VIEWPOINT

DIAGNOSTIC EXCELLENCE

Rethinking Algorithm Performance Metrics for Artificial Intelligence in Diagnostic Medicine

Matthew A. Reyna, PhD, MS
Department of Biomedical Informatics, Emory University, Atlanta, Georgia.

Elaine O. Nsoesie, PhD
Department of Global Health, School of Public Health, Boston University, Boston, Massachusetts.

Gari D. Clifford, DPhil, MA, MSc
Department of Biomedical Informatics, Emory University, Atlanta, Georgia; and Department of Biomedical Engineering, Georgia Institute of Technology and Emory University, Atlanta.

Corresponding Author: Gari Clifford, DPhil, MA, MSc, Department of Biomedical Informatics, Woodruff Memorial Research Building, 101 Woodruff Circle, 4th Floor East, Atlanta, GA 30322 (gari@gatech.edu).

The promise of artificial intelligence (AI) to improve and reduce inequities in access, quality, and appropriateness of high-quality diagnosis remains largely unfulfilled. Vast clinical data sets, extensive computational capacity, and highly developed and accessible machine-learning tools have resulted in numerous publications that describe high-performing algorithmic approaches for a variety of diagnostic tasks. However, such approaches remain largely unadopted in clinical practice.

This discrepancy between promise and practice—the AI chasm—has many causes. Some reasons are endemic to the larger field of AI, including a lack of generalizability and reproducibility for the published algorithms. Other reasons are more specific to clinical AI, such as a lack of gender, racial, and ethnic diversity in clinical data sets and insufficient evaluation of the algorithms in clinical settings. The disconnect between the metrics for algorithm performance and the realities of a clinician's workflow and decision-making process is a fundamental but often overlooked issue. The inclusion of clinical context in AI performance metrics for optimizing and evaluating clinical algorithms could make AI tools more clinically relevant and readily adopted (**Box**).

Clinicians and other health care decision makers have the responsibility to choose algorithms that are transparent, clinically useful, and effective across diverse patient populations.

Performance metrics are used to evaluate AI models and facilitate the interpretation and prioritization of models for clinical use. For example, the sensitivity and specificity of a diagnostic test describe the accuracy of the test for detecting a disease, and the area under the receiver operating characteristic curve reflects the ability of a model to differentiate between healthy patients and patients with disease or illness. However, despite the use of these and other traditional metrics in clinical settings, the metrics have no clinical context: all correct diagnoses are treated as equally positive, and all incorrect diagnoses are considered equally negative. Furthermore, there is no awareness that some algorithms are applied repeatedly to track the condition of a patient over time, triggering repeated false alarms that contribute to mistrust in the algorithm and alarm fatigue.

Few existing performance metrics incorporate clinical context, and different clinical problems require different metrics. Two such metrics, developed for a series of public competitions known as the PhysioNet Challenges, illustrate these issues.

For the 2019 PhysioNet Challenge, teams were asked to develop algorithms for early sepsis prediction.[1] The algorithms made hourly sepsis predictions to identify patients for treatment up to 12 hours before clinical recognition of sepsis onset. A time-dependent performance metric was designed to reward or penalize algorithms, depending on the clinical utility of their predictions and their likelihood of improving patient outcomes.[1] This metric provided high scores for early sepsis predictions to allow for earlier administration of fluids and antibiotics, with higher scores for earlier predictions. The metric provided low scores for late and missed sepsis predictions that would result in delayed treatments. It also provided low scores for false alarms that reduced confidence in the algorithm but less so than late and missed sepsis predictions. Critical care physicians provided input to develop this metric, which quantified their preferences about the value of early treatment, their ability to delay treatment, and the tolerance of staff to false alarms. The precise values of these quantities are open to debate, but they should be chosen to reflect the needs of the users of the algorithms.

For the 2020 and 2021 challenges,[2,3] teams were asked to develop algorithms for identifying 26 cardiac abnormalities from electrocardiograms (ECGs). The algorithms reported conditions that would subsequently be followed up by confirmatory tests, so a performance metric was designed to encourage correct diagnoses but provide different scores for different misdiagnoses.[2,3] This metric provided higher scores for misdiagnoses that resulted in the same follow-up testing and treatment as the correct diagnosis (eg, misclassifying atrial fibrillation as atrial flutter). However, the metric provided much lower scores for missing a more clinically significant arrhythmia that would require urgent attention (eg, misclassifying ventricular fibrillation as atrial fibrillation). Cardiologists were involved in helping to create this metric, and they defined rewards and penalties that reflected the risks and diagnostic similarities of each pair of cardiac abnormalities that were diagnosable from the ECG.

These competitions illustrate examples of generalizable patterns for designing performance metrics with clinical context. The goal is not to entirely replace traditional, one-size-fits-all performance metrics with another set of such metrics. Instead, the goal is to identify the salient features of a clinical problem and design

Box. Key Points for Diagnostic Excellence

Performance metrics for clinical diagnostic algorithms rarely incorporate features relevant to clinical utility and workflow

The development and application of clinically relevant metrics can refine and improve the adoption of artificial intelligence (AI) tools in clinical practice

Performance metrics should explicitly gauge bias and equity in diagnostic algorithms

Engagement between clinicians and algorithm developers is key to developing clinically relevant metrics

performance metrics that improve the clinical utility of the algorithms for clinicians who use them. This requires working with clinicians to define clinically relevant and practically feasible objectives for the algorithms to optimize. Multiple cost functions and constraints are possible and often necessary to describe qualitatively different objectives, such as diagnostic accuracy, timeliness, health care costs, and capacity; they are also necessary to help assess potential biases and differential performance across populations.[4] Reporting both novel and traditional metrics facilitates characterization of the trade-offs between metrics and helps the user understand why slightly lower accuracy values might be tolerated to substantially reduce bias or false alarm rates.

There are many potential and reasonable objections to the introduction of new performance metrics for clinical tasks. Poorly designed metrics can cause more harm than good. For example, health needs are correlated with health care costs, and costs are often easier to quantify. However, directly optimizing for health care costs instead of health needs can contribute to health disparities.[5] The indiscriminate optimization of surrogate metrics can be associated with bias and inequity, which are issues that must be explicitly considered with any metric. A proliferation of performance metrics can also impede the comparison of similar interventions. However, the common framework that traditional metrics provide is partially an illusion, as results for the same metric on different databases or clinical tasks are intrinsically incomparable.

Several suggestions may be helpful for consideration by clinicians and decision makers who are designing and using AI tools. First, clinicians should not assume that traditional metrics, such as the area under the receiver operating characteristic curve, translate to clinical effects because such performance metrics are usually not optimized or evaluated for specific clinical contexts.

Second, clinicians should be involved in guiding the design of metrics to ensure that the algorithms produce outputs that are clinically useful and patient-centered to minimize unintended harms.

Third, clinicians should prioritize the use of AI tools with well-documented and understandable explanations of performance metrics because doing so could enable informed decisions on whether and how best to use the algorithm.

Fourth, clinicians should expect the prospective evaluation of algorithms in clinical settings. Evaluation in varied settings demonstrates the potential utility of an algorithm for actual clinical outcomes.[6]

Fifth, adopters of AI tools should require that AI developers make available the full code for an algorithm, including the training data and code, so that the metrics used to develop the algorithms are explicit and modifiable.

Sixth, diagnostic performance metrics should take into account differential performance in subgroup populations, especially for conditions that may present differently based on race, ethnicity, or sex.

Clinicians and other health care decision makers have the responsibility to choose algorithms that are transparent, clinically useful, and effective across diverse patient populations. To facilitate an informed decision, algorithm development teams should also be diverse and work closely with clinicians to develop and implement AI performance metrics that incorporate clinical context. This process should also recognize and reflect the diversity of objectives and stakeholders in diagnostic medicine to improve the relevance and representation of AI tools in clinical practice.

ARTICLE INFORMATION

Published Online: July 8, 2022. doi:10.1001/jama.2022.10561

Conflict of Interest Disclosures: Dr Clifford reports conducting the PhysioNet Challenges, grants from the National Institute of Biomedical Imaging and Bioengineering (NIBIB) under National Institutes of Health (NIH) grant R01EB030362 (which supports the PhysioNet Challenges) during the conduct of the study. Dr Reyna reports conducting the PhysioNet Challenges, grants from the NIBIB under NIH grant R01EB030362 (which supports the PhysioNet Challenges) during the conduct of the study. No additional disclosures were reported.

Funding/Support: This article was supported, in part, by a grant from NIBIB under NIH (R01EB030362).

Role of the Funder/Sponsor: NIBIB and NIH had no role in the preparation, review, or approval of the manuscript or the decision to submit the manuscript for publication.

REFERENCES

1. Reyna MA, Josef CS, Jeter R, et al. Early prediction of sepsis from clinical data: the PhysioNet/Computing in Cardiology Challenge 2019. *Crit Care Med*. 2020;48(2):210-217. doi:10.1097/CCM.0000000000004145

2. Perez Alday EA, Gu A, J Shah A, et al. Classification of 12-lead ECGs: the PhysioNet/Computing in Cardiology Challenge 2020. *Physiol Meas*. 2021;41(12):124003. doi:10.1088/1361-6579/abc960

3. Reyna MA, Sadr N, Perez Alday EA, et al. Issues in the automated classification of multilead ECGs using heterogeneous labels and populations. Accepted manuscript. Published online July 8, 2022. *Physiol Meas*. doi:10.1088/1361-6579/ac79fd

4. Yang D, Fineberg HV, Cosby K. Diagnostic excellence. *JAMA*. 2021;326(19):1905-1906. doi:10.1001/jama.2021.19493

5. Mullainathan S, Obermeyer Z. On the inequity of predicting A while hoping for B. *AEA Pap Proc*. 2021;111:37-42. doi:10.1257/pandp.20211078

6. Nsoesie EO. Evaluating artificial intelligence applications in clinical settings. *JAMA Netw Open*. 2018;1(5):e182658-e182658. doi:10.1001/jamanetworkopen.2018.2658