JAMIA Open, 2024, **7(4)**, ooae108
https://doi.org/10.1093/jamiaopen/ooae108
Research and Applications

Research and Applications

Addressing ethical issues in healthcare artificial intelligence using a lifecycle-informed process

Benjamin X. Collins, MD[1,2,3,*], Jean-Christophe Bélisle-Pipon, PhD[4], Barbara J. Evans, JD, PhD[5,6], Kadija Ferryman, PhD[7], Xiaoqian Jiang, PhD[8], Camille Nebeker, EdD[9], Laurie Novak, PhD[1], Kirk Roberts, PhD[8], Martin Were, MD[1,3], Zhijun Yin, PhD[1,10], Vardit Ravitsky, PhD[11], Joseph Coco, MS[1], Rachele Hendricks-Sturrup, DHSc[12,13], Ishan Williams, PhD[14], Ellen W. Clayton, MD, JD[2,15], Bradley A. Malin, PhD[1,10]; on behalf of the Bridge2AI Ethics and Trustworthy AI Working Group

[1]Department of Biomedical Informatics, Vanderbilt University Medical Center, Nashville, TN 37203, United States, [2]Center for Biomedical Ethics and Society, Vanderbilt University Medical Center, Nashville, TN 37203, United States, [3]Department of Medicine, Vanderbilt University Medical Center, Nashville, TN 37232, United States, [4]Faculty of Health Sciences, Simon Fraser University, Burnaby, British Columbia V5A 1S6, Canada, [5]Levin College of Law, University of Florida, Gainesville, FL 32611, United States, [6]Herbert Wertheim College of Engineering, University of Florida, Gainesville, FL 32611, United States, [7]Berman Institute of Bioethics, Johns Hopkins University, Baltimore, MD 21205, United States, [8]McWilliams School of Biomedical Informatics, UTHealth Houston, Houston, TX 77030, United States, [9]Herbert Wertheim School of Public Health and Human Longevity Science, University of California, San Diego, La Jolla, CA 92093, United States, [10]Department of Computer Science, Vanderbilt University, Nashville, TN 37212, United States, [11]The Hastings Center, Garrison, NY 10524, United States, [12]National Alliance against Disparities in Patient Health, Woodbridge, VA 22191, United States, [13]Margolis Center for Health Policy, Duke University, Washington, DC 20004, United States, [14]School of Nursing, University of Virginia, Charlottesville, VA 22903, United States, [15]Law School, Vanderbilt University, Nashville, TN 37203, United States

*Corresponding author: Benjamin X. Collins, MD, Vanderbilt University Medical Center, Department of Biomedical Informatics, 2525 West End Ave. Suite 1475, Nashville, TN 37203, United States (benjamin.collins@vumc.org)

Abstract

Objectives: Artificial intelligence (AI) proceeds through an iterative and evaluative process of development, use, and refinement which may be characterized as a lifecycle. Within this context, stakeholders can vary in their interests and perceptions of the ethical issues associated with this rapidly evolving technology in ways that can fail to identify and avert adverse outcomes. Identifying issues throughout the AI lifecycle in a systematic manner can facilitate better-informed ethical deliberation.

Materials and Methods: We analyzed existing lifecycles from within the current literature for ethical issues of AI in healthcare to identify themes, which we relied upon to create a lifecycle that consolidates these themes into a more comprehensive lifecycle. We then considered the potential benefits and harms of AI through this lifecycle to identify ethical questions that can arise at each step and to identify where conflicts and errors could arise in ethical analysis. We illustrated the approach in 3 case studies that highlight how different ethical dilemmas arise at different points in the lifecycle.

Results, Discussion, and Conclusion: Through case studies, we show how a systematic lifecycle-informed approach to the ethical analysis of AI enables mapping of the effects of AI onto different steps to guide deliberations on benefits and harms. The lifecycle-informed approach has broad applicability to different stakeholders and can facilitate communication on ethical issues for patients, healthcare professionals, research participants, and other stakeholders.

Lay Summary

The steps of artificial intelligence in healthcare can be described and visualized from development to widespread use as a lifecycle. Viewing healthcare artificial intelligence (AI) as a lifecycle can help identify and address the ethical issues that arise at each step. This approach addresses barriers in healthcare AI resulting from how people in different positions vary in their interests and perceptions related to AI, which can lead to challenges in how ethical issues are addressed. In this paper, we have built a sample lifecycle using the elements of previously existing lifecycles, filling in gaps to be more comprehensive. We then show how ethical issues arise at different points and present example questions that should be considered throughout the lifecycle. We then use the process of examining ethical issues through the lens of the lifecycle and apply it to 3 case studies. This approach to examining the ethical issues can be applied by people in various positions, including developers of an AI system to end-users, to facilitate communication among different stakeholders to anticipate and address problems.

Key words: artificial intelligence; ethics; healthcare.

Received: February 17, 2024; **Revised:** August 19, 2024; **Editorial Decision:** October 2, 2024; **Accepted:** October 4, 2024

Introduction

Artificial intelligence (AI) offers many potential benefits in healthcare, including improved diagnosis and prognosis, greater decision-making efficiency, and decreased provider work burden, among others.[1–3] Conversely, AI can induce risks of harm that include exacerbating health inequity,[4] introducing inaccurate data,[5] and misuse to supply falsified research as evidence, all with consequences that may injure patients and undermine public trust.[6] The possible benefits and harms are interwoven and requires attention to who weighs risks and benefits as well as how and when they do so. Adding to the complexity, stakeholders may perceive the potential benefits and risks in different ways, creating challenges in how ethical issues are conceptualized and discussed. These differences in perspective can obfuscate the ability of people and organizations to see and understand the full array of effects, ultimately limiting their capacity to maximize the benefits and minimize or mitigate the harms of AI.

In light of the challenges confronted by stakeholders when thinking about the ethical issues, optimizing healthcare AI requires a clear recognition of: (1) when various benefits and risks are salient and to whom; (2) steps that can be taken to maximize benefits and forestall future harms; and (3) how to communicate clearly among various stakeholders about those benefits and harms, recognizing that tradeoffs are often required. The following examples present tradeoffs to consider. Algorithm transparency for healthcare and patient users can disclose information that can be exploited by malicious actors to craft attacks to poison AI models.[7] Strong consent rights may enhance personal control over one's data but at the same time may disproportionately allow individuals from marginalized demographics to avoid enrollment, and often reasonably so given historical injustices, resulting in racial and gender underrepresentation in training data.[8,9] Restricting data access can promote privacy while undermining regulatory AI safety surveillance.[10] Evaluating these ethical tradeoffs requires thoughtful deliberation in a multidisciplinary manner. Communication failures can breed confusion, independently cause patient injury, and contribute to harms.[11]

Ideally, technologies, including AI, go through an iterative process of identifying goals, development, use, and evaluation, which can be characterized as a lifecycle. Lifecycles can be used to identify problems that can arise along the way and opportunities for thoughtful design that anticipates the problems and for meaningful intervention. Thus, in this paper, we introduce an approach to ethical reasoning that maps the benefits and harms of healthcare AI onto the steps of an AI lifecycle. We consolidate features of existing lifecycles into a single framework as a platform to map the ethical issues of healthcare AI and to illustrate the overall value of a lifecycle-informed approach through case studies. We anticipate this approach will facilitate communication across multiple stakeholders and disciplines, enabling people and organizations involved in the development and implementation of AI to know what problems to look for and where to look, so they can begin to formulate solutions, including preventive measures and remediation strategies. We expect the use of the lifecycle approach to be incentivized by the potential to minimize downstream harm and to enable improved responsibility of healthcare organizations and technology collaborators for the AI tools they implement. Although many people are involved in, and are affected by, healthcare AI, including patients, research participants, family members, advocacy groups, researchers, clinicians, and healthcare institutions, we focus this demonstration of the lifecycle-informed approach on how patients are affected as healthcare organizations have strong incentives to prevent patient harm.

Methods

We reviewed the current literature to find existing lifecycles for ethical issues of healthcare AI in English within 1 biomedical focused database (PubMed) and 1 computing focused database (ACM Digital Library). To be considered a lifecycle, the framework needed to provide a chronological and iteratively recurring stepwise description of how a healthcare AI system moves from original conceptualization to widespread use and the major decision points within that process. We excluded lifecycles that were not specific to healthcare or did not prioritize ethics. Two authors independently analyzed the descriptions of individual steps in each existing lifecycle and grouped similar steps together as a theme, with discrepancies discussed until reaching an agreement.

A consolidated lifecycle was constructed through a mixed deductive-inductive process using the identified themes and an understanding of the course of AI development with relation to ethical issues, including those addressed in clinical bioethics, augmented by justice and equity and those presented specifically by the development and use of AI. We then detailed the use of the lifecycle to guide examining ethical issues related to the potential benefits and harms of healthcare AI through questions that may be considered at each step. This lifecycle-informed approach was then applied to 3 case studies.

Results

Our survey of the literature retrieved 13 healthcare AI lifecycles, which varied in the range of ethical issues they addressed and whether they had been implemented in practice (Supplementary Material S1).[12–24] We identified a total of 10 themes. Six themes reflect distinct steps, and four themes cut across the lifecycle. No single lifecycle covered all of the themes identified. We organized the themes into a consolidated lifecycle to guide thorough coverage of the possible ethical issues that may arise with the use of healthcare AI (Figure 1). Each step of the lifecycle (1-6) is represented by a node and progresses in a clockwise direction with cross-cutting themes (A-D) framed around the steps. The black arrows represent the general forward direction of the lifecycle, and the gray arrows represent decisions to return to a previous step. Each step and cross-cutting theme raise questions for ethical reflection that should be considered when thinking about the potential benefits and harms of healthcare AI. At any point, issues may arise that warrant returning to earlier steps, including further back than the immediately preceding step. A comprehensive list of questions that could arise is beyond the scope of this paper which is focused on the process of using a lifecycle tool, but sample questions are presented after the brief description of the lifecycle to provide examples of how one might think about the cross-cutting issues as they arise (Figure 2A) and the issues at each step (Figure 2B). Although some content overlap of the steps and cross-cutting themes occurs, the descriptions keep content

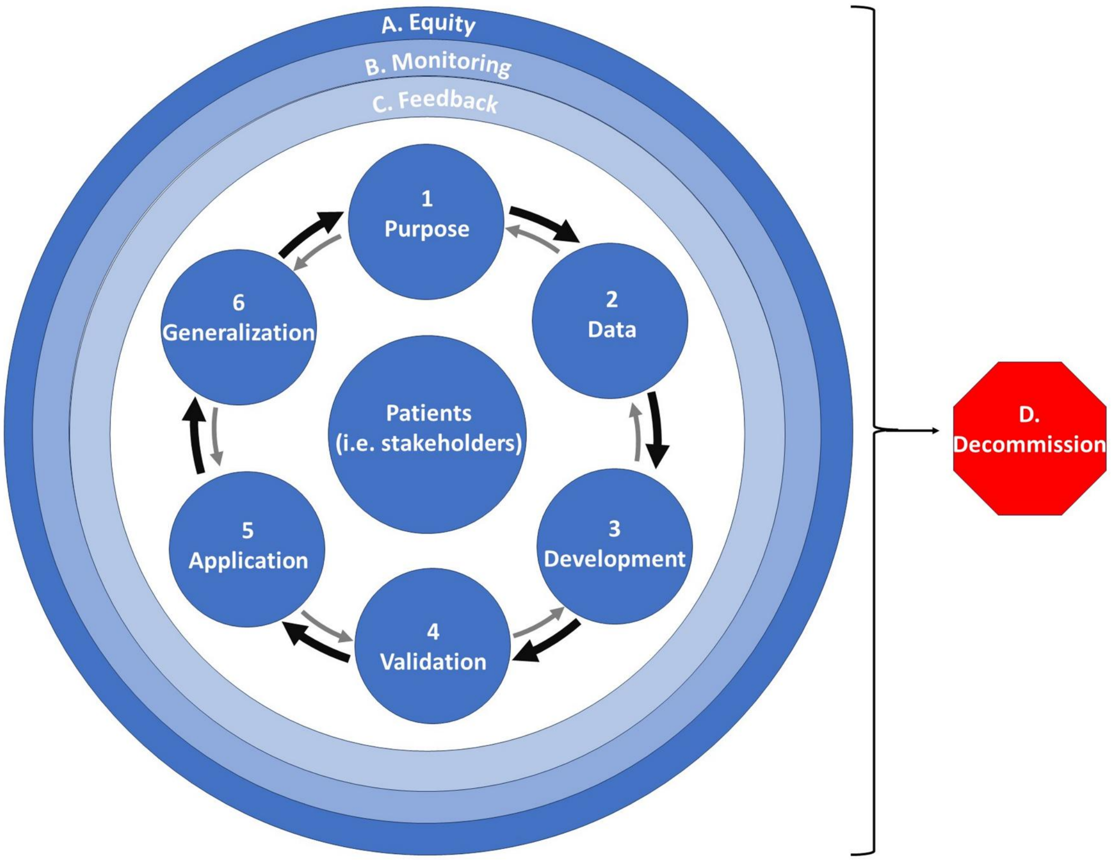
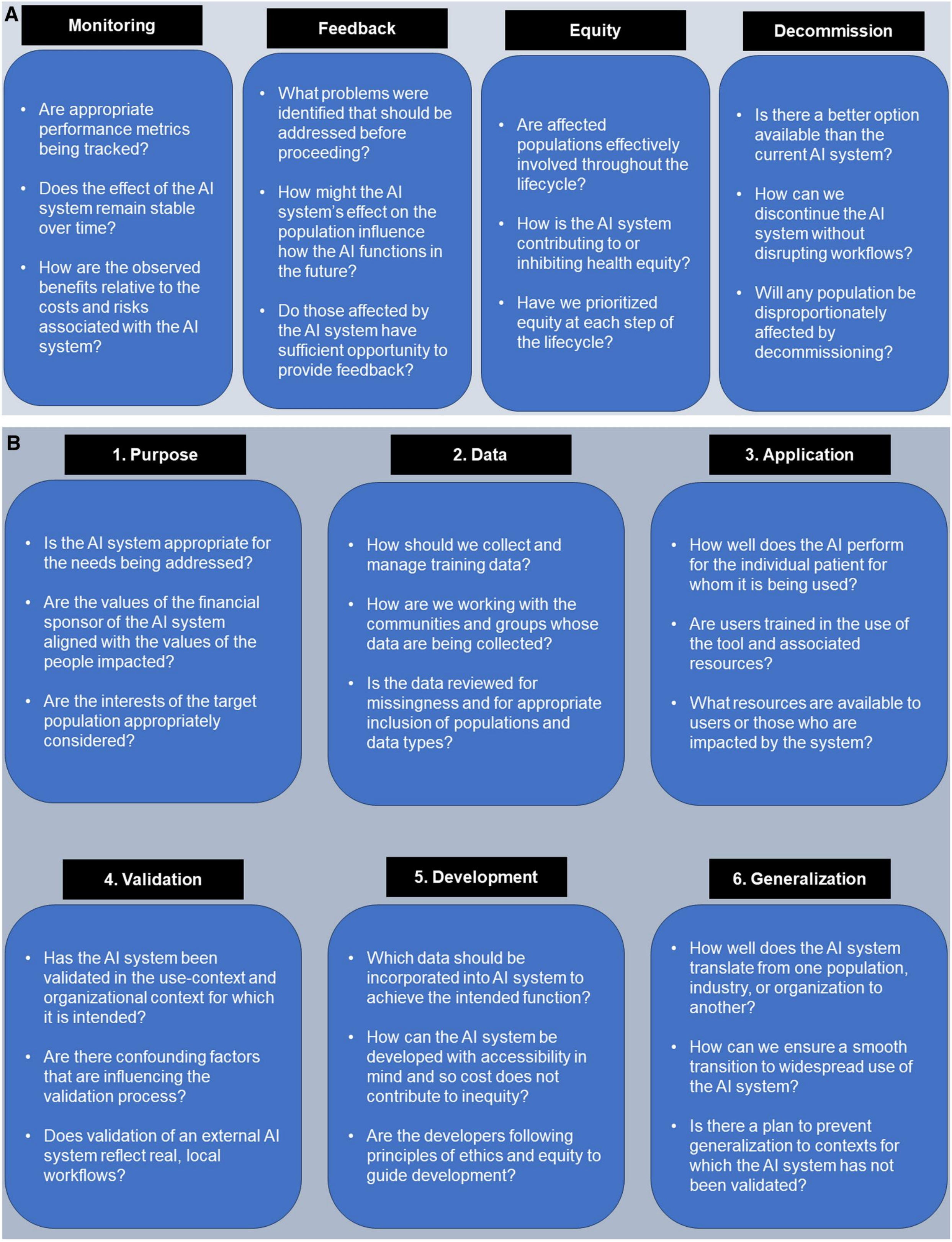


Figure 1. Healthcare AI ethics lifecycle with patients as stakeholders in the center as a use-case.

contained to a single section to avoid repetition in the explanations. Furthermore, the lifecycle we present here is compartmentalized for discussing benefits and harms to patients, indicated by the formation of the lifecycle around patients in the center, to facilitate the demonstration of a lifecycle-informed approach. As needed, the lifecycle can be tailored to other situations, stakeholders, or specific types of AI.

Lifecycle steps

1) The initial step represents the **purpose for which the AI system is built**. This purpose could be a research question, a clinical question, a commercial goal, a government-mandated regulatory goal, or other objectives identified by stakeholders. Any number of various stakeholders may be involved in determining the purpose of the AI system. The development of an AI tool may be motivated by multiple objectives that conflict with each other or even be morally ambiguous (eg, overall population health and optimizing outcomes for individual patients, or minimizing overall costs), especially when multiple stakeholders are involved. Multiple purposes for the AI can be a red flag as this may negatively impact overall performance.
2) The second step revolves around the **generation, collection, storage, provenance, privacy, and security of data** used to build, train, and operate an AI system. Value decisions and accessibility underlie which kinds of data are selected. The quality and reliability of the data as well as their completeness are critical.[25] However, group size is typically considered to be inversely correlated with privacy.[26] Another key consideration is whether the data are representative of the population for which the AI system is intended to apply. The concept of representation can take on several meanings depending on the context and perspective. It is common to consider representation in terms of age, sex, gender, race, ethnicity, socioeconomic status, disability status, or other demographics. However, even in representative data, biases may still be embedded.[27]
3) The third step includes the **technical development of the AI system as well as the physical hardware necessary for the system to function.** Decisions here include the digital and physical structure of the AI system, which features to include in or exclude from the system, and determining how the AI system will function. This step also includes decisions regarding whom to involve in the development process, which part and how much of the data to use for training, and how to program AI to achieve the intended purpose. Environmental costs are also important to consider here given the association between interactions with the environment and human health.[28]
4) The fourth step involves **internal and external validation to ensure the AI system functions as intended.** Internal validity refers to how well the design of a study fits a research question, the ability of the study to answer that question without bias, and how well an AI system matches its intended purpose. External validity refers to how well a study can generalize to other contexts (eg,



Monitoring

- Are appropriate performance metrics being tracked?
- Does the effect of the AI system remain stable over time?
- How are the observed benefits relative to the costs and risks associated with the AI system?

Feedback

- What problems were identified that should be addressed before proceeding?
- How might the AI system's effect on the population influence how the AI functions in the future?
- Do those affected by the AI system have sufficient opportunity to provide feedback?

Equity

- Are affected populations effectively involved throughout the lifecycle?
- How is the AI system contributing to or inhibiting health equity?
- Have we prioritized equity at each step of the lifecycle?

Decommission

- Is there a better option available than the current AI system?
- How can we discontinue the AI system without disrupting workflows?
- Will any population be disproportionately affected by decommissioning?

1. Purpose

- Is the AI system appropriate for the needs being addressed?
- Are the values of the financial sponsor of the AI system aligned with the values of the people impacted?
- Are the interests of the target population appropriately considered?

2. Data

- How should we collect and manage training data?
- How are we working with the communities and groups whose data are being collected?
- Is the data reviewed for missingness and for appropriate inclusion of populations and data types?

3. Application

- How well does the AI perform for the individual patient for whom it is being used?
- Are users trained in the use of the tool and associated resources?
- What resources are available to users or those who are impacted by the system?

4. Validation

- Has the AI system been validated in the use-context and organizational context for which it is intended?
- Are there confounding factors that are influencing the validation process?
- Does validation of an external AI system reflect real, local workflows?

5. Development

- Which data should be incorporated into AI system to achieve the intended function?
- How can the AI system be developed with accessibility in mind and so cost does not contribute to inequity?
- Are the developers following principles of ethics and equity to guide development?

6. Generalization

- How well does the AI system translate from one population, industry, or organization to another?
- How can we ensure a smooth transition to widespread use of the AI system?
- Is there a plan to prevent generalization to contexts for which the AI system has not been validated?

Figure 2. (A) Sample questions for ethical reflection to guide discussion of the possible benefits and harms within the cross-cutting issues. Decommission refers to decisions to discontinue or replace an AI system. (B) Sample questions for ethical reflection to guide discussion of the possible benefits and harms across the steps of the lifecycle.

different locations or populations) and how well an AI system functions for a broader sample. Internal validation alone is insufficient because models may align too closely with the training data so that other data cannot be incorporated into the model well, thus decreasing the performance of the AI.[29] Validation is essential for establishing the efficacy and safety of using an AI system, providing the basis for FDA approval of AI-based software as a medical device.[24] If the AI system fails validation, it may need to be retrained and/or the purpose(s) may need to be reconsidered.

5) The fifth step is the **application of AI in practice**. Application includes defining thresholds for implementation into healthcare workflows, ongoing decisions about when and how to apply AI, and the usability of an AI system. Questions of whether those who will be using AI and those for whom the AI is used are informed and prepared are crucial.[30] Organizations play an important role here,[31] needing to consider the potential consequences of using AI for purposes beyond their intended use and to implement suitable controls for managing the risks such uses could pose. For patient care, the users of AI systems will often be healthcare professionals, while smartphone AI healthcare apps allow for patient users.[32]
6) The final step, **generalization, represents the larger-scale effects of the AI on people and society**, including the sociocultural context in which the AI has been applied and the generalizability of AI to new contexts. The effects of generalizability are often measured through population health and epidemiological methods, but understanding the nuances of how AI systems affect populations can require investigating qualitative perspectives. This is especially the case for looking at how AI systems affect different populations. At the stage of generalization, there needs to be cognizance regarding translating population level statistical claims to the needs of individuals. This is a problem of quantifying the lives of people, not specific to AI.

Cross-cutting themes

A. **Monitoring involves overseeing the development process and the impact of the AI system.** Monitoring helps ensure that AI achieves its intended purpose for all who may be affected. Each step should incorporate intentional evaluation to determine if it is appropriate to move forward to the next step. Once systems are implemented into the workflow of healthcare practices, their efficacy must be monitored over time as their applicability may change over time as they learn and circumstances change.[33,34] A discrepancy between the expected and actual effect suggests unaccounted factors or other problems that require analysis and possibly modification of the AI. There should also be a process for reporting outcomes and errors that undergoes review.

B. **Feedback includes information returned at each step and learning of the AI system from iterative use.** Monitoring throughout the lifecycle can identify information that should influence prior steps, either by returning to a previous step or integrating that information on a future iteration. Feedback also occurs as the final step of generalization inherently feeds back into the first step. The function of AI can change over time in ways that require modification.

C. **Equity refers to the balance of an AI system's effect on populations affected.** Principles to promote equity should be intentionally considered at every step,[35] along with involving populations who are going to be affected, assembling diverse, representative community participants and a development team capable of understanding and honoring different perspectives. For their part, AI-capable organizations need to account for existing structural inequities,[31] as like all of healthcare, AI exists in a sociocultural context in which people are not equally situated from the outset. These preexisting disparities are root causes of many harms or unfairly distributed benefits associated with the use of the AI. While the historical factors that influence the impact of AI often are not within the control of the people and organizations involved in developing and using healthcare AI, these issues should be acknowledged and, when possible, addressed by means that are within their control. For example, historical data are often biased by poor data collection methods,[36] and the social problems that led to this circumstance are not altered by practices within the healthcare AI lifecycle. Nonetheless, the potential harms of such data should be identified, and actions taken for equitable data collection, to use debiasing methods,[37] or other means of limiting the harm of the historical data.

D. **Decommissioning (or termination) is the decision to discontinue or replace an AI system.** An AI system may be decommissioned at any point in the lifecycle if judged to be inadequate. It is important to note that the function of AI systems can be replaced with non-AI or non-technological solutions. Decisions to decommission a particular tool may come about if the AI has problems that cannot be sufficiently overcome, the availability of better alternatives, costs that cannot be covered, or if the tool is no longer needed. These decisions may come from developers or users of an AI system but could also originate from authority (ie, government), market dynamics, or other factors.

Case studies

Here we report on 3 cases, identifying where in the lifecycle the benefits and harms of AI for patients occur in each case to illustrate how this process can improve AI in healthcare. These cases were selected to be representative of the issues faced related to healthcare AI but given the allotted space not every important issue can be covered here.

Case 1—AI algorithm for distribution of healthcare resources

The first case is based on a fundamental, early example of potential harm from the use of AI algorithms in healthcare unearthed when Obermeyer and colleagues analyzed the use of a commercial risk-prediction algorithm to identify patients for inclusion in "high-risk care management" programs intended to improve the care for patients at high risk by shifting the distribution of resources to those who need them most.[38]

Benefits at the steps of purpose and generalization

The beneficial purpose was to support a fair allocation of resources across society so that patients with similar cases are treated in a similar manner was a beneficial purpose. This benefit could have accrued at the generalization step through the equitable need-based allocation of benefits, risks, and costs across society.[39] It is at a larger scale where the main benefit of resource distribution can be observed. Targeting additional resources to patients who need complex care can make healthcare more cost-effective and equitable which is particularly compelling in the United States where the healthcare system is experiencing significant rise in costs. Since resources are limited, it is the fair allocation of resources to patients in most need that is in itself a benefit.

Harms at the steps of data and development

Harms originated at the data step in this example because the available data reflected the influence of prior discriminatory practices and so did not represent the population to whom the algorithm was applied or who was actually in need of support. Then at the development step, the algorithm developers did not realize this data was afflicted by historical inequities. This ultimately led to the allocation of fewer resources for sick Black patients than similarly or less sick White patients. The data misled the algorithm because Black patients previously received less care than White patients, even though the former had higher disease burdens.

Without a lifecycle perspective, it may be possible to perceive harm only at the population level since this is where the effect becomes noticeable. However, placing a fix here, such as a debiasing of the initial results, does not fully address the problem because the bias cannot be entirely removed. The lifecycle allows recognizing harm at the root, providing an opportunity to reflect on how to address the problem. In this case, one may have determined that it would be suitable to identify and use more appropriate data that aligns better with the purpose of the AI and accurately reflects patients' severity of illness, thereby promoting equity.

Case 2—generative AI to automate clinical explanations

Liu and colleagues demonstrate the use of generative AI in test scenarios for automating the optimization of clinical decision support (CDS) with noninferiority compared to human CDS.[38] Existing CDS alert logic was input into ChatGPT, a large language model, so that it would generate suggestions for how the alert could be improved.

Benefits at the step of application

Automated optimization of CDS can be beneficial at the point of application if understandable, accurate alerts are produced when appropriate, allowing patients to receive more attention from their clinicians. As a whole, CDS generated by AI has been judged to be of better quality than those developed by humans,[40] which would also lead to more effective, personalized care for patients.

Harms at the step of application

Generative AI can produce false information, and thus have the capacity to place inaccurate information into CDS alert logic. Liu and colleagues describe a "hallucination" of ChatGPT where it suggested using a nonexistent biologic agent "etanerfigut" to treat a clinical problem. Generative AI created unique ideas or concepts from patterns of language data that do not simply replicate the content of the input data. Although "etanerfigut" is not a real medication, the name follows the pattern of other medications' names and is similar to "etanercept," which is an actual medication. It is the application of generative AI itself that presents the risk of harm.

In this instance, both the benefits and the harms of AI originate at the step of application. However, without considering the lifecycle framework, it can be more difficult to think about the benefits and harms in context. For generative AI, the preceding steps of the lifecycle do not have the same role as they do in other forms of AI because they do not determine the output in the same way. All forms of AI have the potential to be incorrect, but not all forms of AI can create output that is not aligned with the input data in a way that creates something new. Thinking about generative AI in the same way as thinking about other forms of AI can lead to applying fixes that may not be effective. Unlike the previous case, supplying more, or different, data does not necessarily resolve the issue of generative AI's hallucination. It is difficult to balance the accuracy of an AI system while also allowing it the freedom to generate unique content. Placing safeguards at the step of application to limit the potential for false outputs from AI in particular domains and support users to recognize when false outputs have been generated may allow generative AI in healthcare to be more trustworthy. For example, the AI in this case could be restricted to using actual medication names when it brings up medications. This does not minimize the role of having sufficient, quality data, but suggests there is still a need to guide how data are used. Ultimately, many applications of AI continue to require human oversight.

Case 3—a predictive AI model for patient deterioration

Singh et al document the evaluation of a proprietary AI model to predict patients' deterioration from COVID-19 and other causes.[41] The Epic Deterioration Index developed by Epic Systems Corporation (Epic), was trained on data from 3 healthcare systems to calculate the risk of patient deterioration, and promoted by Epic for adoption through financial incentives, and implemented in hundreds of hospitals.[42]

Benefits at the step of application

Initially developed prior to COVID-19 pandemic, the deterioration model was applied to COVID-19 at the onset of the pandemic, making use of the data from the Epic electronic health record (EHR). The accurate prediction of clinical deterioration could allow preemptive action to address the cause of deterioration, leading to better patient outcomes.

Harms at the steps of purpose, validation, generalization

This case has 3 separate primary origin points of harm: (1) purpose, (2) validation, and (3) generalization. Epic's promotion of the deterioration model through financial incentives for healthcare systems that use the Epic EHR partially shifted the purpose of the AI model from care of individual patients to Epic's financial gain.[42] This does not mean that models cannot be used by commercial organizations, but this can introduce the potential for bias from a conflict of interest, which may have affected decisions by healthcare organizations to implement the deterioration model. Even then, there was the opportunity to perform full validation of the AI model, a step bypassed by many healthcare organizations in part because the proprietary nature of the algorithm made validation more difficult.[41] The model failed to meet performance expectations as populations of different healthcare systems were not well represented in the populations on which the AI model was trained, leading many healthcare systems to decommission it from use.

This case represents the challenges associated with evaluating the risks and benefits of AI use in healthcare and provides a strong example of the value of the lifecycle-informed approach. When the lifecycle is not taken into account, important decision-points can blur to make multiple decisions appear to be one decision. For example, the steps of validation and generalization are both related to the harms in this case, and addressing those harms requires multiple decisions. Organizations need to test run AI locally as a proof of concept prior to full implementation and should be open to backing away from a problematic AI, whether that means returning to a previous step in the lifecycle or discontinuing use. Viewed as a lifecycle, many healthcare systems could have recognized more quickly had the appropriate steps been completed, enabling them to intercept the issues before they occurred. The lifecycle framework can also help decision-makers to recognize places where harms may not be as overt, such as with the potential for commercial bias at the initial step in this case. Intentionally thinking about commercial biases, or other cognitive biases, when determining the purpose of an AI system can strengthen the ethical foundation of the AI.

Discussion

A lifecycle-informed approach for examining the ethical issues of healthcare AI facilitates recognizing where benefits and harms may occur, using that knowledge to refine AI, and communicating about ethical decisions across disciplines. Following this approach acknowledges that single, dedicated fixes applied at later points of the lifecycle, such as bias detection, may be incapable of solving all the problems and that focusing on one aspect of AI cannot maximize the possible benefits.[43] If stakeholders of AI are able to recognize the benefits and harms in the context of where they originate in the lifecycle, it may be possible to formulate more effective solutions. Although a lifecycle-informed framework will not answer questions such as how accurate an AI system needs to be so that its benefits outweigh its risks, it can guide collaborative discussions on the accuracy of a system and its impact on various stakeholders toward ethically sound decisions. Ultimately, evaluation of the lifecycle in practice will take time and effort.

Through the case studies, we mapped a representative set of benefits and harms onto a healthcare AI lifecycle. Case 1 involved aspects of data bias and distributive justice; case 2 required weighing the potential benefits and harms at the same step of the lifecycle; and case 3 displayed how the interconnectivity of benefits and harms can lead to a complex web of ethical issues. The reality that it is not possible to choose benefits without the potential for harm cements the need to use a lifecycle-informed approach, which in practice may focus on additional factors such as the environment in which the AI is placed, the type of research involved, or who is using the lifecycle.

Next steps for stakeholders involved in the ethical integration of AI systems in healthcare include addressing how healthcare organizations may adopt the lifecycle perspective for decisions about AI in healthcare, mapping the ethical issues of other stakeholders in addition to patients, developing a more extensive list of questions that should be asked at each step, and applying the lifecycle framework to more complex issues such as the role of race and ethnicity as variables in healthcare AI. When ethical issues are identified, comprehensive and inclusive governance practices and structures, which are beyond the scope of this paper, will be needed to resolve those issues.

Conclusion

The lifecycle-informed approach provides a framework to foster discussion about the benefits and harms of healthcare AI. By facilitating proactive communication across disciplines, this framework opens the door to improved ethical decisions about AI in healthcare.

Acknowledgments

Members of the NIH Bridge2AI Ethics Core contributed to this paper.

Author contributions

Original conceptualization of the paper by Bradley A. Malin. Benjamin X. Collins led writing of the original draft and edits to the manuscript. Benjamin X. Collins and Ellen W. Clayton planned the methodology and performed data curation. Benjamin X. Collins prepared content visualization. All authors contributed to writing the original draft and review & editing of manuscript revisions.

Supplementary material

Supplementary material is available at *JAMIA Open* online.

Funding

This work was supported by NIH grants U54HG012510, U54HG012513, U54HG012517, OT2OD032720, OT2OD032701, OT2OD032644, and OT2OD032742 and NLM grant T15M007450.

Conflicts of interest

The authors have no competing interests to report.

Data availability

The data underlying this article are available in the article and in its online supplementary material.

References

1. Adler-Milstein J, Aggarwal N, Ahmed M, et al. Meeting the moment: addressing barriers and facilitating clinical adoption of artificial intelligence in medical diagnosis. NAM Perspectives. 2022. Accessed July 20, 2023. https://nam.edu/meeting-the-moment-addressing-barriers-and-facilitating-clinical-adoption-of-artificial-intelligence-in-medical-diagnosis
2. Moore J. AI in health care: the risks and benefits. Med Econ. 2023. Accessed May 5, 2023. https://www.medicaleconomics.com/view/ai-in-health-care-the-risks-and-benefits
3. Rajpurkar P, Chen E, Banerjee O, et al. AI in health and medicine. *Nat Med*. 2022;28:31-38.
4. Nazer LH, Zatarah R, Waldrip S, et al. Bias in artificial intelligence algorithms and recommendations for mitigation. *PLOS Digit Health*. 2023;2:e0000278.

5. Bhattacharyya M, Miller VM, Bhattacharyya D, et al. High rates of fabricated and inaccurate references in ChatGPT-generated medical content. *Cureus*. 2023;15:e39238.
6. Cascella M, Montomoli J, Bellini V, et al. Evaluating the feasibility of ChatGPT in healthcare: an analysis of multiple clinical and research scenarios. *J Med Syst*. 2023;47:33.
7. Nguyen T, Lai P, Phan N, et al. XRAND: differentially private defense against explanation-guided attacks. arXiv:2212.04454. 2022. Accessed July 28, 2023. https://arxiv.org/pdf/2212.04454
8. Evans BJ. Rules for robots, and why medical AI breaks them. *J Law Biosci*. 2023;10:lsad001.
9. Jagsi R, Griffith KA, Sabolch A, et al. Perspectives of patients with cancer on the ethics of rapid-learning health systems. *J Clin Oncol*. 2017;35:2315-2323.
10. Zusterzeel R, Goldstein BA, Evans BJ, et al. *Evaluating AI-Enabled Clinical Decision Clinical Decision and Diagnostic Support Tools Using Real-World Data*. Duke-Margolis Center for Health Policy; 2022.
11. Rosen MA, DiazGranados D, Dietz AS, et al. Teamwork in healthcare: key discoveries enabling safer, high-quality care. *Am Psychol*. 2018;73:433-450.
12. Abràmoff MD, Tarver ME, Loyo-Berrios N, et al.; Foundational Principles of Ophthalmic Imaging and Algorithmic Interpretation Working Group of the Collaborative Community for Ophthalmic Imaging Foundation, Washington, D.C. Considerations for addressing bias in artificial intelligence for health equity. *NPJ Digit Med*. 2023;6:170.
13. Ahmad MA, Eckert CM. Show your work: responsible model reporting in health care artificial intelligence. *Surg Clin North Am*. 2023;103:e1-e11.
14. Assadi A, Laussen PC, Goodwin AJ, et al. An integration engineering framework for machine learning in healthcare. *Front Digit Health*. 2022;4:932411.
15. Bedoya AD, Economou-Zavlanos NJ, Goldstein BA, et al. A framework for the oversight and local deployment of safe and high-quality prediction models. *J Am Med Inform Assoc*. 2022;29:1631-1636.
16. Chen IY, Pierson E, Rose S, et al. Ethical machine learning in healthcare. *Annu Rev Biomed Data Sci*. 2021;4:123-144.
17. Chen Y, Clayton EW, Novak LL, et al. Human-centered design to address biases in artificial intelligence. *J Med Internet Res*. 2023;25:e43251.
18. Dankwa-Mullan I, Scheufele EL, Matheny ME, et al. A proposed framework on integrating health equity and racial justice into the artificial intelligence development lifecycle. *J Health Care Poor Underserved*. 2021;32:300-317.
19. Economou-Zavlanos NJ, Bessias S, Cary MP, et al. Translating ethical and quality principles for the effective, safe and fair development, deployment and use of artificial intelligence technologies in healthcare. *J Am Med Inform Assoc*. 2023;31:ocad221.
20. McCradden MD, Odusi O, Joshi S, et al. What's fair is… fair? Presenting JustEFAB, an ethical framework for operationalizing medical ethics and social justice in the integration of clinical machine learning. In: *FAccT '23*. Association for Computing Machinery; 2023.
21. Ng MY, Kapur S, Blizinsky KD, et al. The AI life cycle: a holistic approach to creating ethical AI for health decisions. *Nat Med*. 2022;28:2247-2249.
22. Rojas JC, Fahrenbach J, Makhni S, et al. Framework for integrating equity into machine learning models: a case study. *Chest*. 2022;161:1621-1627.
23. Solanki P, Grundy J, Hussain H. Operationalising ethics in artificial intelligence for healthcare: a framework for AI developers. *AI Ethics*. 2022;3:223-240.
24. United States Food and Drug Administration. Proposed regulatory framework for modifications to artificial intelligence/machine learning (AI/ML)-based software as a medical device (SaMD). United States Food and Drug Administration. 2019. FDA-2019-N-1185-0001. Accessed July 31, 2023. https://www.fda.gov/media/122535/download
25. Ranjbar A, Ravn J. Data quality in healthcare for the purpose of artificial intelligence: a case study on ECG digitalization. *Stud Health Technol Inform*. 2023;305:471-474.
26. Sweeny L. k-anonymity: a model for protecting privacy. *Int J Uncertain Fuzziness Knowl-Based Syst*. 2002;10:557-570.
27. Ferryman K, Mackintosh M, Ghassemi M. Considering biased data as informative artifacts in AI-assisted health care. *N Engl J Med*. 2023;389:833-838.
28. Chevance G, Hekler EB, Efoui-Hess M, et al. Digital health at the age of the antropocene. *Lancet Digit Health*. 2020;2:e290-e291.
29. Ramspek CL, Jager KJ, Dekker FW, et al. External validation of prognostic models: what, why, how, when and where? *Clin Kidney J*. 2021;14:49-58.
30. Russell RG, Novak LL, Patel M, et al. Competencies for the use of artificial intelligence-based tools by health care professionals. *Acad Med*. 2023;98:348-356.
31. Novak LL, Russell RG, Garvey K, et al. Clinical use of artificial intelligence requires AI-capable organizations. *JAMIA Open*. 2023;6:ooad028.
32. Kane O, Ferryman K. Applying the ethical data practices framework to digital therapeutics. *Am J Bioeth*. 2023;23:53-56.
33. Wu DTY, Barrick L, Ozkaynak M, et al. Principles for designing and developing a workflow monitoring tool to enable and enhance clinical workflow automation. *Appl Clin Inform*. 2022;12:132-138.
34. Zheng K, Ratwani RM, Adler-Milstein J. Studying workflow and workarounds in electronic health record-supported work to improve health system performance. *Ann Intern Med*. 2020;172:S116-S122.
35. Hendricks-Sturrup R, Simmons M, Anders S, et al. Developing ethics and equity principles, terms, and engagement tools to advance health equity and researcher diversity in AI and machine learning: modified Delphi approach. *JMIR AI*. 2023;2:e52888.
36. Nagurney JT, Brown DFM, Sane S, et al. The accuracy and completeness of data collected by prospective and retrospective methods. *Acad Emerg Med*. 2005;12:884-895.
37. Nazer LH, Zatarah R, Waldrip S, et al. Bias in artificial intelligence algorithms and recommendations for mitigation. *PLOS Digit Health*. 2023;2:e0000278.
38. Obermeyer Z, Powers B, Vogeli C, et al. Dissecting racial bias in an algorithm used to manage the health of populations. *Science*. 2019;366:447-453.
39. Fisher OM, Brown KGM, Coker DJ, et al. Distributive justice during the coronavirus disease 2019 pandemic in Australia. *ANZ J Surg*. 2020;90:961-962.
40. Liu S, Wright AP, Patterson BL, et al. Using AI-generated suggestions from ChatGPT to optimize clinical decision support. *J Am Med Inform Assoc*. 2023;30:1237-1245.
41. Singh K, Valley TS, Tang S, et al. Evaluating a widely implemented proprietary deterioration index model among hospitalized patients with COVID-19. *Ann Am Thorac Soc*. 2021;18:1129-1137.
42. Drees J. Epic pays hospitals that use its EHR algorithms, report finds. Becker's Hospital Review. 2021. Accessed April 14, 2023. https://www.beckershospitalreview.com/ehrs/epic-pays-hospitals-that-use-its-ehr-algorithms-report-finds.html
43. Giovanola B, Tiribelli S. Beyond bias and discrimination: redefining the AI ethics principle of fairness in healthcare machine-learning algorithms. *AI Soc*. 2023;38:549-563.

JAMIA Open, 2024, 7, 1–8
https://doi.org/10.1093/jamiaopen/ooae108
Research and Applications