMed
Google Scholar
Embase
Scopus
See Supplemental Appendix for full list of search queries
2271 titles and abstracts identified from search results
Manual screening by AJR and SA
446 articles chosen for in-depth review
Every article was reviewed and graded 0-3 for relevance, as follows:
0: No relevance to MLOps
1: Brief mention of MLOps
2: Commentary or editorial on MLOps
3: Review or original research on MLOps
145 articles scored 2 or 3
98 articles from search results were deemed most relevant
Search results were supplemented with:
9 guideline articles and government whitepapers
29 additional peer-reviewed articles
12 non peer-reviewed sources
148 articles were included in the final review

FIGURE 1. Consolidated Standards of Reporting Trials (CONSORT) diagram depicting article selection process. MLOps, machine learning operations.

ML Model Monitoring

Model performance and drift monitoring have likely been the subject of the most health care MLOps publications to date.[7-11] We suspect this is partly because these issues are fairly easy to quantify and rectify. Although others have offered precise definitions for various types of model or data drift,[12] we prefer to unify the discussion of these issues around the notion that the data and circumstances used for model training no longer mirror the present environment, with potentially deleterious consequences for a model's performance and its effective use. Existing literature has shown that the relative stability of model performance is dependent on population factors,[11] clinical practice patterns,[10] model type,[11] and the prediction being made.[8,13]

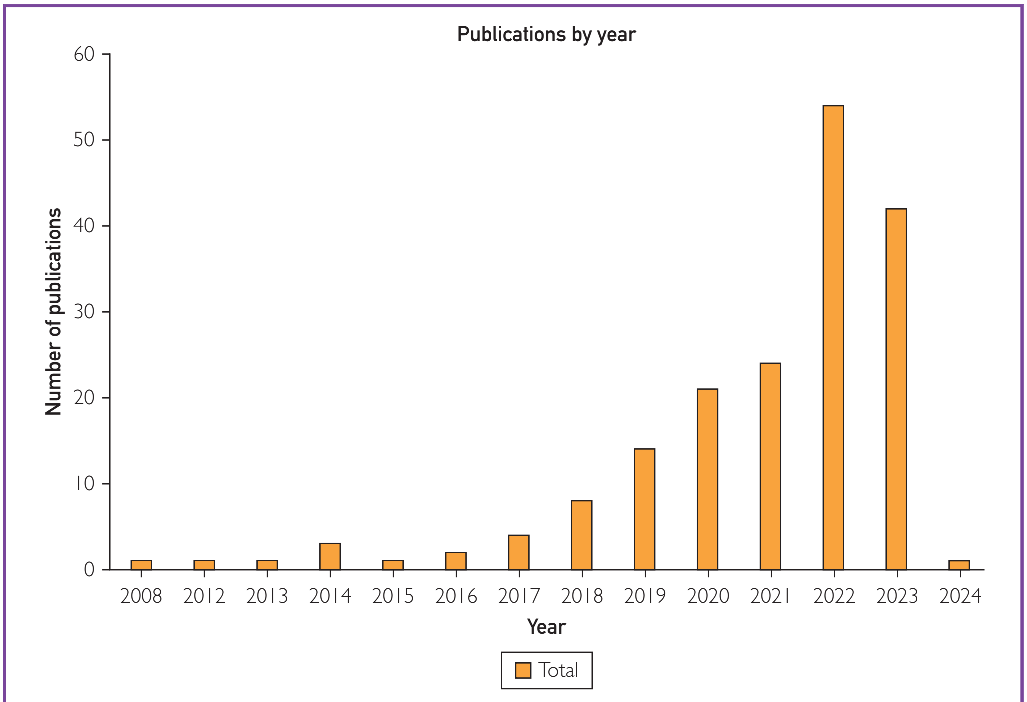

FIGURE 2. Bar graph of year of publication of articles eligible for inclusion in the review.

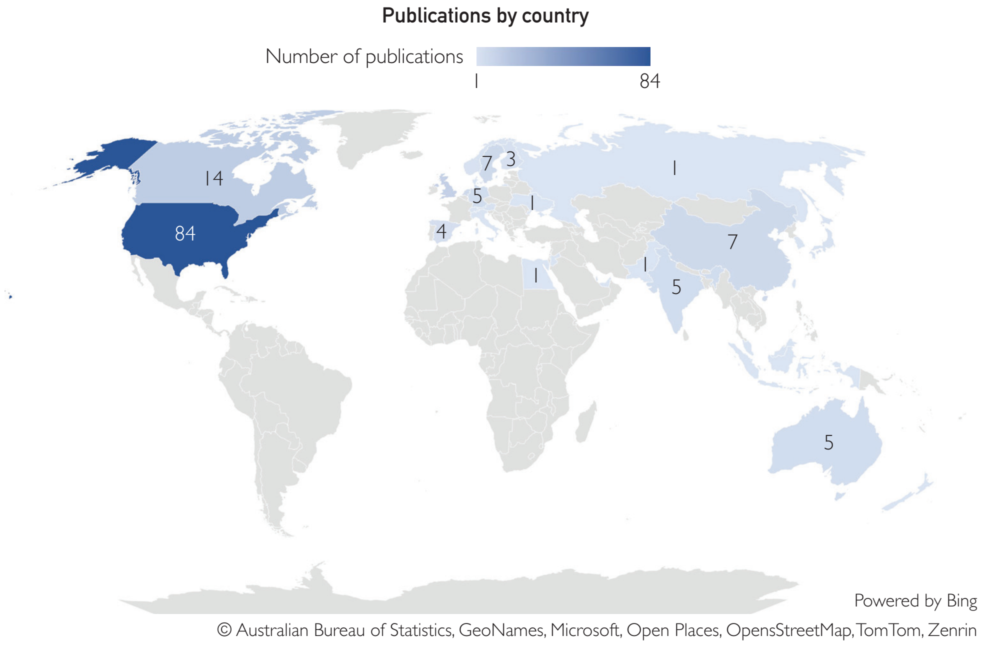
Publications by country
Number of publications
1
84
14
84
7
3
5
1
4
1
1
1
7
5
5
Powered by Bing
© Australian Bureau of Statistics, GeoNames, Microsoft, Open Places, OpensStreetMap, TomTom, Zenrin

FIGURE 3. Heat map of lead author country affiliation across articles eligible for inclusion in the review.

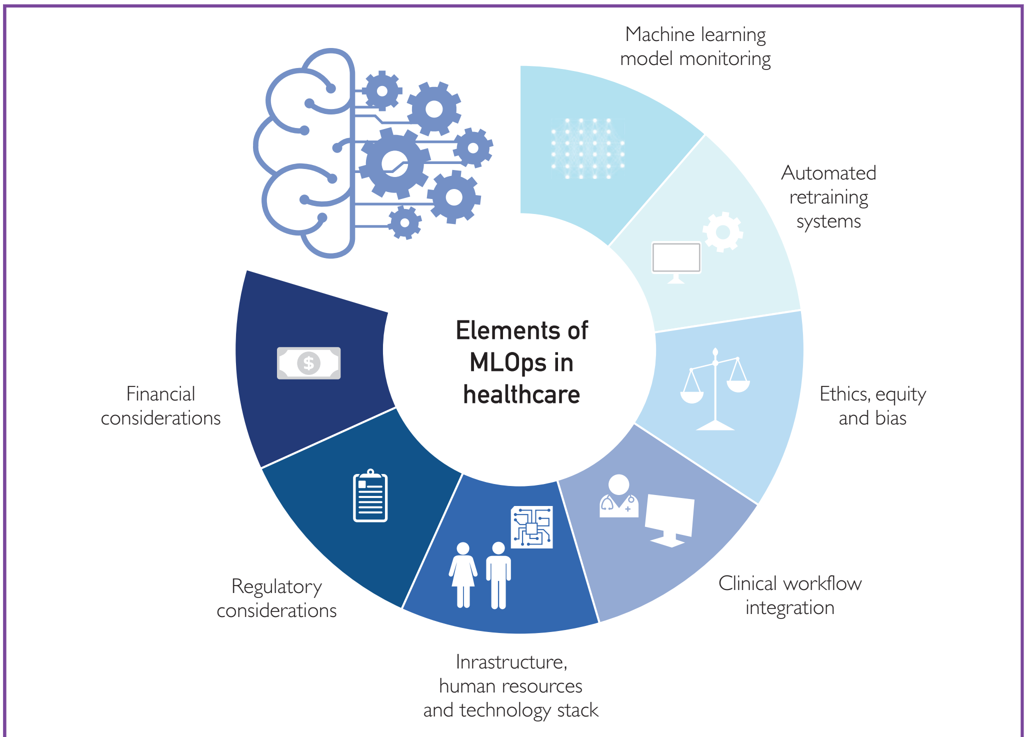
Machine learning model monitoring
Automated retraining systems
Ethics, equity and bias
Clinical workflow integration
Inrastructure, human resources and technology stack
Regulatory considerations
Financial considerations
Elements of MLOps in healthcare

FIGURE 4. MLOps topic areas covered by the review. Created with BioRender.com. MLOps, machine learning operations.

Multiple types of data and model drift have been defined, all of which contribute to model performance that is worse in live operation than in retrospective validation and testing. Notably, the terminology pertaining to different types of drift remains quite varied, with multiple terms being used interchangeably for the same issue. We therefore limit our discussion to the most commonly used and high-level terms that practitioners are most likely to encounter, borrowing from the study by Moreno-Torres et al.[14] Covariate shift is a type of data drift that occurs when the distributions of a model's features differ between training data and live data.[14] This may become apparent gradually over time or immediately when a model is deployed with live data (also known as train-serving skew, in the latter case). Similarly, previous probability shift (also known as prediction shift) occurs when the distribution of the output variable changes, either gradually or abruptly. Concept shift occurs when the relationship between features and outputs changes. ML practitioners must take heed to incorporate sufficient clinical understanding to proactively mitigate concept shift that can occur as the result of changing disease definitions, such as with sepsis, or with changes in clinical standards of care, as found with COVID-19.[15-17] Finally, Vela et al[18] describe a novel approach to understanding temporal performance degradation of AI models, which they term model aging. They note that degradation in model performance cannot be explained solely by data drift but, instead, can be attributed to unrelated factors, such as training set size or hyperparameter selection.[18] They note a need for periodic model retraining, with consideration that the appropriate retraining schedule is highly context dependent.[18]

Further, our own experience implementing an ML model for hospital admission prediction with emergency department patients highlighted multiple challenges with diagnosing and fixing issues with data drift. In our case, we had to reverse engineer data pipelines to isolate a particular extract, transform, load process that led to data mismatches between live and training data, and ultimately data drift. Further, as we implemented our model across emergency department sites with widely varying patient volumes, it became apparent that the optimal time window used for drift monitoring must be adjusted to the volume of observations occurring for a given clinical prediction.[19]

Consistent with general best practices for prospectively evaluating interventions, implemented models should have performance thresholds and assessments determined a priori, which will be used to trigger retraining or at least a reevaluation of the model and its use. Explanations for drift should be sought wherever possible because the solutions for various types of drift may differ depending on the causative factor.[11,16,20]

One particularly nuanced issue that appears insufficiently addressed in the current literature is how to optimally handle model retraining when a model's output has been used to prompt an intervention that perturbs subsequent data.[12] In such situations, retraining models on perturbed postinterventional data without appropriately accounting for the existence of this feedback loop may lead to an undoing of the change the model was meant to effect. This was elegantly simulated in studies by Adam et al[21] and Vaid et al.[22] In their simulations, they consider the effects of clinicians' adherence to model outputs, efficacy of model outputs on improving clinical outcomes, and various feedback loop mitigation strategies on model performance. Scenarios including retraining of a single model on perturbed data and the implementation of multiple predictive models along a single clinical event pathway (eg models for heart attack risk and mortality risk prediction) were modeled. Mitigation strategies that selectively sample data for model retraining or limit the application of multiple models to a single patient were considered. Ultimately, no single optimal retraining strategy was found to be broadly applicable. Nonetheless, these studies highlight important considerations for which studies will prospectively also be crucial.[21,22]

Automated Retraining Systems

A related MLOps topic that has received considerable attention is automated monitoring and retraining systems.[23,24] Automated retraining systems periodically collect model performance metrics and trigger retraining according to preset model performance thresholds. These approaches stand in contrast to monitoring and retraining at fixed time intervals, which may be suboptimal owing to the complex and, at times, unpredictable factors that contribute to diminishing model performance over time.[24]

These self-monitoring systems typically rely on a sliding window in which performance statistics are calculated over a set number of trailing observations, and when performance drops below a prespecified threshold, adjustments to the model are made.[25,26] Depending on the model type, such adjustments may include intercept correction or varying types of model recalibration or refitting.[25,27-31] Davis et al[27] additionally developed a framework for selecting the least complex model update to ameliorate specific types of drift detected in various simulations. Others have also proposed more complex methods for detecting and addressing performance drift, including transfer learning that combines insights from old and new data batches, domain generalization, and unsupervised domain adaptation methods.[32,33] Of these approaches, results from the transfer learning method were promising but were more complex to achieve than simple refitting.

Several shortcomings are apparent in the existing literature. Most research in this area focused primarily on assessing statistically significant drift, which may not always correlate with clinically significant drift.[25] Additionally, most investigations of these systems used retrospective or synthetic data, whereas studies of live implementations were scarce.[23,26-28] As with most ML tools, additional hurdles and insights will likely emerge after implementation.

Ethics, Equity, and Bias

The fair and ethical usage of AI in health care must be informed by technical, sociological, and philosophical considerations, which have received much attention in recent literature.[34-42] MLOps practices that uphold fairness and equity while minimizing bias are still evolving.[34] Notably, many issues pertaining to model equity and bias must be addressed upstream of model implementation, during data acquisition and model training.[8,20,35,43-46] In the scope of this review, we will focus on post-implementational considerations of bias and fairness, which include measuring and overcoming disparities as well as ensuring sustained appropriate use of AI tools.[39,46] Additionally, multiple definitions of AI model fairness have been described, and selecting the most appropriate lens for a given AI model will be context dependent. Common statistical definitions of fairness include equal odds and equal opportunity.[36,37] Importantly, recent publications have gone beyond purely statistical definitions of fairness to argue that a model's implementation can be fair, even if its predictions are unfair, generally in cases where the model somehow favors accuracy or a better outcome in an underprivileged group.[38]

For a model to perform with a minimum level of bias, we assume that during the preimplementation phase, every effort has been made to adhere to distributive justice principles during model design, training, and validation before model deployment.[40,41,47] These principles are encapsulated in 3 foundational axioms: equal outcomes, equal performance, and equal allocation.[39] Equal outcomes refer to underprivileged groups benefiting equally due to the ML model, whereas equal performance refers to the model's statistical performance of groups, including underprivileged groups. Finally, equal allocation ensures that resources are allocated equally among all groups, including underprivileged groups.[39,48,49]

Model agency bias (underprivileged groups having no input in the development of the model), privilege bias (the model not being available to underprivileged groups), informed mistrust (underprivileged groups believing the model is biased against them), automation bias (overdependence on model;[50] providers defaulting to model decision that may be less reliable for underprivileged groups), and dismissal bias (providers disregard model decisions in underprivileged groups owing to a higher error rate) are all additional considerations for fair model implementation.[39]

One must also consider the final usability of the model in terms of its impact on underprivileged groups. A model's features must be carefully examined along with any expected downstream actions to be taken, so as not to perpetuate or worsen biases. For example, consider a situation in which patients belonging to an underprivileged group tended to have a higher failure rate of treatment because of factors related to their underprivileged group status. If one were to design a treatment prioritization algorithm including these patients and others, it would likely be more fair and equitable to make predictions on the basis of clinical need, without consideration of the anticipated treatment failure rate.

As others have proposed, a methodical pipeline approach is recommended to recognize biases and ethical loopholes throughout the model life cycle.[42,47,51] Diverse stakeholders are needed for all-around oversight.[52]

Clinical Workflow Integration

The current practice landscape is fraught with numerous software tools at the clinician's disposal in a background of ever-growing documentation and administrative tasks.[53] The quintessential ML model is well-validated, augments decision-making, eases cognitive and clerical burdens, and is straightforward to act on. The overall clinical impact must justify the learning curve and initial challenges with adoption. Although theoretically sound, this is often challenging to implement. Watson et al[54] in their study on overcoming barriers to implementing ML in US academic medical centers, note that a high-performing model that is disregarded by a clinician is ineffectual, as is a poorly performing model that is consistently used.[54] Depending on the type of solution being deployed, patient perceptions of AI may influence adoption as well.[55] The deployment of AI/ML tools into clinical practice continues to lag the development of such tools, which unfortunately often wind up stranded as retrospective investigations rather than being used with live data to impact patient care positively.[56] Successful workflow integration requires, first and foremost, an actionable model output, followed by consideration of users' digital and physical workflows, users' understanding of an AI/ML tool's outputs, organizational support for adoption, and the capabilities and constraints of a given AI/ML infrastructure stack.[57,58]

To comprehensively understand users' workflows, most AI/ML implementation studies have used primarily ad hoc, observational methodologies.[59,60] Although several evaluation frameworks have been developed in recent years,[61-63] they are mainly centered around the reporting of clinical AI studies as opposed to real-world operational guidelines.[64] One exception is a recent study from the University Medical Center of Utrecht, which highlighted real-world operational integration of an in-house developed model for preterm neonate monitoring.[65] In their manuscript, the authors propose a framework for requirements of model creators and users to ensure safe implementation. More generally, there is convergence particularly around ensuring that model outputs are not disruptive to workflows and carefully placing such outputs at the right place and the right time workflows where they can be acted on.[53]

The extent to which a ML model is used in clinical practice greatly relies on its usability and its ability to deliver some direct benefit to the user.[66] Achieving this can be facilitated by priming stakeholders before introduction,[67] offering intuitive interfaces, providing continuous education, implementing in phases (particularly for complex interventions),[68,69] and iterating based on periodic feedback.[70,71] Uptake is also likely impacted by prospective users' preexisting attitudes toward and knowledge about AI/ML. Several studies have assessed which factors may be impactful in various user groups, including pharmacists, physicians, and health care executives. Potentially important factors included baseline knowledge about AI, trust in AI, familiarity with AI adoption strategies, as well as concerns about cost, job replacement, and overreliance on AI tools.[20,72-78] In aggregate, these studies suggested that health care workers generally share positive views about the adoption of AI.

Regarding users' understanding of AI tool outputs, limited studies have compared different AI score output formats for their ease of use or comprehension. At least some authors report that most clinicians prefer graphical user interfaces to command-line interfaces, although better comparative studies of user interfaces must follow.[23] Survey studies focused on nurse and health sciences students also indicated a growing need to incorporate basic AI education into core curricula.[78,79]

Infrastructure, Human Resources, and Technology Stack

Another vital aspect of implementing successful MLOps in health care is the thoughtful investment in the required infrastructure, technologies, and personnel throughout the entire lifecycle of ML algorithms. Although likely to be considerable in cost,[7,20,80] institutions that thoughtfully account for the resources will likely be well-positioned to realize the potential of ML models in practice. In addition, knowledge in implementation science and integration of some of its elements in the deployment of models facilitates resource allocation and better integration into workflows.[81]

Infrastructure/Technology Stack. Successful implementation of an MLOps program in health care necessitates several unique infrastructure considerations. Paramount among these is ensuring the data security of patient health information. Implementing safeguards like data encryption and de-identification, access management, monitoring, security assessments, and auditing are essential.[82,83] Equally important are provisioning computing resources for training models, including central processing units, graphics processing units, and more advanced custom processing units (eg, tensor processing units),[84] tailored to the size and complexity of the models being developed. The decision to host computing resources on-premises, in the cloud, or via a hybrid approach must be thoughtfully balanced, considering factors such as fixed vs variable costs, performance expectations, scalability, data security, and maintenance.[85-88]

Accurate, reliable, and accessible storage and integration of disparate data sources, such as those from electronic health records (EHRs), picture archiving and communication system, laboratory information system, and other clinical systems, will be vital for the utilization of ML models.[89,90] Moreover, data pipelines must be designed to ensure high availability and fidelity of data, while achieving high performance in extract, transform, and load processes.[91] Developing and deploying ML models will also necessitate selecting suitable ML frameworks, incorporating version control, containerization, orchestration, continuous integration/deployment systems,[92] and performance monitoring.[93-95] These common infrastructure needs may present an opportunity for the development of an ML model and MLOps platform, in which various customers could pay to use models that are rigorously evaluated and maintained by a third party. Such a concept has been explored theoretically,[96—98] as well as practically.[99,100]

Of note, limited study has been devoted to comparing different infrastructure stacks for running AI models in clinical practice. One study describes the development and implementation of a framework to support researcher developed retrospectively validated models directly into the EHR.[101] Similarly, there has been a few articles describing the development of frameworks allowing for the implementation of ML models outside of the EHR.[13,102,103] In our experience, most health care systems are constrained to use the limited set of infrastructure providers with which their organization has use agreements, such as Google, Microsoft, Amazon, or Epic. Homegrown infrastructure stacks are also an option for institutions with deep IT capabilities. Overall, it is likely that comparative studies between infrastructure stacks in health care will continue to be limited by the fact that institutions generally have insufficient resources to implement and compare the effectiveness of major infrastructure products from multiple vendors.[101]

Human Resources. Even with the most advanced infrastructure and comprehensive technology stack, success in MLOps cannot be realized without the right people. Technical roles such as data scientists, ML engineers, data engineers, MLOps engineers, and security/privacy experts to manage the infrastructure and technology stack are fundamental. Just as critical for success is the input of clinical domain experts, clinical informaticists, and affected stakeholders[104-106] who can provide necessary insights and feedback. Fostering an interdisciplinary culture of collaboration, continuous learning, and ethical AI use is key to maximizing the value of these human resources.

Although the importance of stakeholder collaboration has been underscored in the literature,[8,68,107,108] real-world examples of successful interdisciplinary collaboration remain sparse. Zhang et al[109] point out the need to shift the focus from academic-centered groups (eg, clinicians, data experts) to multidisciplinary teams with members from typical MLOps processes (engineers, implementation specialists, and others) receiving equal footing. Aligning views and goals on model use among teams from different domains remains a considerable challenge, which may occur partly because of the insufficient involvement of various representatives throughout the model cycle.[54] Furthermore, despite the ubiquity of model development experts, specialists appear to be scarce for long-term operations and maintenance. This becomes apparent when employees with institution-specific knowledge leave an organization, creating a deployment gap.[54] Recruitment, retention, and succession planning of an effective multidisciplinary team thus are critical elements for the long-term success of an MLOps endeavor.[110]

Regulatory Considerations

Thoughtful regulation of AI/ML requires a multidisciplinary approach. To date, literature on AI/ML regulation has primarily taken the form of white papers from regulatory agencies or expert commentaries.[111-115] The Food and Drug Administration (FDA)[112] and the European Commission (EC)[116-119] are attempting to make strides in this regard. However, the flexible, evolving nature of the technology challenges traditional standards and regulations. Both agencies have released statements to provide a philosophical framework for the future and what regulations may look like. The current framework is to treat software as medical devices with premarket approval (for the FDA, this is the 510[k] clearance) and the expectation that solution developers will plan for potential expected modifications. Such updates would then be reported at regular intervals that the software as medical devices is evolving and performing in line with expectations.[120] A major hurdle for regulatory standardization is the lack of shared, generalizable training and evaluation data sets.[121] We note that multiple large health care organizations have undertaken efforts to curate such a data set, which may enable standardized evaluation and regulation in the future.[122,123]

To complement national and international regulatory efforts, there is also a need for institutionally based regulatory bodies. One novel approach to this could be the development of AI institutional boards, similar to institutional regulatory boards.[124] Such review boards would represent the various stakeholders involved in MLOps.[81] This would help to ensure that all facets of development, implementation, and maintenance are being considered.

There is also the issue of accountability when using AI systems for health care decision-making.[125] In a general sense, clinicians are ultimately responsible for the treatment plans and outcomes of their patients. The difficulty arises, however, with decision support applications based on AI/ML. Conventional wisdom would suggest that the provider is the one who will assume the most risk in using these applications, rather than the AI/ML technology team. However, it would be naïve to assume a faulty AI/ML model that leads to patient harm would not have potential legal liability. To navigate these uncertain waters most effectively, clinicians must have education in interpreting AI tool outputs with an understanding of their limitations and risks, which has not always been the case to date.[126,127]

Financial Considerations

In terms of financial considerations in MLOps, there have been only a few detailed accounts of the costs to train, deploy, and maintain an AI model.[128,129]

Undoubtedly, financial estimations are complicated by dynamic costs that continue to evolve with ML technologies and platforms.[130] However, we recommend that specific categories of financial factors to consider include the following:[82] First, there is (1) the initial cost to develop a model, followed by (2) the costs of implementing the model in the real-world. These implementation costs include setting up cloud or server infrastructure as well as implementing interfaces necessary to communicate between data storage systems, model housing, and end-user software programs. Next are the costs of (3) end-user training for go-live and the associated feedback and early troubleshooting process.

After this initial rollout, often underestimated is (4) the ongoing costs to maintain, recalibrate, and retrain models as per the considerations laid out in this review. Next is the potential lengthy FDA clearance process, which may entail marked legal and other specialized assistance. Furthermore, suppose a model can be applied to other geographic locations or organizations, in that case, there is (6) the assessment of model appropriateness, calibration, and possible retraining at each additional site and (7) the actual implementation, end-user instruction, and maintenance at each additional site.

Of note, there can be orders of magnitude variation in the costs to train a model. A simple XGBoost model using open-source software such as Python and a previously acquired data set could have minimal model training costs. Depending on the approach and thoroughness of a local project, operational costs of a simple ML solution, including model infrastructure, data support, and engineering deployment, could range from $60,750 to $94,500 for a small-scale operation by one estimate,[131] to amounts far larger for more complex models or larger health care systems.

In contrast, training large language models like generative pretrained transformers can cost tens of millions.[80] However, once trained, these expensive, larger models can be used for multiple tasks, retrained on specific health care use cases at lower costs, or deployed as software as a service. For example, Microsoft announced in July 2023 that it would charge $30 per user per month for its AI Copilot product in Microsoft 365, which leverages its highly complex AI model that required billions of US dollars in development costs.[132] Although many models in the academic medical literature have been funded by grants or local organizations initially, ML health care models are increasingly the subject of industry and venture capital funding after demonstrating effectiveness or at earlier stages. These models may then be used off-the-shelf, after some fine-tuning or as part of a software or ML model stack.

After assessing the full spectrum of costs, any ML operationalization must accordingly weigh these costs vs the possible benefits to patients, providers, or implementing organizations. Benefits may include improved patient care for its own sake, recouped costs via secondary cost savings, or AI model-associated income generation.

Currently, business models in medical AI are still developing as developers of AI tools work to build clinical and business cases for the value their tools provide.[110] Regarding

TABLE. Strengths and Limitations of Existing Literature on MLOPs

Key Topic	Strengths	Weaknesses
ML model monitoring	● Relatively more mature literature with quantitative original research, including prospective data ● Open-source modules have been published	● Understandings of clinically significant model changes, as opposed to purely statistically significant changes, are still being defined
Automated retraining systems	● Sophisticated drift detection and retraining packages have been developed, quantitatively evaluated, and open-sourced	● Most of investigation has involved retrospective or simulated data ● Lack of consensus on which retraining systems to use
Ethics, equity, and bias	● Numerous experts have offered ethics and equity frameworks, which offer model developers a breadth of lenses to consider	● Currently, there is nearly limitless complexity and nuance in the ethics, equity, and bias issues discussed; literature should mature around the most relevant applications of ML
Clinical workflow integration	● The importance of workflow integration is frequently discussed ● Associated with the helpful “5 Rights” acronym[138]	● Limited standardization of workflow mapping and user interface best practices within MLOps literature to date
Infrastructure, human resources, and technology stack	● Despite relatively sparse literature, infrastructure requirements are rapidly standardizing, driven by industry developments	● Limited published data on real-world infrastructure requirements or on comparison between vendors of infrastructure components
Regulatory considerations	● Regulatory agencies are attempting to move quickly and proactively, with major regulatory bodies issuing briefings relevant to MLOps	● Pace of new AI discoveries continues to outpace regulation
Financial considerations	● Literature is rapidly evolving, with useful information coming from both academia and industry	● Very limited data from real-world clinical ML deployments

Abbreviations: AI, artificial intelligence; ML, machine learning; MLOPs, machine learning operations.

reimbursement and income, it is essential to determine who will fund a model's initial costs and maintenance, whether it is the implementing health care organization itself, an insurer, another payor, a grant-funding agency, or the consumer or patient. The sustainability of this funding source should also be considered in light of the necessary long-term demands of model maintenance and reassessment.

Secondary estimated cost savings can include increased provider or administrative efficiency, reduced hospitalizations, reduced length of stay, or other broader reductions in health care costs or utilization. As an example of secondary cost savings, 1 AI colorectal cancer diagnostic tool estimated a nearly $400 million cost savings in the United States by using a model to more accurately and rapidly identify who may be most likely to benefit from expensive cancer immunotherapies and who may not need them.[133] In another example, AI software for polysomnogram analysis, EnsoData, which is widely used, reports an average of 56% time savings for sleep analysts using their software, helping sleep specialty centers save on staffing expenses.[134] In a third example, another widely implemented ML health care solution, RapidAI, reported that interhospital transfers of patients for stroke care that were clinically unnecessary could be reduced by 92% with more accurate AI diagnostics, demonstrating their model's value through secondary cost savings.[135]

Next, although ML efforts have become almost ubiquitous in health care, it is worth noting that with increased competition for DevOps personnel hiring skilled staff can be expensive.[136] Consequently, the difficulty and expense of hiring personnel with the required skill set can lead to longer implementation times, poorer quality implementation, and increased overall long-term costs.[136]

Similarly, the role that clinicians play is also paramount in aiding the creation, troubleshooting, and implementation of MLOps. Lack of engagement has often led to underused ML tools, and neglecting their input can lead to ML solutions in search of problems to fix . Providing clinician subject matter experts with sufficient time to have input on an ML operationalization may be a costly but key ingredient to success.

Finally, it will be imperative for organizations to be wary of profit incentives to use clinical decision support systems for monetary gain via algorithms that may encourage additional costs, drugs, tests, diagnostics, or devices used without proportional benefit—and possibly even harm—to patient care. Indeed, AI technologies will not be immune to the tension between ethical medical care and generating profit.[137] Properly assessing the full costs of each MLOps endeavor—and ensuring they provide tangible and clear patient benefits—will be essential to the successful implementation of ML in health care.

DISCUSSION

In this review, we highlighted 7 key topic areas of emerging MLOps literature. We also attempted to present best practices, where available, and gaps in understanding, all of which may be useful to ML practitioners and consumers in health care. A key theme pervading all aspects of this review is the relative nascence of MLOps in health care, with few rigorous studies involving more downstream metrics such as patient outcomes and financial costs. This finding is consistent with an assessment by Zhang et al,[109] who noted that most published AI work to date is heavily "model-centric," focused on statistical assessments of model performance, rather than end-to-end understanding of AI tool impacts. We have summarized the strengths and weaknesses of existing literature across our 7 topic areas in Table.[138]

Our review was limited by its scoping, as opposed to a systematic approach. We chose this approach owing to the relative newness of the MLOps field, the ongoing evolution of terminology in this relatively new field of study, and the need to characterize the breadth of studies pertaining to this topic. We also did not include articles published in primary languages other than English. Nonetheless, we hope this review can serve as a useful guide to other ML practitioners.

CONCLUSION

Through this scoping review, we emphasize the critical importance of convergence on MLOps best practices and a better understanding of the current limitations in MLOps, which are essential as more ML models are introduced into live environments. We outline how MLOps literature continues to evolve, coalescing around 7 key content areas, namely (1) ML model monitoring, (2) automated retraining systems, (3) ethics, equity, and bias, (4) clinical workflow integration, (5) infrastructure, human resources and technology stack, (6) regulatory considerations, and (7) financial considerations. We note that MLOps studies using prospective data are still relatively sparse but suspect that this will evolve to include prospective evaluations of models' impacts on patient outcomes and care equity in addition to statistical model performance.

POTENTIAL COMPETING INTERESTS

Dr Kawamoto reports research grants and contracts involving machine learning from the National Institutes of Health and Hitachi Inc. All other authors report no competing interests.

ACKNOWLEDGMENTS

We thank Stephanie A. McFarland, MSIS, from the Plummer library at Mayo Clinic for her invaluable assistance with the literature search process.

SUPPLEMENTAL ONLINE MATERIAL

Supplemental material can be found online at https://www.mcpdigitalhealth.org/.

Supplemental material attached to journal articles has not been edited, and the authors take responsibility for the accuracy of all data.

Abbreviations and Acronyms: **AI,** artificial intelligence; **ML,** machine learning; **MLOps,** machine learning operations

Correspondence: Address to Alexander J Ryu, MD, Division of Hospital Internal Medicine, Department of Medicine, 200 1st ST SW, Mayo Clinic, Rochester, MN 55905 (ryu.alexander@mayo.edu).

ORCID

Alexander J. Ryu: https://orcid.org/0000-0002-0138-5112

REFERENCES

1. Kreuzberger D, Kühl N, Hirschl S. Machine learning operations (MLOps): overview, definition, and architecture. *IEEE Access*. 2023;11:31866-31879. https://doi.org/10.1109/ACCESS.2023.3262138.
2. Treveil M, Omont N, Stenac C, et al. *Introducing MLOps*. O'Reilly Media; 2020.
3. Mello MM, Guha N. Understanding liability risk from using health care artificial intelligence tools. *N Engl J Med*. 2024;390(3):271-278. https://doi.org/10.1056/NEJMhle2308901.
4. 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. https://doi.org/10.1126/science.aax2342.
5. Ebert C, Gallardo G, Hernantes J, Serrano N. DevOps. *IEEE Softw*. 2016;33(3):94-100. https://doi.org/10.1109/MS.2016.68.
6. Stirbu V, Granlund T, Mikkonen T. Continuous design control for machine learning in certified medical systems. *Softw. Qual J*. 2023;31(2):307-333. https://doi.org/10.1007/s11219-022-09601-5.
7. Sendak M, Gao M, Nichols M, Lin A, Balu S. Machine learning in health care: a critical appraisal of challenges and opportunities. *EGEMS (Wash DC)*. 2019;7(1):1. https://doi.org/10.5334/egems.287.
8. de Hond AAH, Leeuwenberg AM, Hooft L, et al. Guidelines and quality criteria for artificial intelligence-based prediction models in healthcare: a scoping review. *NPJ Digit Med*. 2022;5(1):1-13. https://doi.org/10.1038/s41746-021-00549-7.
9. Kelly CJ, Karthikesalingam A, Suleyman M, Corrado G, King D. Key challenges for delivering clinical impact with artificial intelligence. *BMC Med*. 2019;17(1):195. https://doi.org/10.1186/s12916-019-1426-2.
10. Doyen S, Dadario NB. 12 Plagues of AI in healthcare: a practical guide to current issues with using machine learning in a medical context. *Front Digit Health*. 2022;4:765406.
11. Rahmani K, Thapa R, Tsou P, et al. Assessing the effects of data drift on the performance of machine learning models used in clinical sepsis prediction. *Int J Med Inform*. 2023;173:104930. https://doi.org/10.1016/j.ijmedinf.2022.104930.
12. Sahiner B, Chen W, Samala RK, Petrick N. Data drift in medical machine learning: implications and potential remedies. *Br J Radiol*. 2023;96(1150):20220878. https://doi.org/10.1259/bjr.20220878.
13. Veeranki SPK, Kramer D, Hayn D, et al. Is regular re-training of a predictive delirium model necessary after deployment in routine care? *Stud Health Technol Inform*. 2019;260:186-191.
14. Moreno-Torres JG, Raeder T, Alaiz-Rodríguez R, Chawla NV, Herrera F. A unifying view on dataset shift in classification. *Pattern Recognit*. 2012;45(1):521-530. https://doi.org/10.1016/j.patcog.2011.06.019.
15. Singer M, Deutschman CS, Seymour CW, et al. The Third International consensus definitions for sepsis and septic shock (Sepsis-3). *JAMA*. 2016;315(8):801-810. https://doi.org/10.1001/jama.2016.0287.
16. Duckworth C, Chmiel FP, Burns DK, et al. Using explainable machine learning to characterise data drift and detect emergent health risks for emergency department admissions during COVID-19. *Sci Rep*. 2021;11(1):23017. https://doi.org/10.1038/s41598-021-02481-y.
17. Vasilev Y, Vladzymyrskyy A, Arzamasov K, et al. Clinical application of radiological AI for pulmonary nodule evaluation: replicability and susceptibility to the population shift caused by the COVID-19 pandemic. *Int J Med Inform*. 2023;178:105190. https://doi.org/10.1016/j.ijmedinf.2023.105190.
18. Vela D, Sharp A, Zhang R, Nguyen T, Hoang A, Pianykh OS. Temporal quality degradation in AI models. *Sci Rep*. 2022;12(1):11654. https://doi.org/10.1038/s41598-022-15245-z.
19. Ryu AJ, Ayanian S, Qian R, et al. A Clinician's guide to running custom machine-learning models in an electronic health record environment. *Mayo Clin Proc*. 2023;98(3):445-450. https://doi.org/10.1016/j.mayocp.2022.11.019.
20. van der Vegt AH, Scott IA, Dermawan K, Schnetler RJ, Kalke VR, Lane PJ. Implementation frameworks for end-to-end clinical AI: derivation of the SALIENT framework. *J Am Med Inform Assoc*. 2023;30(9):1503-1515. https://doi.org/10.1093/jamia/ocad088.
21. Adam GA, Chang CHK, Haibe-Kains B, Goldenberg A. Hidden risks of machine learning applied to healthcare: unintended feedback loops between models and future data causing model degradation. In: *Proceedings of the 5th Machine Learning for Healthcare Conference*. PMLR; 2020:710-731. https://proceedings.mlr.press/v126/adam20a.html. Accessed May 16, 2024.
22. Vaid A, Sawant A, Suarez-Farinas M, et al. Implications of the use of artificial intelligence predictive models in health care settings : a simulation study. *Ann Intern Med*. 2023;176(10):1358-1369. https://doi.org/10.7326/M23-0949.
23. Bai E, Song SL, Fraser HSF, Ranney ML. A graphical toolkit for longitudinal dataset maintenance and predictive model training in health care. *Appl Clin Inform*. 2022;13(1):56-66. https://doi.org/10.1055/s-0041-1740923.
24. Waring J, Lindvall C, Umeton R. Automated machine learning: review of the state-of-the-art and opportunities for healthcare. *Artif Intell Med*. 2020;104:101822. https://doi.org/10.1016/j.artmed.2020.101822.
25. Davis SE, Greevy RA, Lasko TA, Walsh CG, Matheny ME. Detection of calibration drift in clinical prediction models to inform model updating. *J Biomed Inform*. 2020;112:103611. https://doi.org/10.1016/j.jbi.2020.103611.
26. Levy TJ, Coppa K, Cang J, et al. Development and validation of self-monitoring auto-updating prognostic models of survival for hospitalized COVID-19 patients. *Nat Commun*. 2022;13(1):6812. https://doi.org/10.1038/s41467-022-34646-2.
27. Davis SE, Greevy RA Jr, Fonnesbeck C, Lasko TA, Walsh CG, Matheny ME. A nonparametric updating method to correct clinical prediction model drift. *J Am Med Inform Assoc*. 2019;26(12):1448-1457. https://doi.org/10.1093/jamia/ocz127.
28. Davis SE, Lasko TA, Chen G, Siew ED, Matheny ME. Calibration drift in regression and machine learning models for acute kidney injury. *J Am Med Inform Assoc*. 2017;24(6):1052-1061. https://doi.org/10.1093/jamia/ocx030.
29. Del Fiol G, Haug PJ. Infobuttons and classification models: a method for the automatic selection of on-line information resources to fulfill clinicians' information needs. *J Biomed Inform*. 2008;41(4):655-666. https://doi.org/10.1016/j.jbi.2007.11.007.
30. Chen J, Zheng Y, Liang Y, et al. Edge2Analysis: a novel AIoT platform for atrial fibrillation recognition and detection. *IEEE J Biomed Health Inform*. 2022;26(12):5772-5782. https://doi.org/10.1109/JBHI.2022.3171918.

31. Toor AA, Usman M, Younas F, M Fong AC, Khan SA, Fong S. Mining massive E-health data streams for IoMT enabled healthcare systems. *Sensors*. 2020;20(7):2131. https://doi.org/10.3390/s20072131.

32. Zhang X, Xue Y, Su X, et al. A transfer learning approach to correct the temporal performance drift of clinical prediction models: retrospective cohort study. *JMIR Med Inform*. 2022; 10(11):e38053. https://doi.org/10.2196/38053.

33. Guo LL, Pfohl SR, Fries J, et al. Evaluation of domain generalization and adaptation on improving model robustness to temporal dataset shift in clinical medicine. *Sci Rep*. 2022; 12(1):2726. https://doi.org/10.1038/s41598-022-06484-1.

34. Ethics and governance of artificial intelligence for health: WHO guidance Executive summary. World Health Organization. https://www.who.int/publications-detail-redirect/9789240037403. Accessed October 15, 2023.

35. Wiens J, Saria S, Sendak M, et al. Do no harm: a roadmap for responsible machine learning for health care. *Nat Med*. 2019; 25(9):1337-1340. https://doi.org/10.1038/s41591-019-0548-6.

36. Mehrabi N, Morstatter F, Saxena N, Lerman K, Galstyan A. A survey on bias and fairness in machine learning. *ACM Comput Surv*. 2019;54(6):1-35. https://doi.org/10.1145/3457607.

37. Hardt M, Price E, Price E, Srebro N. Equality of opportunity in supervised learning. In: Advances in Neural Information Processing Systems. Vol 29. Curran Associates. 2016. https://proceedings.neurips.cc/paper_files/paper/2016/hash/9d2682367c3935defcb1f9e247a97c0d-Abstract.html. Accessed June 27, 2024.

38. Grote T, Keeling G. Enabling fairness in healthcare through machine learning. *Ethics Inf Technol*. 2022;24(3):39. https://doi.org/10.1007/s10676-022-09658-7.

39. Rajkomar A, Hardt M, Howell MD, Corrado G, Chin MH. Ensuring fairness in machine learning to advance health equity. *Ann Intern Med*. 2018;169(12):866-872. https://doi.org/10.7326/M18-1990.

40. Rueda J, Rodríguez JD, Jounou IP, Hortal-Carmona J, Ausín T, Rodríguez-Arias D. "Just" accuracy? Procedural fairness demands explainability in AI-based medical resource allocations. *AI Soc*. Published online December 21, 2022. https://doi.org/10.1007/s00146-022-01614-9.

41. Fletcher RR, Nakeshimana A, Olubeko O. Addressing fairness, bias, and appropriate use of artificial intelligence and machine learning in global health. *Front Artif Intell*. 2021;3:561802. https://doi.org/10.3389/frai.2020.561802.

42. Char DS, Abràmoff MD, Feudtner C. Identifying ethical considerations for machine learning healthcare applications. *Am J Bioeth*. 2020;20(11):7-17. https://doi.org/10.1080/15265161.2020.1819469.

43. DeCamp M, Lindvall C. Latent bias and the implementation of artificial intelligence in medicine. *J Am Med Inform Assoc*. 2020; 27(12):2020-2023. https://doi.org/10.1093/jamia/ocaa094.

44. McCradden MD, Anderson JA, A Stephenson E, et al. A research ethics framework for the clinical translation of healthcare machine learning. *Am J Bioeth*. 2022;22(5):8-22. https://doi.org/10.1080/15265161.2021.2013977.

45. Kleppe A, Skrede OJ, De Raedt S, Liestøl K, Kerr DJ, Danielsen HE. Designing deep learning studies in cancer diagnostics. *Nat Rev Cancer*. 2021;21(3):199-211. https://doi.org/10.1038/s41568-020-00327-9.

46. Drukker K, Chen W, Gichoya J, et al. Toward fairness in artificial intelligence for medical image analysis: identification and mitigation of potential biases in the roadmap from data collection to model deployment. *J Med Imaging (Bellingham)*. 2023; 10(6):061104. https://doi.org/10.1117/1.JMI.10.6.061104.

47. Yang J, Soltan AAS, Eyre DW, Yang Y, Clifton DA. An adversarial training framework for mitigating algorithmic biases in clinical machine learning. *NPJ Digit Med*. 2023;6(1):55. https://doi.org/10.1038/s41746-023-00805-y.

48. Cohen IG, Amarasingham R, Shah A, Xie B, Lo B. The legal and ethical concerns that arise from using complex predictive analytics in health care. *Health Aff (Millwood)*. 2014;33(7): 1139-1147. https://doi.org/10.1377/hlthaff.2014.0048.

49. Pleiss G, Raghavan M, Wu F, Kleinberg J, Weinberger KQ. *On fairness and calibration*. In: *Advances in Neural Information Processing Systems*. Curran Associates; 2017;30. https://proceedings.neurips.cc/paper_files/paper/2017/hash/b8b9c74ac526fffbeb2d39ab038d1cd7-Abstract.html. Accessed November 17, 2023.

50. Chomutare T, Tejedor M, Svenning TO, et al. Artificial intelligence implementation in healthcare: a theory-based scoping review of barriers and facilitators. *Int J Environ Res Public Health*. 2022;19(23):16359. https://doi.org/10.3390/ijerph192316359.

51. Yogarajan V, Dobbie G, Leitch S, et al. Data and model bias in artificial intelligence for healthcare applications in New Zealand. *Front Comput Sci*. 2022;4. https://doi.org/10.3389/fcomp.2022.1070493.

52. Meetings examine impact of healthcare algorithms on racial and ethnic disparities in health and healthcare. Effective Health Care (EHC) Program. https://effectivehealthcare.ahrq.gov/news/meetings. Accessed December 22, 2023.

53. Kawamoto K, Finkelstein J, Del Fiol G. Implementing machine learning in the electronic health record: checklist of essential considerations. *Mayo Clin Proc*. 2023;98(3):366-369. https://doi.org/10.1016/j.mayocp.2023.01.013.

54. Watson J, Hutyra CA, Clancy SM, et al. Overcoming barriers to the adoption and implementation of predictive modeling and machine learning in clinical care: what can we learn from US academic medical centers? *JAMIA Open*. 2020;3(2): 167-172. https://doi.org/10.1093/jamiaopen/ooz046.

55. Schaarup JFR, Aggarwal R, Dalsgaard EM, et al. Perception of artificial intelligence-based solutions in healthcare among people with and without diabetes: a cross-sectional survey from the health in Central Denmark cohort. *Diabetes Epidemiol Manag*. 2023;9:100114. https://doi.org/10.1016/j.deman.2022.100114.

56. McIntosh C, Conroy L, Tjong MC, et al. Clinical integration of machine learning for curative-intent radiation treatment of patients with prostate cancer. *Nat Med*. 2021;27(6):999-1005. https://doi.org/10.1038/s41591-021-01359-w.

57. Wang SM, Hogg HDJ, Sangvai D, et al. Development and integration of machine learning algorithm to identify peripheral arterial disease: multistakeholder qualitative study. *JMIR Form Res*. 2023;7:e43963. https://doi.org/10.2196/43963.

58. Ng R, Tan KB. Implementing an individual-centric discharge process across Singapore public hospitals. *Int J Environ Res Public Health*. 2021;18(16):8700. https://doi.org/10.3390/ijerph18168700.

59. Engstrom CJ, Adelaine S, Liao F, Jacobsohn GC, Patterson BW. Operationalizing a real-time scoring model to predict fall risk among older adults in the emergency department. *Front Digit Health*. 2022;4:958663. https://www.frontiersin.org/articles/10.3389/fdgth.2022.958663. Accessed July 28, 2023.

60. Moorman LP. Principles for real-world implementation of bedside predictive analytics monitoring. *Appl Clin Inform*. 2021;12(4):888-896. https://doi.org/10.1055/s-0041-1735183.

61. Wolff RF, Moons KGM, Riley RD, et al. PROBAST: a tool to assess the risk of bias and applicability of prediction model studies. *Ann Intern Med*. 2019;170(1):51-58. https://doi.org/10.7326/M18-1376.

62. Collins GS, Reitsma JB, Altman DG, Moons KGM. Transparent reporting of a multivariable prediction model for individual prognosis or diagnosis (TRIPOD): the TRIPOD statement. *Ann Intern Med*. 2015;162(1):55-63. https://doi.org/10.7326/M14-0697.

63. Vasey B, Nagendran M, Campbell B, et al. Reporting guideline for the early-stage clinical evaluation of decision support systems driven by artificial intelligence: DECIDE-AI. *Nat Med*. 2022;28(5):924-933. https://doi.org/10.1038/s41591-022-01772-9.

64. Reddy S, Rogers W, Makinen VP, et al. Evaluation framework to guide implementation of AI systems into healthcare settings. *BMJ Health Care Inform*. 2021;28(1):e100444. https://doi.org/10.1136/bmjhci-2021-100444.

65. Bartels R, Dudink J, Haitjema S, Oberski D, van 't Veen A. A perspective on a quality management system for AI/ML-based clinical decision support in hospital care. *Front Digit Health*. 2022;4:942588. https://doi.org/10.3389/fdgth.2022.942588.

66. Cutillo CM, Sharma KR, Foschini L, Kundu S, Mackintosh M, Mandl KD, et al. Machine intelligence in healthcare-perspectives on trustworthiness, explainability, usability, and transparency. *NPJ Digit Med*. 2020;3:47. https://doi.org/10.1038/s41746-020-0254-2.

67. Cai CJ, Winter S, Steiner D, Wilcox L, Terry M. "Hello AI": uncovering the onboarding needs of medical practitioners for human-AI collaborative decision-making. *Proc ACM Hum-Comput Interact*. 2019;3(CSCW):1-24. https://doi.org/10.1145/3359206.

68. Verma AA, Murray J, Greiner R, et al. Implementing machine learning in medicine. *CMAJ*. 2021;193(34):E1351-E1357. https://doi.org/10.1503/cmaj.202434.

69. Sendak MP, Ratliff W, Sarro D, et al. Real-world integration of a sepsis deep learning technology into routine clinical care: implementation study. *JMIR Med Inform*. 2020;8(7):e15182. https://doi.org/10.2196/15182.

70. Hribar MR, Read-Brown S, Goldstein IH, et al. Secondary use of electronic health record data for clinical workflow analysis. *J Am Med Inform Assoc*. 2018;25(1):40-46. https://doi.org/10.1093/jamia/ocx098.

71. Vankipuram A, Traub S, Patel VL. A method for the analysis and visualization of clinical workflow in dynamic environments. *J Biomed Inform*. 2018;79:20-31. https://doi.org/10.1016/j.jbi.2018.01.007.

72. Jarab AS, Al-Qerem W, Alzoubi KH, et al. Artificial intelligence in pharmacy practice: attitude and willingness of the community pharmacists and the barriers for its implementation. *Saudi Pharm J*. 2023;31(8):101700. https://doi.org/10.1016/j.jsps.2023.101700.

73. Neher M, Petersson L, Nygren JM, Svedberg P, Larsson I, Nilsen P. Innovation in healthcare: leadership perceptions about the innovation characteristics of artificial intelligence-a qualitative interview study with healthcare leaders in Sweden. *Implement Sci Commun*. 2023;4(1):81. https://doi.org/10.1186/s43058-023-00458-8.

74. Akudjedu TN, Torre S, Khine R, Katsifarakis D, Newman D, Malamateniou C. Knowledge, perceptions, and expectations of artificial intelligence in radiography practice: a global radiography workforce survey. *J Med Imaging Radiat Sci*. 2023;54(1):104-116. https://doi.org/10.1016/j.jmir.2022.11.016.

75. Al-Medfa MK, Al-Ansari AMS, Darwish AH, Qreeballa TA, Jahrami H. Physicians' attitudes and knowledge toward artificial intelligence in medicine: benefits and drawbacks. *Heliyon*. 2023;9(4):e14744. https://doi.org/10.1016/j.heliyon.2023.e14744.

76. Chen Y, Wu Z, Wang P, et al. Radiology residents' perceptions of artificial intelligence: nationwide cross-sectional survey study. *J Med Internet Res*. 2023;25:e48249. https://doi.org/10.2196/48249.

77. Tanaka M, Matsumura S, Bito S. Roles and competencies of doctors in artificial intelligence implementation: qualitative analysis through physician interviews. *JMIR Form Res*. 2023;7:e46020. https://doi.org/10.2196/46020.

78. Ahmad MN, Abdallah SA, Abbasi SA, Abdallah AM. Student perspectives on the integration of artificial intelligence into healthcare services. *Digit Health*. 2023;9:20552076231174095. https://doi.org/10.1177/20552076231174095.

79. Abuzaid MM, Elshami W, Fadden SM. Integration of artificial intelligence into nursing practice. *Health Technol (Berl)*. 2022;12(6):1109-1115. https://doi.org/10.1007/s12553-022-00697-0.

80. Sezgin E, Sirrianni J, Linwood SL. Operationalizing and implementing pretrained, large artificial intelligence linguistic models in the US health care system: outlook of generative pretrained transformer 3 (GPT-3) as a service model. *JMIR Med Inform*. 2022;10(2):e32875. https://doi.org/10.2196/32875.

81. Hogg HDJ, Al-Zubaidy M; Technology Enhanced Macular Services Study Reference Group, Talks J, Denniston AK, Kelly CJ, et al. Stakeholder perspectives of clinical artificial intelligence implementation: systematic review of qualitative evidence. *J Med Internet Res*. 2023;25(1):e39742. https://doi.org/10.2196/39742.

82. Kruse CS, Smith B, Vanderlinden H, Nealand A. Security techniques for the electronic health records. *J Med Syst*. 2017;41(8):127. https://doi.org/10.1007/s10916-017-0778-4.

83. Price WN, Cohen IG. Privacy in the age of medical big data. *Nat Med*. 2019;25(1):37-43. https://doi.org/10.1038/s41591-018-0272-7.

84. Marinescu DC. *Cloud Computing: Theory and Practice*. Morgan Kaufmann; 2022.

85. Darwish A, Hassanien AE, Elhoseny M, Sangaiah AK, Muhammad K. The impact of the hybrid platform of internet of things and cloud computing on healthcare systems: opportunities, challenges, and open problems. *J Ambient Intell Human Comput*. 2019;10(10):4151-4166. https://doi.org/10.1007/s12652-017-0659-1.

86. Kuo AM. Opportunities and challenges of cloud computing to improve health care services. *J Med Internet Res*. 2011;13(3):e67. https://doi.org/10.2196/jmir.1867.

87. Hummer W, Muthusamy V, Rausch T, et al. ModelOps: cloud-based lifecycle management for reliable and trusted AI. In: *2019 IEEE International Conference on Cloud Engineering (IC2E)*. IEEE; 2019:113-120. https://doi.org/10.1109/IC2E.2019.00025.

88. Kwasniewska A, Raghava S, Davila C, Sevenier M, Gamba D, Ruminski J. Preferred benchmarking criteria for systematic taxonomy of embedded platforms (STEP) in human system interaction systems. In: *2022 15th International Conference on Human System Interaction (HSI)*. IEEE; 2022:1-7. https://doi.org/10.1109/HSI55341.2022.9869470.

89. Pianykh OS, Guitron S, Parke D, et al. Improving healthcare operations management with machine learning. *Nat Mach Intell*. 2020;2(5):266-273. https://doi.org/10.1038/s42256-020-0176-3.

90. Baier L, Jöhren F, Seebacher S. *Challenges in the deployment and operation of machine learning in practice*. Stockholm: Presented at: ECIS 2019—27th European Conference on Information Systems; May 2019.

91. Vassiliadis P, Simitsis A, Skiadopoulos S. Conceptual modeling for ETL processes. In: Proceedings of the 5th ACM International Workshop on Data Warehousing and OLAP. *DOLAP '02*. Association for Computing Machinery; 2002:14-21. https://doi.org/10.1145/583890.583893.

92. Garg S, Pundir P, Rathee G, Gupta PK, Garg S, Ahlawat S. On continuous integration/continuous delivery for automated deployment of machine learning models using MLOps. In: *2021 IEEE Fourth International Conference on Artificial Intelligence and Knowledge Engineering (AIKE)*. IEEE; 2021:25-28. https://doi.org/10.1109/AIKE52691.2021.00010.

93. Granlund T, Stirbu V, Mikkonen T. Towards regulatory-compliant MLOps: Oravizio's journey from a machine learning experiment to a deployed certified medical product. *SN Comput Sci*. 2021;2(5):342. https://doi.org/10.1007/s42979-021-00726-1.

94. Kleftakis S, Mavrogiorgou A, Mavrogiorgos K, Kiourtis A, Kyriazis D. Digital twin in healthcare through the eyes of the vitruvian man. In: Chen YW, Tanaka S, Howlett RJ, Jain LC,

eds. Innovation in Medicine and Healthcare. *Smart Innovation, Systems and Technologies*. Springer Nature; 2022:75-85. https://doi.org/10.1007/978-981-19-3440-7_7.

95. Soh J, Singh P. Machine learning operations. In: Soh J, Singh P, eds. *Data Science Solutions on Azure: Tools and Techniques Using Databricks and MLOps*. Apress; 2020:259-279. https://doi.org/10.1007/978-1-4842-6405-8_8.
96. Bhavsar K, Vishwakarma H, Pawar BR, Shinde S, Kimbahune S, Ghose A. A platform to enable algorithms as service model aimed at digital health service delivery. In: *2022 14th International Conference on Communication Systems & Networks (COMSNETS)*. IEEE; 2022:241-245. https://doi.org/10.1109/COMSNETS53615.2022.9668453.
97. Yoo J, Lee J, Min JY, et al. Development of an interoperable and easily transferable clinical decision support system deployment platform: system design and development study. *J Med Internet Res*. 2022;24(7):e37928. https://doi.org/10.2196/37928.
98. Cresswell K, Rigby M, Magrabi F, et al. The need to strengthen the evaluation of the impact of artificial intelligence-based decision support systems on healthcare provision. *Health Policy*. 2023;136:104889. https://doi.org/10.1016/j.healthpol.2023.104889.
99. Colantonio S, Berti A, Buongiorno R, et al. AI trustworthiness in prostate cancer imaging: a look at algorithmic and system transparency. In: *2023 IEEE EMBS Special Topic Conference on Data Science and Engineering in Healthcare, Medicine and Biology*. IEEE; 2023:79-80. https://doi.org/10.1109/IEEECONF58974.2023.10404432.
100. Epic unveils new app "showroom" for 3rd-party vendors. Fierce Healthcare. https://www.fiercehealthcare.com/health-tech/epic-unveils-new-app-showroom-third-party-vendors. Accessed May 16, 2024.
101. Tarabichi Y, Higginbotham J, Riley N, Kaelber DC, Watts B. Reducing disparities in no show rates using predictive model-driven live appointment reminders for at-risk patients: a randomized controlled quality improvement initiative. *J Gen Intern Med*. 2023;38(13):2921-2927. https://doi.org/10.1007/s11606-023-08209-0.
102. Assadi A, Laussen PC, Goodwin AJ, et al. An integration engineering framework for machine learning in healthcare. *Front Digit Health*. 2022;4:932411. https://doi.org/10.3389/fdgth.2022.932411.
103. Sperrin M, Martin GP, Sisk R, Peek N. Missing data should be handled differently for prediction than for description or causal explanation. *J Clin Epidemiol*. 2020;125:183-187. https://doi.org/10.1016/j.jclinepi.2020.03.028.
104. Li RC, Asch SM, Shah NH. Developing a delivery science for artificial intelligence in healthcare. *NPJ Digit Med*. 2020;3(1):1-3. https://doi.org/10.1038/s41746-020-00318-y.
105. Wiesenfeld BM, Aphinyanaphongs Y, Nov O. AI model transferability in healthcare: a sociotechnical perspective. *Nat Mach Intell*. 2022;4(10):807-809. https://doi.org/10.1038/s42256-022-00544-x.
106. Hofer IS, Burns M, Kendale S, Wanderer JP. Realistically integrating machine learning into clinical practice: a road map of opportunities, challenges, and a potential future. *Anesth Analg*. 2020;130(5):1115-1118. https://doi.org/10.1213/ANE.0000000000004575.
107. Shaw J, Rudzicz F, Jamieson T, Goldfarb A. Artificial intelligence and the implementation challenge. *J Med Internet Res*. 2019;21(7):e13659. https://doi.org/10.2196/13659.
108. Shah P, Kendall F, Khozin S, et al. Artificial intelligence and machine learning in clinical development: a translational perspective. *NPJ Digit Med*. 2019;2(1):69. https://doi.org/10.1038/s41746-019-0148-3.
109. Zhang J, Budhdeo S, William W, et al. Moving towards vertically integrated artificial intelligence development. *NPJ Digit Med*. 2022;5(1):143. https://doi.org/10.1038/s41746-022-00690-x.
110. Leming MJ, Bron EE, Bruffaerts R, et al. Challenges of implementing computer-aided diagnostic models for neuroimages in a clinical setting. *NPJ Digit Med*. 2023;6(1):129. https://doi.org/10.1038/s41746-023-00868-x.
111. Gerke S, Babic B, Evgeniou T, Cohen IG. The need for a system view to regulate artificial intelligence/machine learning-based software as medical device. *NPJ Digit Med*. 2020;3(1):53. https://doi.org/10.1038/s41746-020-0262-2.
112. US FDA. Proposed regulatory framework for modifications to Artificial Intelligence/Machine Learning (AI/ML)-based Software as a Medical Device (SaMD). Analysis and Policy Observatory. https://apo.org.au/node/228371. Accessed December 21, 2023.
113. White paper on artificial intelligence. a European approach to excellence and trust. European Commission. https://commission.europa.eu/document/d2ec4039-c5be-423a-81ef-b9e44e79825b_en. Accessed December 22, 2023.
114. Roberts H, Cowls J, Morley J, Taddeo M, Wang V, Floridi L. The Chinese approach to artificial intelligence: an analysis of policy, ethics, and regulation. *AI Soc*. 2021;36(1):59-77. https://doi.org/10.1007/s00146-020-00992-2.
115. Buiten MC. Towards intelligent regulation of artificial intelligence. *Eur J Risk Regulat*. 2019;10(1):41-59. https://doi.org/10.1017/err.2019.8.
116. White paper on artificial intelligence commission-white-paper-artificial-intelligence-feb2020_en.pdf. European Commission. https://commission.europa.eu/system/files/2020-02/commission-white-paper-artificial-intelligence-feb2020_en.pdf. Accessed December 21, 2023.
117. Artificial Intelligence Act, Corrigendum, 19 April 2024. European Union. https://artificialintelligenceact.eu/the-act/. Accessed July 4, 2024.
118. Nilsen P, Reed J, Nair M, et al. Realizing the potential of artificial intelligence in healthcare: learning from intervention, innovation, implementation and improvement sciences. *Front Health Serv*. 2022;2:961475. https://doi.org/10.3389/frhs.2022.961475.
119. Shashikumar SP, Amrollahi F, Nemati S. Unsupervised detection and correction of model calibration shift at test-time. *Annu Int Conf IEEE Eng Med Biol Soc*. 2023;2023:1-4. https://doi.org/10.1109/EMBC40787.2023.10341086.
120. US FDA. Premarket Notification 510(k). Food and Drug Administration. https://www.fda.gov/medical-devices/premarket-submissions-selecting-and-preparing-correct-submission/premarket-notification-510k. Accessed December 22, 2023.
121. McKee M, Wouters OJ. The challenges of regulating artificial intelligence in healthcare comment on "clinical decision support and new regulatory frameworks for medical devices: are we ready for it? - a viewpoint paper". *Int J Health Policy Manag*. 2023;12:7261. https://doi.org/10.34172/ijhpm.2022.7261.
122. Pica F. AI, genetic testing: what washed up at Epic's global conference. The Cap Times. https://captimes.com/news/business/ai-genetic-testing-what-washed-up-at-epics-global-conference/article_4cc21fb9-79f6-533a-b4f7-300fddcce032.html. Accessed December 21, 2023.
123. Mayo Clinic platform expands distributed data network, partnerships. TechTarget. https://healthitanalytics.com/news/mayo-clinic-platform-expands-distributed-data-network-partnerships. Accessed December 21, 2023.
124. Liao F, Adelaine S, Afshar M, Patterson BW. Governance of clinical AI applications to facilitate safe and equitable deployment in a large health system: key elements and early successes. *Front Digit Health*. 2022;4:931439. https://doi.org/10.3389/fdgth.2022.931439.
125. Naik N, Hameed BMZ, Shetty DK, et al. Legal and ethical consideration in artificial intelligence in healthcare: who takes responsibility? *Front Surg*. 2022;9:862322. https://doi.org/10.3389/fsurg.2022.862322.
126. Grunhut J, Wyatt AT, Marques O. Educating future physicians in artificial intelligence (AI): an integrative review and proposed changes. *J Med Educ Curric Dev*. 2021;8:23821205211036836. https://doi.org/10.1177/23821205211036836.

127. Tolentino R, Baradaran A, Gore G, Pluye P, Abbasgholizadeh-Rahimi S. Curriculum frameworks and educational programs in artificial intelligence for medical students, residents, and practicing physicians: a scoping review protocol. *JBI Evid Synth*. 2023;21(7):1477-1484. https://doi.org/10.11124/JBIES-22-00374.

128. Strubell E, Ganesh A, McCallum A. Energy and policy considerations for deep learning in NLP. *Preprint.* Posted online June 5, 2019. https://doi.org/10.48550/arXiv.1906.02243.

129. Sharir O, Peleg B, Shoham Y. The cost of training NLP models: a concise overview. *Preprint.* Published online April 19, 2020. https://doi.org/10.48550/arXiv.2004.08900.

130. Khanna NN, Maindarkar MA, Viswanathan V, et al. Economics of artificial intelligence in healthcare: diagnosis vs. treatment. *Healthcare (Basel)*. 2022;10(12):2493. https://doi.org/10.3390/healthcare10122493.

131. Coop R. What is the cost to deploy and maintain a machine learning model? phData. https://www.phdata.io/blog/what-is-the-cost-to-deploy-and-maintain-a-machine-learning-model/. Accessed October 17, 2023.

132. Microsoft to charge more for AI in office, secure Bing from leaks. Reuters. https://www.reuters.com/technology/microsoft-charge-more-ai-office-secure-bing-leaks-2023-07-18/. Accessed October 17, 2023.

133. Kacew AJ, Strohbehn GW, Saulsberry L, et al. Artificial intelligence can cut costs while maintaining accuracy in colorectal cancer genotyping. *Front Oncol*. 2021;11:630953. https://doi.org/10.3389/fonc.2021.630953.

134. AI scoring for PSGs and HSATs. EnsoData. https://www.ensodata.com/ai-scoring/. Accessed October 17, 2023.

135. Inc iSchemaView. Stroke patient care. Rapid stroke. https://www.rapidai.com/stroke. Accessed October 17, 2023.

136. Integrating DevOps with existing healthcare IT infrastructure and processes: challenges and key considerations. Empirical quests for management essences. https://researchberg.com/index.php/eqme/article/view/98. Accessed October 17, 2023.

137. He J, Baxter SL, Xu J, Xu J, Zhou X, Zhang K. The practical implementation of artificial intelligence technologies in medicine. *Nat Med*. 2019;25(1):30-36. https://doi.org/10.1038/s41591-018-0307-0.

138. Campbell RJ. The five "rights" of clinical decision support. *J AHIMA*. 2013;84(10):42-47 [quiz 48].