Journal of the American Medical Informatics Association, 2024, **31(3)**, 705–713
https://doi.org/10.1093/jamia/ocad221
Advance access publication 29 November 2023
Brief Communication

Brief Communication

Translating ethical and quality principles for the effective, safe and fair development, deployment and use of artificial intelligence technologies in healthcare

Nicoleta J. Economou-Zavlanos, PhD[1,*], Sophia Bessias, MPH, MSA[1], Michael P. Cary Jr, PhD, RN[1,2], Armando D. Bedoya, MD, MMCi[3,4], Benjamin A. Goldstein, PhD[1,5], John E. Jelovsek, MD[6], Cara L. O'Brien, MD[3,4], Nancy Walden, BS[1], Matthew Elmore, ThD[1], Amanda B. Parrish, PhD, RAC[7], Scott Elengold, JD[8], Kay S. Lytle, DNP, RN-BC[2,3], Suresh Balu, MS, MBA[9], Michael E. Lipkin, MD[10], Afreen Idris Shariff, MD, MBBS[4,11], Michael Gao, MS[9], David Leverenz, MD, MEd[4], Ricardo Henao, PhD[5,12], David Y. Ming, MD[4,13,14], David M. Gallagher, MD[4], Michael J. Pencina, PhD[1,5], Eric G. Poon, MD, MPH[3,4,5]

[1]Duke AI Health, Duke University School of Medicine, Durham, NC 27705, United States, [2]Duke University School of Nursing, Durham, NC 27710, United States, [3]Duke Health Technology Solutions, Duke University Health System, Durham, NC 27705, United States, [4]Department of Medicine, Duke University School of Medicine, Durham, NC 27710, United States, [5]Department of Biostatistics and Bioinformatics, Duke University School of Medicine, Durham, NC 27705, United States, [6]Department of Obstetrics and Gynecology, Duke University School of Medicine, Durham, NC 27710, United States, [7]Office of Regulatory Affairs and Quality, Duke University School of Medicine, Durham, NC 27705, United States, [8]Office of Counsel, Duke University, Durham, NC 27701, United States, [9]Duke Institute for Health Innovation, Duke University, Durham, NC 27701, United States, [10]Department of Urology, Duke University School of Medicine, Durham, NC 27710, United States, [11]Duke Endocrine-Oncology Program, Duke University Health System, Durham, NC 27710, United States, [12]Department of Bioengineering, King Abdullah University of Science and Technology, Thuwal 23955, Saudi Arabia, [13]Duke Department of Pediatrics, Duke University Health System, Durham, NC 27705, United States, [14]Department of Population Health Sciences, Duke University School of Medicine, Durham, NC 27701, United States

*Corresponding author: Nicoleta J. Economou-Zavlanos, PhD, Algorithm-Based Clinical Decision Support (ABCDS) Oversight Director, Duke AI Health, Duke University School of Medicine, Hock Plaza, 2424 Erwin Road, Durham, NC 27705 (nicoleta.economou@duke.edu)

Abstract

Objective: The complexity and rapid pace of development of algorithmic technologies pose challenges for their regulation and oversight in healthcare settings. We sought to improve our institution's approach to evaluation and governance of algorithmic technologies used in clinical care and operations by creating an Implementation Guide that standardizes evaluation criteria so that local oversight is performed in an objective fashion.

Materials and Methods: Building on a framework that applies key ethical and quality principles (clinical value and safety, fairness and equity, usability and adoption, transparency and accountability, and regulatory compliance), we created concrete guidelines for evaluating algorithmic technologies at our institution.

Results: An Implementation Guide articulates evaluation criteria used during review of algorithmic technologies and details what evidence supports the implementation of ethical and quality principles for trustworthy health AI. Application of the processes described in the Implementation Guide can lead to algorithms that are safer as well as more effective, fair, and equitable upon implementation, as illustrated through 4 examples of technologies at different phases of the algorithmic lifecycle that underwent evaluation at our academic medical center.

Discussion: By providing clear descriptions/definitions of evaluation criteria and embedding them within standardized processes, we streamlined oversight processes and educated communities using and developing algorithmic technologies within our institution.

Conclusions: We developed a scalable, adaptable framework for translating principles into evaluation criteria and specific requirements that support trustworthy implementation of algorithmic technologies in patient care and healthcare operations.

Key words: artificial intelligence; algorithms; technology assessment; health equity; quality assurance healthcare.

Introduction

Algorithmic technologies, including those that use artificial intelligence (AI), can enhance efficiency, reduce error, and provide evidence-based decision support in healthcare settings. However, they may lack transparency and can contribute to unintended harms such as misdiagnosis, inappropriate treatment decisions, loss of patient trust, and disparities in care.[1] Clear, robust requirements are urgently needed to ensure their implementation is effective, safe, fair, and equitable.

Federal and state agencies have provided guidance regarding health AI[2–4] and multiple frameworks have been advanced for developing, deploying, evaluating, and reporting on these

Received: June 22, 2023; **Revised:** October 6, 2023; **Editorial Decision:** October 29, 2023; **Accepted:** November 3, 2023

technologies.[5–10] However, health systems face practical and operational difficulties when deploying algorithmic technologies and must balance innovation with accountability and regulatory compliance. They also face an array of standards and regulations (current and proposed) and may not understand how these affect development and deployment of algorithmic technologies in healthcare settings.[11]

We previously reported on Duke University's Algorithm-Based Clinical Decision Support (ABCDS) Oversight, which provides governance, evaluation, and monitoring of algorithms used in clinical care and operations at Duke.[12] To facilitate application of key governance principles,[2,13] ABCDS Oversight developed an Implementation Guide that translates these principles into formal evaluation criteria and requirements for development teams. This principles-driven approach can be applied to AI solutions powered by predictive algorithms and can accommodate generative AI. Here, we describe this guide and its application to promoting clinical benefit, safety, and equitable impact of algorithmic technologies deployed at our institution.

Materials and methods

The ABCDS governance framework

Oversight Committee

The ABCDS Oversight Committee (OC) provides governance and oversight of algorithms proposed for use in the Duke University Health System. Its 14 members include clinical and operational leaders, biostatisticians and AI engineers, data scientists, and experts in law, regulation, and health equity[12] selected to represent key areas of knowledge and expertise across the Duke University School of Medicine, School of Nursing, and Health System. All algorithms intended for use in patient care settings at Duke must be registered with and evaluated by the OC before they can be deployed for clinical use.

Algorithm registration and evaluation

Development teams must notify the OC of intentions to deploy any algorithmic technology (whether developed locally or sourced from a vendor) that could affect patient care or clinical operations. Teams complete a registration form (Exhibit S1) detailing the technology's purpose, training, and proposed implementation. Registration information includes operational accountability with designated points of contact for questions, along with concise descriptions of appropriate use, modeling methods, data inputs, output(s), and plans for clinical and operational support. The development team must describe how the technology's design and proposed workflow integration may influence its potential to be considered Software as a Medical Device (SaMD) by the Food and Drug Administration (FDA).[4] The OC completes the registration process with a request for evaluation material.

Evaluation checkpoints

Evaluation of algorithmic technologies is based on 5 guiding principles applied across the development lifecycle: (1) transparency and accountability; (2) clinical value and safety; (3) fairness and equity; (4) usability and adoption; and (5) regulatory compliance (Figure 1). Just-in-time evaluation checkpoint gates (G_x) are placed between lifecycle phases and at regular postdeployment intervals. Before proceeding to each new phase of development or deployment, teams must provide evidence to demonstrate alignment with these principles. The review process (Figure 2) facilitates communication and collaboration between OC members and development teams and creates multiple opportunities for refining algorithms and use cases before clinical use.

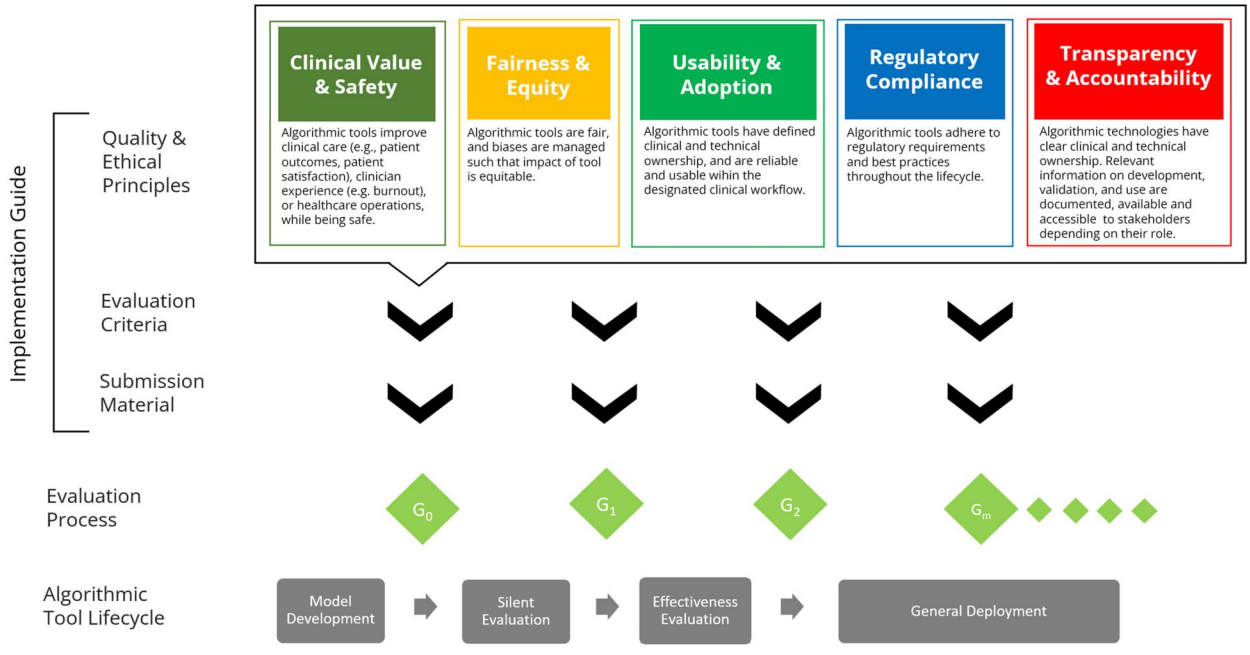

Figure 1. A framework for translating ethical and quality principles to evaluation criteria and submission requirements for the trustworthy development and deployment of health AI. An Implementation Guide introduces key ethical and quality principles into the lifecycle of algorithmic technologies through the checkpoint gates (G0, G1, G2, Gm) of evaluation processes represented as diamonds. The Implementation Guide is detailed in the Supplementary Material section of this manuscript.

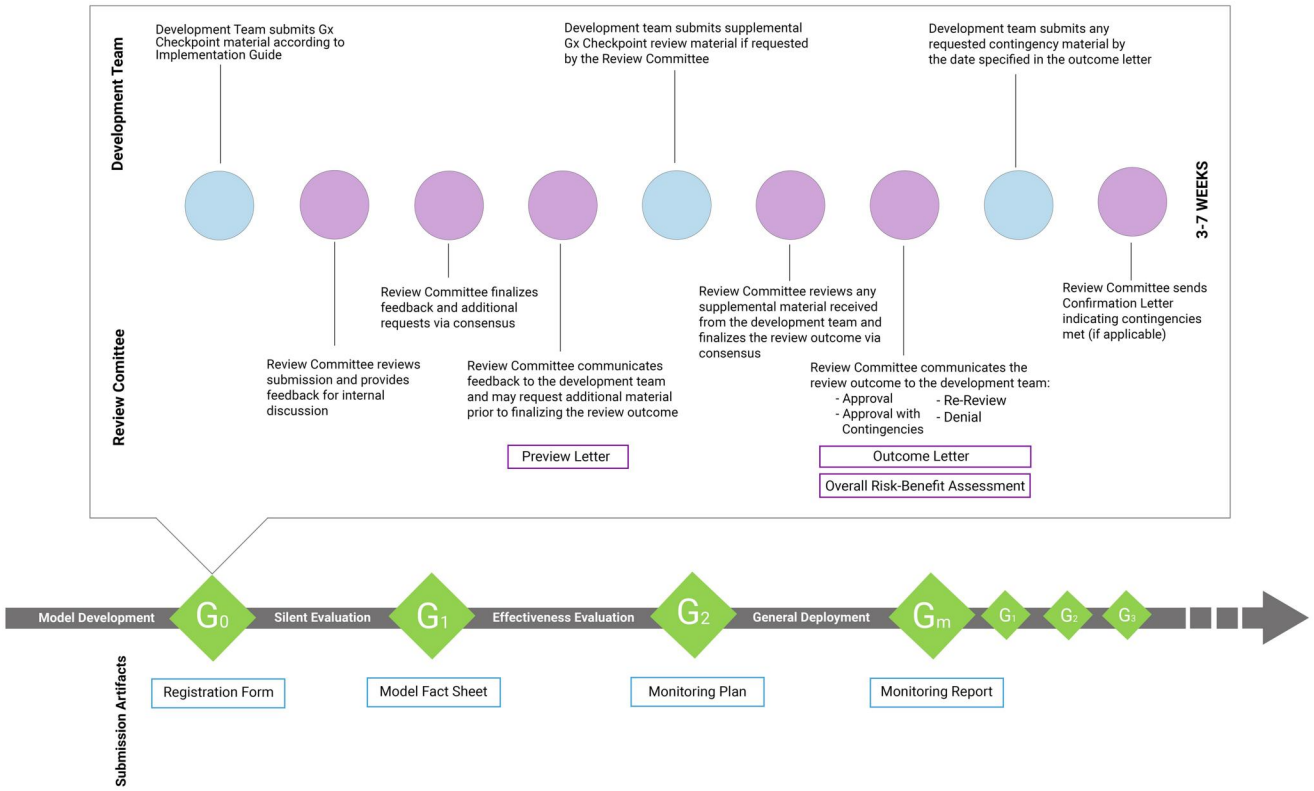

Figure 2. The (Gx) checkpoint review is a collaborative process of information sharing between the development team and the Review Committee (RC). Roles and responsibilities of the development team and the RC are shared in blue and purple bubbles, respectively. The Implementation Guide details standard information the RC requests from the development team to assess the trustworthy design and implementation of algorithmic technologies. Representative artifacts that are generated by the development team and the RC are highlighted in blue and purple boxes, respectively. Depending on the maturity of the algorithmic technology, different submission artifacts are presented over time to the RC throughout the lifecycle (color-filled boxes) as described within the Implementation Guide (Supplementary Material).

At each checkpoint, the development team initiates review by providing the submission material detailed in the Implementation Guide (Exhibit S2) plus updated registration information. A subset of OC members designated as the Review Committee (RC) then evaluates results from completed phases, assesses readiness for subsequent phases, and provides structured feedback in a short survey (Exhibit S3); subject matter experts are consulted as needed. In collaboration with institutional regulatory advisors, the RC also provides guidance on appropriate regulatory pathways and, where applicable, FDA engagement.

RC feedback is shared with the development team in a preview letter (Exhibit S3) detailing contingencies for proceeding to the next lifecycle phase and optional considerations for improvement. The preview letter serves as a vehicle to work with development teams and gather additional evidence the RC requires to assess risks and benefits of deployment and to determine the review outcome. Development teams have an opportunity to respond to RC requests with updated submission material. Finally, they receive an outcome letter (Exhibit S3) indicating the RC's decision: (1) approval; (2) approval with contingencies; (3) re-review; or (4) denial. Once the RC reaches an outcome decision, they document a risk-benefit analysis (Exhibit S3) to justify continued development. The time needed to complete a review varies, depending on the RC's initial findings and how quickly the development team can provide Supplementary Material; most reviews are completed within 3-7 weeks. Unless the RC requests a meeting, the process is fully asynchronous.

Creation and iterative refinement of the Implementation Guide

Initial evaluation criteria were based on reporting standards from the Transparent Reporting of a multivariable prediction model for Individual Prognosis or Diagnosis (TRIPOD) checklist[14] and informed by learnings from 3 to 4 algorithm reviews over a 6-month period. We established transparency requirements based on FDA guidance[15] and adopted model facts sheets[16] to provide minimal transparency information for end users. All criteria were endorsed by the OC.

Early feedback from development teams undergoing review indicated a desire for more concrete articulation of the types of evidence required for checkpoint evaluation. The OC responded with a review of program materials, culminating in the creation of a detailed Implementation Guide describing evaluation processes, requirements, and standards. Fairness and equity requirements, including bias identification and mitigation strategies, were incorporated based on independent work.[12,17]

The OC continues to iteratively refine the Implementation Guide, suggesting changes based on federal/local policies and/or ongoing learnings from quality improvement efforts.[18] Feedback and learnings are captured through reviews, education sessions, and daily communications with project teams

and logged in a centralized tracker to inform future updates. All oversight processes are updated based on decisions to ensure consistency across reviews, triage, and decision-making.

Results

The ABCDS Oversight portfolio currently comprises 54 registered algorithmic technologies at various lifecycle phases, including 17 AI algorithms in clinical use. Across multiple iterations of the Implementation Guide, the OC has conducted 36 reviews of 32 discrete algorithmic technologies.

The ABCDS Implementation Guide provides development teams with transparency into the evaluation criteria used for health AI technology review and aligns submission requirements with the ABCDS Oversight mission and principles. The 5 key principles noted above are mapped to evaluation criteria and other requirements specific to each lifecycle phase (Figure 1). This 3-tiered method encompasses: (1) quality and ethical principles, (2) evaluation criteria, and (3) submission materials. Before an algorithmic technology proceeds to a new lifecycle phase, evaluation criteria mapped to these governance principles must be met. The guide details the kinds of evidence that development teams must provide to determine whether candidate technologies meet criteria for all key principles (Table 1).

ABCDS evaluation has established a baseline for the quality of algorithms deployed at Duke Health, ensuring that they reflect core governance principles. The checkpoint process also informs determinations about whether a technology should continue to be developed or deployed and supports documentation of risks and benefits. Below, we illustrate the value added through the evaluation process with use cases taken from our experience with ABCDS review at Duke Health (Table 2).

G_0 checkpoint review

After submitting a registration form (Figure 2), the algorithmic technology is triaged and queued for review. At the G_0 checkpoint, the development team presents selected predictors and output variables (including those sensitive to discrimination, such as race, ethnicity, and sex, among others); proposed metrics for evaluating model performance, including calibration across the full range of predictions; justification for clinical decision thresholds; metrics for evaluating impact on outcomes or processes; and, where applicable, comparison with an existing solution or workflow the algorithm is intended to replace.[19]

G_0: The Immune-Related Adverse Events (IRAE) Algorithm

The IRAE algorithm predicts unplanned hospital admissions and emergency department visits in an at-risk specialty cohort of cancer patients receiving immune-checkpoint modulators. It is proposed to support a new workflow for proactive follow-up by the oncology care team based on targeted chart review.

The RC approved the IRAE algorithm to proceed from Model Development to Silent Deployment. As part of the G_0 evaluation, the RC requested additional calibration results, which revealed that although the algorithm effectively distinguished between patients who were likely versus unlikely to have adverse events, the predicted risk score was not proportional to the clinical probability of the adverse event. The RC suggested recalibrating the algorithm to better align probability output with true clinical risk. This sparked a constructive discussion with the development team, who implemented a technical solution that improved calibration prior to proceeding to silent evaluation with real-time patient data (Exhibit S4). The improvements supported the principles of "clinical value and safety" as well as "usability and adoption" by aligning the algorithm's output more closely with clinical judgment of risk and reducing need for additional interpretation steps.

G_1 checkpoint review

At the G_1 checkpoint following silent evaluation, the development team demonstrates that the data flows within the production environment meet expectations and the algorithm performs consistently when evaluated prospectively and retrospectively. The team also presents plans for a limited clinical implementation (effectiveness evaluation) to demonstrate real-world impact in a small, representative patient cohort before proceeding to general deployment. A finalized workflow diagram specifying roles and responsibilities for the clinical team and end user transparency information (eg, model facts sheets; training materials) are provided. The development team must share methods and findings from the silent evaluation reflecting clinical users' confidence in the algorithm's safety, fairness, and impact on processes and/or outcomes. At minimum, patient-level assessment in the form of chart review offers a pragmatic approach for determining whether the algorithm works appropriately, as well as how false positives/negatives may affect care. This informs understanding of real-world deployment risks, which can be shared with end users via model facts sheets, evaluation plans, and mitigation strategies. By this checkpoint, the regulatory pathway for each algorithmic technology should be established.

G_1: The Telehealth Appropriateness Algorithm

Duke Rheumatology developed a predictive algorithm based on EHR data and patient-reported disease activity to recommend visits appropriate for telehealth.[20,21] Providers are presented with the algorithm's output and work with the scheduling team to suggest a telehealth visit to rheumatology patients, who then choose either to accept or proceed with an in-person appointment.

In the initial G_1 submission, the development team demonstrated that the algorithm was successfully operationalized within its platform and that feedback mechanisms could be used to evaluate end user acceptance within the new workflow. A training plan for educating end users on appropriate use was developed, and a quantitative assessment of potential algorithmic biases provided.[22–24] Key details of the algorithm's data sources, development, and validation were described in a model facts sheet (Exhibit S4). Analyses from the silent evaluation phase suggested that the algorithm's performance would be acceptable across population subgroups in deployment. Measures of effectiveness and impact were established for ongoing monitoring.

Noting that the algorithm included sex as a significant predictor of appropriateness for telehealth care, the RC requested an exploration of potential challenges related to "fairness and equity" and asked the team to test for performance differences with and without sex as an input. Given the potential of the technology to contribute to differential offers of telehealth versus in-person care by sex, the development

Table 1. Governance principles, evaluation criteria, and submission requirements for G_0 checkpoint review.

Principle	Evaluation criteria	Submission material
Clinical value and safety	The ABCDS software, in comparison to current state, has potential to improve clinical outcomes and/or processes at Duke Health.	Results of a retrospective validation on Duke patient data, including: Description of key metrics used to evaluate model performance (concordance index, calibration metrics, sensitivity, PPV, etc.) Calibration of model performance across the full range of predictions (calibration curve) Justification for clinical decision thresholds, if applicable Description of key metrics used to evaluate impact on clinical outcomes and/or processes (net benefit) Comparison to the existing solution or workflow the algorithm is intended to replace, if applicable
	Plans for Silent Evaluation will inform the decision to proceed with pilot implementation in clinic.	Plans for prospective Silent Evaluation, including: Success criteria to proceed to clinical implementation Study design, including in/exclusion criteria, timeframe, and sample size considerations Plan for data analysis and data quality assessment Description of chosen evaluation metrics and shell tables Description of how the algorithm will deal with missing data
Usability and adoption	The proposed workflow links the tool's output(s) to clear decisions/actions and reflects understanding of the current workflow.	Workflow SWIM diagram describing how the new tool will change the current process or standard of care List of actions clinicians may take based on the tool's output Mock-up of the user interface
	Operational and clinical resources are in place and timelines established to proceed to the next ABCDS Lifecycle Phase.	Completed resource table for the upcoming ABCDS Lifecycle Phase
	The technical environment is ready for Silent Evaluation.	Technical feasibility summary with estimated timeline Name of DHTS contact and short description of build requirements (if applicable) Plans for operational qualification/testing of tool Data flow diagram
Fairness and equity	The project team has considered potential sources of bias and taken steps to prevent or mitigate associated harms.	Completed or updated bias analysis with bias management strategies Rationale for decisions regarding sensitive variables as inputs Summary of any new considerations re: fairness and equity
	A strategy has been articulated for ongoing fairness and equity subgroup analysis. Performance of the algorithm is acceptable across subgroups.	List of clinical impact and model performance metrics to examine across subgroups with rationale List of stratification variables with rationale Results of fairness and equity subgroup analysis with interpretation of findings
Regulatory compliance	Project team understands how design, implementation, and proposed use may influence the tool's potential to be considered SaMD.	Initial regulatory assessment re: SaMD with justification (may be provided within an updated registration form).
	IRB application plans (if any) are articulated for development. High-level plans are in place for commercialization (if applicable).	Brief comment and IRB submission reference number(s) for model development and/or implementation. Brief summary of commercialization plans and partners, including timing for integration of any external partners Data sharing agreement(s)
Transparency and accountability	ABCDS registration is complete and up-to-date.	Completed or updated ABCDS registration form
	Plans are in place for ongoing clinical and operational support.	Description of staff and financial resources needed to ensure the tool's sustainability Line of support by an individual with a budget
	Modeling methods, data inputs, and output(s) are clearly explained.	Description of training dataset with an explanation of time period, inclusion/exclusion criteria List of all data elements (predictors, outcomes, other) used to develop or evaluate the model Comments on data quality and limitations Description of modeling methods and variable importance

Abbreviations: ABCDS = algorithm-based clinical decision support; SaMD = software as a medical device.

Table 2. Impact of Implementation Guide on 4 algorithmic technologies.

G_x checkpoint	Tool name and use case	Impact of Implementation Guide during ABCDS review	Principles
G_0 checkpoint between model development and silent evaluation	The Immune-Related Adverse Events (IRAE) tool is a Duke-developed tool that predicts unplanned hospital admissions and emergency department visits in an at-risk specialty cohort of cancer patients. It is proposed to support a new, more proactive workflow for preventive follow-up by the oncology care team. The underlying algorithm is a lightGBM classifier.	The RC requested additional calibration results (Supplementary Material), which revealed that the model's predicted probabilities did not correspond well to true clinical risk. The RC suggested recalibrating the model, which led the development team to implement a technical solution resulting in improved calibration. Improved calibration renders the output of the model more interpretable and better aligned to clinical risk.	Clinical value and safety Usability and adoption
G_1 checkpoint between silent evaluation and effectiveness evaluation	The Telehealth Appropriateness Algorithm is a Duke-developed tool that recommends candidates for telehealth rheumatology appointments. Providers work with the scheduling team to offer telehealth appointments to patients identified by the model as appropriate candidates for virtual follow-up. Patients then have the option to accept or proceed with an in-person appointment. The underlying algorithm is a lasso logistic regression.	The RC requested visibility into the general impact of the tool and how that would be measured. The RC encouraged the development team to test the performance of the model with and without sex as an input. After considering potential equity implications and confirming that model performance was similar across the 2 candidate models, the development team opted to remove the sex variable. The RC shared input on implementation methods for the small-scale deployment of the model and requested monitoring of the impact of the tool on patients of different sociodemographic groups.	Fairness and equity Usability and adoption
G_2 checkpoint between effectiveness evaluation and general deployment	The Pediatric Complex Care Integration (PCCI) tool is a Duke-developed tool that provides population-level risk stratification among pediatric patients, allowing care managers to identify children who are at highest risk for hospitalization and most likely to benefit from complex care management services. Risk is reported as a percentage with thresholds for high, medium, and low risk. Thresholds are determined based on outreach capacity and probable clinical benefit. The underlying algorithm is an xgboost classifier.	The RC helped clarify the definition of clinical outcomes measured to demonstrate impact during monitoring phase The RC helped with the definition of the thresholds and requested additional calibration data to support the tool's generalizability to other contexts or resource levels. The RC requested transparency information, including a description of how the algorithm is presented within the interface and how the end user is trained.	Clinical value and safety Usability and adoption Transparency and accountability Regulatory compliance
G_m checkpoint in general deployment	The Readmission tool is a vendor-sourced tool used to stratify hospitalized patients according to risk of readmission over a 30-day period. This enables the case manager to discuss high-risk patients with the provider and prioritize follow-up calls within a week of discharge. The underlying algorithm is a logistic regression.	The RC requested additional transparency information, including who developed the model, the population in which the model is used, and the methods used for development and threshold selection. The RC requested more information about how the algorithm is presented within the interface, how the end user is trained, and what transparency information the end-user has at hand. The RC requested visibility into how the algorithm performs in different sociodemographic populations.	Fairness and equity Transparency and accountability Regulatory compliance

Abbreviation: RC = Review Committee.

team opted to remove the variable after confirming that the model continued to perform well without it (Exhibit S4). The development team recognized the capacity of the ABCDS independent review process to safeguard an element of fairness and equity that was initially overlooked. The OC also worked collaboratively with team members to refine implementation methods to deploy in a small-scale pilot study and meet project timelines, supporting the principles of "usability and adoption."

G_2 checkpoint review

At the G_2 checkpoint, results from the small-scale pilot, including an assessment of adoption by intended users, are evaluated. A maintenance and monitoring plan including metrics for performance, stability, clinical impact, fairness, and equity is shared, along with adoption measures for use cases, recommendations regarding monitoring frequency, and criteria for updating or retiring technologies. The team also provides a plan for ongoing fairness and equity management and regulatory compliance. Before real-world deployment, the algorithm's performance must be acceptable across population subgroups. Transparency materials are updated to inform end users about the technology and its appropriate use.

G_2: The Pediatric Complex Care Integration Algorithm (PCCI)

The PCCI algorithm is a population-level risk stratification technology that uses EHR data to allow care managers to identify children at highest risk for hospitalization and most likely to benefit most from complex care management services.[25] Risk is calculated as a percentage, with thresholds for high, medium, and low risk determined by outreach capacity and probable clinical benefit.

During the G_2 checkpoint evaluation, the RC requested calibration and decision curves showing algorithm performance across the full range of thresholds. This additional performance data further justified the team's definitions of low, medium, and high risk, supporting the principles of "clinical value and safety" as well as "usability and adoption" in case of shifting resources or application of the technology outside of Duke. Input from the RC also helped refine the monitoring plan with a strategy to measure "clinical value and safety" after deployment (Exhibit S4). Finally, the RC requested training material and screenshots of the user interface to support the principles of "regulatory compliance" and "transparency and accountability" through visibility into proposed workflow implementation.

G_m checkpoint review for continuous monitoring phase

Once fully deployed, the algorithm is continuously monitored for alignment with expectations regarding performance, stability, clinical impact, and adoption, and the monitoring plan is updated as needed. Plans for decommissioning and refinement are clearly delineated by the development team.

G_m: The Readmission Algorithm

The RC reviewed an algorithm, already in general deployment at the time ABCDS processes were established, that stratifies patients according to likelihood of readmission within 30 days of initial discharge.[26] The algorithm was developed using logistic regression; its inputs include demographics, diagnoses, labs, medications, and utilization data. A monitoring report was shared based on a plan the development team established for the readmission algorithm deployed within the EHR, and additional user interface and implementation details to assess evolving standards in "regulatory compliance."

Discussion

The ABCDS Oversight Implementation Guide aligns ethical and quality principles, evaluation criteria, and supportive evidence needed to ensure safe and effective clinical use of algorithms throughout their lifecycles. By providing clear descriptions and definitions of essential elements for evaluating and monitoring algorithmic technologies and embedding them within standardized processes, we have educated the different communities within our institution currently working with or developing such technologies. The guide also streamlines oversight processes and provides transparency and visibility to health system leadership on how algorithmic technologies, such as those using AI, affect patient care. In particular, actively working with key institutional offices can improve compliance with submission and review requirements and help disseminate awareness of ABCDS' mission and its context within different environments.

Objective criteria for quality, equity, and safety are essential for translating broad standards to fit the needs of individual health systems and hospitals without imposing unnecessarily cumbersome procedures. The processes encompassed by the Implementation Guide can be adapted to meet evolving standards and apply them to use cases within individual health systems. This makes the approach piloted here applicable beyond our local institution and extendable to emerging AI technologies. The 3-tiered, principles-based methodology is being applied to further iterate on the current framework to accommodate large language models as they are integrated into clinical and administrative workflows. A similar approach can be applied to demonstrate safety, effectiveness, and fairness for rules-based algorithms, local validation of externally validated or FDA-cleared algorithms, and even simple calculators used for standard care.

When reviews were originally performed, the committee required retroactive collection of evidence by development teams, who were also required to share reasonable plans for developing and managing an algorithm across its lifecycle. Although development teams do share the common goal to protect patient safety and deliver high-quality care, the considerable effort required to provide the necessary documentation for this review process was originally perceived as burdensome. To address this concern, the committee put emphasis on internal education and communication around ABCDS Oversight's mission, goals, requirements, and processes. Along with the continuing development of government standards and regulations, this effort to educate and communicate with stakeholders has now led to widespread institutional recognition that the additional steps needed to provide this evidence are valued.

A key priority during development of the Implementation Guide was to enable documentation of quality and effectiveness without stifling innovation by mandating an inflexible, burdensome review process. With the evolving landscape of AI technologies and methodologies, the guide is not rigidly

applied in isolation but is complemented by information about local algorithmic governance and processes.[12]

We recognize that scalability of this review process may be limited outside of settings with dedicated AI and data engineering resources; however, the creation of a network of national assurance labs combining resources from various organizations could provide technical assistance and lifecycle quality and safety assurance as a service in the future.[27] The creation of technological platforms to support governance, evaluation, and monitoring may also benefit lower-resource health systems. Nevertheless, lower-resource health systems can leverage this Implementation Guide as a starting point in assessing third party AI technologies.

Conclusion

The generalizable framework described here can be used to translate principles into continuously updated evaluation criteria and submission requirements that help all stakeholders engaged in developing and deploying algorithmic technologies to understand and implement their roles and responsibilities, while AI methodologies continuously evolve. Expert consensus on evaluating algorithmic technology will be key to standardizing objective review processes at larger scales, ultimately leading to impactful deployments of equitable and safe technologies in healthcare settings.

Acknowledgments

The authors thank Jonathan McCall, MS, and Rabail Baig, MA, for their editorial and creative support for this manuscript. The authors also thank ABCDS Oversight Committee member Sharon Ellison, PharmD, for supporting and serving ABCDS Oversight's mission.

Author contributions

All authors made substantial contributions to: conception and design of framework; interpretation of data; revising it critically for important intellectual content; and final approval of the version to be published. The first draft was written by N.J.E.-Z. and S.B.

Supplementary material

Supplementary material is available at *Journal of the American Medical Informatics Association* online.

Funding

This work was supported in part by the National Center for Advancing Translational Sciences of the National Institutes of Health (Award Number UL1TR002553). The content is solely the responsibility of the authors and does not necessarily represent the official views of the National Institutes of Health.

Conflicts of interest

 M. J.P. is a paid research collaborator at McGill University Health Centre in Montreal, Canada (stipend). He serves as a paid expert advisor for the National Medical Research Agency (Polska Agencja Badan Medycznych), Warsaw, Poland; consultant for Cleerly and Azra; and speaker for the Janssen Speaker's Bureau. He is also the recipient of the "Measuring Artificial Intelligence Development and Deployment Maturity in Health Care Organizations" grant from the Gordon and Betty Moore Foundation (grant number 12048). He is an unpaid adjunct professor of Biostatistics at Boston University, Boston MA. N.J.E.-Z. is the recipient of the "Measuring Artificial Intelligence Development and Deployment Maturity in Health Care Organizations" grant from the Gordon and Betty Moore Foundation (grant number 12048). D.L. has received an independent quality improvement and medical education grant from Pfizer, a medical education grant from the Rheumatology Research Foundation, and has served as a consultant for Sanofi. A.I.S. serves as a speaker and consultant for Merck and is also a consultant for Bristol Myers Squibb.

Data availability

This manuscript did not involve analysis or acquisition of datasets; however, all data relevant to this manuscript are shared and important relevant information about the use cases is also shared in the Supplementary Material.

References

1. 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
2. White House. *Blueprint for an AI Bill of Rights: Making Automated Systems Work for the American People*. Nimble Books; 2022.
3. Bonta R. *Attorney General Bonta Launches Inquiry into Racial and Ethnic Bias in Healthcare Algorithms*. State of California Department of Justice Office of the Attorney General; 2022. Accessed June 31, 2022. https://oag.ca.gov/news/press-releases/attorney-general-bonta-launches-inquiry-raci al-and-ethnic-bias-healthcare.
4. US Food and Drug Administration Website. Center for Devices and Radiological Health. *Artificial Intelligence and Machine Learning in Software as a Medical Device*. U.S. Food and Drug Administration; 2021. Accessed June 15, 2023. https://www.fda.gov/medical-devices/software-medical-device-samd/artificial-intelligence-and-machine-learning-software-medical-device.
5. Vasey B, Nagendran M, Campbell B, et al.; DECIDE-AI Expert Group. 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.
6. Collins GS, Dhiman P, Navarro CLA, et al. Protocol for development of a reporting guideline (TRIPOD-AI) and risk of bias tool (PROBAST-AI) for diagnostic and prognostic prediction model studies based on artificial intelligence. *BMJ Open*. 2021;11(7):e048008. https://doi.org/10.1136/bmjopen-2020-048008
7. World Health Organization. *Ethics and Governance of Artificial Intelligence for Health*. World Health Organization; 2021. Accessed June 15, 2023. https://www.who.int/publications/i/item/9789240029200.
8. Sanderson C, Lu Q, Douglas D, Xu X, Zhu L, Whittle J. Towards implementing responsible AI. *IEEE International Conference on Big Data*, Conference Publication. IEEE; December 2022; Osaka, Japan. Accessed May 23, 2023. https://ieeexplore.ieee.org/

document/10021121; https://doi.org/10.1109/BigData55660.2022.10021121
9. Microsoft Responsible AI Standard, V2. *General Requirements*. Microsoft; 2022. Accessed May 23, 2023. https://blogs.microsoft.com/wp-content/uploads/prod/sites/5/2022/06/Microsoft-Responsible-AI-Standard-v2-General-Requirements-3.pdf.
10. Google Responsible AI Practices. Google AI. Accessed May 23, 2023. https://ai.google/responsibility/responsible-ai-practices/.
11. National Institute of Standards and Technology. *NIST AI Risk Management Framework (AI RMF 1.0) Playbook*. NIST; 2023. Accessed May 24, 2023. https://airc.nist.gov/AI_RMF_Knowledge_Base/Playbook.
12. 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(9):1631-1636. https://doi.org/10.1093/jamia/ocac078
13. European Commission. *Ethics by Design and Ethics of Use Approaches for Artificial Intelligence (version 1.0)*. European Commission; 2021. Accessed May 23, 2023. https://ec.europa.eu/info/funding-tenders/opportunities/docs/2021-2027/horizon/guidance/ethics-by-design-and-ethics-of-use-approaches-for-artificial-intelligence_he_en.pdf.
14. Moons KG, Altman DG, Reitsma JB, et al. Transparent Reporting of a multivariable prediction model for Individual Prognosis or Diagnosis (TRIPOD): explanation and elaboration. *Ann Intern Med*. 2015;162(1):W1-W73.
15. U.S. Food and Drug Administration. *Clinical Decision Support Software. Final Guidance for Industry and FDA Staff*. U.S. Food and Drug Administration; 2022. Accessed June 12, 2023. https://www.fda.gov/media/109618/download.
16. Sendak MP, Gao M, Brajer N, Balu S. Presenting machine learning model information to clinical end users with model facts labels. *NPJ Digit Med*. 2020;3(1):41.
17. Cary MP Jr, Zink A, Wei S, et al. Mitigating racial and ethnic bias and advancing health equity in clinical algorithms: a scoping review. *Health Aff (Millwood)*. 2023;42(10):1359-1368.
18. Chandrasekaran A, Toussaint JS. *Creating a Culture of Continuous Improvement*. Harvard Business Review; 2019. Accessed May 23, 2023. https://hbr.org/2019/05/creating-a-culture-of-continuous-improvement.
19. Pencina MJ, Goldstein BA, D'Agostino RB. Prediction models—development, evaluation, and clinical application. *N Engl J Med*. 2020;382(17):1583-1586. https://doi.org/10.1056/NEJMp2000589
20. Solomon M, Henao R, Economou-Zavlanos N, et al. EASY model: development and pilot implementation of a predictive model to identify visits appropriate for telehealth in rheumatology. *Arthritis Care Res*. 2023; e-print. https://doi.org/10.1002/acr.25247
21. Smith ID, Coles TM, Howe C, et al. Telehealth Made EASY: understanding provider perceptions of telehealth appropriateness in outpatient rheumatology encounters. *ACR Open Rheumatol*. 2022;4(10):845-852.
22. Cerrato P, Halamka J, Pencina M. A proposal for developing a platform that evaluates algorithmic equity and accuracy. *BMJ Health Care Inform*. 2022;29(1):e100423.
23. Suresh H, Guttag J. A framework for understanding sources of harm throughout the machine learning life cycle. *Equity and Access in Algorithms, Mechanisms, and Optimization*. ACM; 2021:1-9.
24. Wang HE, Landers M, Adams R, et al. A bias evaluation checklist for predictive models and its pilot application for 30-day hospital readmission models. *J Am Med Inform Assoc*. 2022;29(8):1323-1333. Erratum in: *J Am Med Inform Assoc* 2022 Jun 17.
25. Ming DY, Zhao C, Tang X, et al. Predictive modeling to identify children with complex health needs at risk for hospitalization. *Hosp Pediatr*. 2023;13(5):357-369.
26. Gallagher D, Zhao C, Brucker A, et al. Implementation and continuous monitoring of an electronic health record embedded readmissions clinical decision support tool. *J Pers Med*. 2020;10(3):103.
27. Coalition for Health AI. Blueprint for Trustworthy AI Implementation Guidance and Assurance for Healthcare. Coalition for Health AI; 2023. Accessed September 25, 2023. https://www.coalitionforhealthai.org/papers/blueprint-for-trustworthy-ai_V1.0pdf.