Journal of the American Medical Informatics Association, 27(12), 2020, 2011–2015
doi: 10.1093/jamia/ocaa088
Advance Access Publication Date: 28 June 2020
Perspective

Perspective

MINIMAR (MINimum Information for Medical AI Reporting): Developing reporting standards for artificial intelligence in health care

Tina Hernandez-Boussard,[1,2,3]* Selen Bozkurt,[1] John P.A. Ioannidis,[1,4–6] and Nigam H. Shah[1,2]

[1]Department of Medicine, Stanford University, Stanford, California, USA, [2]Department of Biomedical Data Science, Stanford University, Stanford, California, USA, [3]Department of Surgery, Stanford University, Stanford, California, USA, [4]Department of Statistics, Stanford University, Stanford, California, USA, and [5]Meta-Research Innovation Center at Stanford, Stanford University, Stanford, California, USA

Corresponding Author: Tina Hernandez-Boussard, PhD, Medicine (Biomedical Informatics), Stanford School of Medicine, 1265 Welch Road, #245, Stanford University, Stanford, CA 94306, USA; boussard@stanford.edu

Received 16 April 2020; Revised 24 April 2020; Editorial Decision 24 April 2020; Accepted 29 April 2020

ABSTRACT

The rise of digital data and computing power have contributed to significant advancements in artificial intelligence (AI), leading to the use of classification and prediction models in health care to enhance clinical decision-making for diagnosis, treatment and prognosis. However, such advances are limited by the lack of reporting standards for the data used to develop those models, the model architecture, and the model evaluation and validation processes. Here, we present MINIMAR (MINimum Information for Medical AI Reporting), a proposal describing the minimum information necessary to understand intended predictions, target populations, and hidden biases, and the ability to generalize these emerging technologies. We call for a standard to accurately and responsibly report on AI in health care. This will facilitate the design and implementation of these models and promote the development and use of associated clinical decision support tools, as well as manage concerns regarding accuracy and bias.

Key words: reporting standards, electronic health records, artificial intelligence, clinical decision support

INTRODUCTION

The rise of digital data and advances in computing power have contributed to significant advancements in artificial intelligence (AI), including machine learning (ML), for clinical decision support for diagnosis, treatment, and prognosis.[1,2] The literature suggests that these methods may approach or exceed the performance of expert clinicians, particularly in the fields of signal processing, image classification, and spotting medication errors.[3–5] These advances bring hopes for better personalized and value-based care. The healthcare industry is becoming comfortable with AI-based solutions, which are rapidly emerging at the point of care.

However, the influx of AI models into the healthcare setting presents a fundamental shift in the use of data to guide clinical care and treatment decisions. Up until now, most models have been fed select input variables that were often handpicked by clinicians because they are known or suspected to have a valid clinical association with the outcome of interest. There are currently over 250,000 publications based on these kind of clinical scoring systems.[6] With

the increasing use of machine learning, the machine decides what input variables or features are important and related to the outcome of interest. Therefore, the data used for training and the definition of the task—be it classification or prediction—become more important than the specifics of the machine learning algorithm.[7] Detailed knowledge of the data used to train the model (ie, the training data) and the population those data represent—or often, does not represent—is essential to understanding the validity and generalizability of the "AI solution."

New reports suggest that biases hidden in the training data used for model development could have negative consequences in certain populations.[8,9] It is clear that the performance of any AI model broadly depends on its reliability and its ability to generalize to the setting and population in which it is applied, rather than its performance represented by the training and test data alone.[10] However, the characteristics of the data necessary to assess how these predictive models perform are not being adequately reported in the literature,[11] leaving uncertainty and doubt about the application in the broader healthcare setting. An empirical evaluation of 81 studies comparing AI models against clinicians showed major problems with lack of transparency, bias, and unjustified claims, likely because key details about the studies were often missing.[12]

Given the fast-evolving pace of AI solutions in health care, regulating them is complicated and global efforts are emerging to safely and efficiently standardize this regulatory task. The current regulatory environment is developing rapidly, with regulatory leaders and diverse stakeholders (eg, healthcare systems, clinicians, patients) developing a framework that both promotes innovation and ensures safety, privacy, and good intent.[13] There is a global consensus that AI solutions must be fair and nondiscriminatory and that AI solutions in health care should have a positive impact across all sectors of social and economic life.[2,14,15] However, through a lack of incentives, restrictions around data sharing and data privacy, and the acceptance of stealth science in industry (eg, science that is not backed by peer-reviewed evidence),[16] we have created a healthcare environment that allows AI solutions to be disseminated and deployed at point of care without understanding how the model was developed, from what data was the model learned, and using what data was the model deemed satisfactory for use.

Transparency is needed across 3 main categories: the population from which the data were acquired; model design and development, including training data; and model evaluation and validation. A lack of transparency regarding the training data used for model development directly affects the reproducibility, generalizability, and interpretability of a proposed model. Indeed, our recent study showed an alarming lack of transparency of ML models developed in research studies.[11] Therefore, we need transparency in the reporting of the design, development, evaluation, and validation of AI models in health care to achieve and retain confidence and trust for all the stakeholders.

Minimal standards for reporting scientific information are common and have improved the standards of biomedical as well as clinical research. From MIAME (Minimum Information About a Microarray Experiment) for gene expression microarrays to PRISMA (Preferred Reporting Items for Systematic Reviews and Meta-Analyses) meta-analyses,[17,18] reporting standards have emerged from communities to promote replication, validation and the use of secondary resources. These standards not only ensure transparent reporting of findings, but also guide authors in preparing their manuscripts, and allow journals to critically evaluate and appraise the findings, thus aiding the general interpretation of scientific information.

Many standards comprise a short checklist of minimal information required, such as the 25-item CONSORT (Consolidated Standards of Reporting Trials) statement for clinical trials, the 22-item STROBE (Strengthening the Reporting of Observational Studies in Epidemiology) checklist for observational studies, and the 33-item SPIRIT (Standard Protocol Items: Recommendations and Intervention Trials) checklist for interventional trials.[19–21] Importantly, both CONSORT and SPIRIT will be extending their checklists to include guidelines for trials that include an ML or AI component.[22] This will complement a new initiative from TRIPOD, TRIPOD-ML (Transparent Reporting of a Multivariable Prediction Model for Individual Prognosis or Diagnosis for Machine Learning). Feeding into these ongoing initiatives, we propose MINIMAR (MINimum Information for Medical AI Reporting), as a starting point for a broader community discussion. We believe that the adoption of such a standard will help the dissemination of such algorithms across healthcare systems and provide transparency to address potential biases and unintended consequences. MINIMAR will also promote external validation, encouraging the use of secondary resources.

GENERAL PRINCIPLES OF THE MINIMAR DESIGN

As a starting point, such a standard should satisfy the following requirements: (1) include information on the population providing the training data, in terms of data sources, cohort selection; (2) include training data demographics in a way that enables a comparison with the population the model is applied to; (3) provide detailed information about the model architecture and development so as to interpret the intent of the model and compare it to similar models and permit replication; and (4) transparently report model evaluation, optimization, and validation to clarify how local model optimization can be achieved and enable replication and resource sharing (Table 1).

The first requirement is related to the study population and setting, including patient demographics and cohort selection. It is essential to know the target population and how the training data were derived from this target population. This includes the need to understand the data that were used to develop (and train) the model, including the target patient population, the study setting, and data source, and how the final cohort was selected. These details provide the information on the data that a model is trained to anticipate potential biases and equity issues. As the second requirement, this should include the detailed documentation of patient characteristics and sensitive variables in the population, such as race and socioeconomic status. For example, a model that predicts general maternal mortality that is then applied to a black community must include a significant proportion of black patients in the training data, as well as risk factors applicable to them, such as sickle cell disease or high blood pressure, in order to adequately predict outcomes in the black community.[25] Data transparency is essential to promote fair and equitable models.

The third requirement would serve to provide a detailed explanation of the design and development of the AI or ML model in every publication. To evaluate any AI solution, it is essential to know the model task (ie, classification or prediction), the intended model output (eg, risk score for 30-day mortality), and the model beneficiary, if any. Currently, this is not widely done, which has led to important misinterpretations of model outcomes. For example, a recent study highlighted downstream bias in an ML model that was developed to

Table 1. Reporting standards for 4 essential components of artificial intelligence solutions in health care

Features	Description	Example[23]	Example[24]
1. Study population and setting			
Population	Population from which study sample was drawn	Patients undergoing elective surgery	All patients
Study setting	The setting in which the study was conducted (eg, academic medical left, community healthcare system, rural healthcare clinic)	U.S. academic, tertiary care hospital	2 U.S. academic medical lefts
Data source	The source from which data were collected	EHRs	EHRs
Cohort selection	Exclusion/inclusion criteria	Adult patients; Patients were excluded if they died during hospitalization.	All admissions for adult patients. Hospitalizations of 24 h or longer.
2. Patient demographic characteristics			
Age	Age of patients included in the study	Mean 58.34 y	Median ~56 y
Sex	Sex breakdown of study cohort	Female: 73.0% Male: 27.0%	Female 55.0%
Race	Race characteristics of patients included in the study	White: 69.0% Black: 3.1% Asian: 5.9%	Not provided
Ethnicity	Ethnicity breakdown of patients included in the study	Hispanic: 13.2%	Not provided
Socioeconomic status	A measure or proxy measure of the socioeconomic status of patients included in the study	Private: 31.9% Medicare: 47.8% Medicaid: 11.7%	Not provided
3. Model architecture			
Model output	The computed result of the model	Postoperative pain scores	In-hospital deaths, 30-day unplanned readmission, length of stay, discharge status
Target user	The indented user of the model output (eg, clinician, hospital management team, insurance company)	Risks scores produced by the model will be used by the hospital team for pain management	Predictions produced by the model will be used by hospitals for care management
Data splitting	How data were split for training, testing, and validation	10-fold cross-validation	80%/10%10% (train/validation/test)
Gold standard	Labeled data used to train and test the model	100 manually annotated clinical notes and pain scores recorded in EHR	Death, readmission and ICD codes in EHRs
Model task	Classification or prediction	Prediction	Prediction
Model architecture	Algorithm type (eg, machine learning, deep learning, etc.)	ElasticNet regularized regression	Recurrent neural networks, attention-based time-aware neural network model, and neural network
Features	List of variables used in the model and how they were used in the model in terms of categories or transformation	65 predictive features including age, race, ethnicity, sex, insurance type (as public and private) and preoperative pain (log transformation was applied)	Provided in detail for all models
Missingness	How missingness was addressed: reported, imputed, or corrected	Missing data were imputed using median of the variable distribution	Not provided

(continued)

Table 1. continued

Features	Description	Example[23]	Example[24]
4. Model evaluation			
Optimization	Model or parameter tuning applied	Generated vectors with a dimension of 300 and a window size of 5	Documented and provided for all models in detail
Internal model validation	Study internal validation	Internal 10-fold cross-validation	Hold-out validation set
External model validation	External validation using data from another setting	Not performed	Not performed
Transparency	How code and data are shared with the community.	Code and sample data available via GitHub	Data is not available; code is available via GitHub

EHR: electronic health record; ICD: International Classification of Diseases.

predict costs of care yet was implemented in the healthcare setting to predict need of care.[7] This misinterpretation resulted in allocating more intensive care resources to patients who had higher reimbursement rates, rather than to patients who had higher clinical need for those resources. Other necessary model details, such as modeling technique, feature selection, and the handling of missing values, should be transparent to appropriately apply an AI model in health care.

The fourth requirement is related to information on model evaluation, including optimization and validation. Model evaluation strategies should be defined in detail, in terms of data used for both internal and external validation as well as the adopted approach adopted for evaluation (eg, 5-fold cross-validation or 80/20 split). The choice of validation metrics, such as sensitivity, specificity, positive predictive value, or area under the receiver-operating characteristic curve, also needs to be defined. In addition, the overall model performance metrics and the hyperparameters chosen for the final best model optimizations should be reported. Finally, as part of model evaluation, transparency is necessary for broad AI application in health care in order to achieve and retain confidence and trust from all the stakeholders. Indeed, recent studies show an alarming difficulty in reproducing models developed in research studies and suggest that even if the training data cannot be shared due to privacy issues, the source code of the model should be shared publicly.[26] Therefore, in order to demonstrate the provenance and authenticity of the data and knowledge used to make decisions by AI models, promoting access to training data and source code is crucial to ensure that ML in biomedicine can be broadly applied and generalized. This is essential not only for choosing the best model for the given setting, but also for the unbiased comparison of different models or different settings.

DISCUSSION

Our goal is to set forth a standard for minimum information necessary to understand intended predictions, target populations, and hidden biases of an AI and ML clinical decision tool for both research scientists and medical practitioners. To that end, we hope that this description will stimulate discussion of the proposed MINIMAR standards and encourage the medical informatics community, as well as the general research community, to provide us with their views on how this standard can be improved.

Clearly, the consequences of making wrong or inaccurate classifications or predictions in health care can be fatal. To address this, we need clear reporting of the training data, the model architecture, and evaluation and validation procedures. For that, we need reporting standards. Here, we start this conversation by proposing MINIMAR, the minimal information for medical AI reporting. We believe it would be valuable if groups producing these studies would strive for a level of transparency in their methods that supports the reproducibility of results, in particular on different underlying population representations. This information can help prioritize research agendas and highlight populations underrepresented in this wave of medical informatics. We call for a standard to accurately and responsibly report on AI in health care. This will facilitate the design and implementation of these models and promote the development and use of associated clinical decision support tools, as well as managing concerns regarding accuracy and bias. In this era of data-driven medicine, establishing minimum standards for developing and reporting methodologies, sharing algorithms and tools, and establishing other resources is essential to ensure transparency and equity are at the forefront of AI-augmented health care. This is a necessary step in a larger agenda that will help assess the ethics, regulation, and effectiveness of AI models in transforming health care.

AUTHOR CONTRIBUTIONS

TH-B attests that all listed authors meet authorship criteria and that no others meeting the criteria have been omitted. TH-B affirms that the article is an honest, accurate, and transparent account of the study being reported; that no important aspects of the study have been omitted; and that any discrepancies from the study as planned have been explained. TH-B was involved in study concept and design; drafting of the article; administrative, technical, or material support; and study supervision. TH-B, SB, JPAI, and NHS were involved in critical revision of the manuscript for important intellectual content.

CONFLICT OF INTEREST STATEMENT

None declared.

REFERENCES

1. Rothman B, Leonard JC, Vigoda MM. Future of electronic health records: implications for decision support. *Mt Sinai J Med* 2012; 79 (6): 757–68.
2. He J, Baxter SL, Xu J, Xu J, Zhou X, Zhang K. The practical implementation of artificial intelligence technologies in medicine. *Nat Med* 2019; 25 (1): 30–6.
3. Hannun AY, Rajpurkar P, Haghpanahi M, *et al*. Cardiologist-level arrhythmia detection and classification in ambulatory electrocardiograms using a deep neural network. *Nat Med* 2019; 25 (1): 65–9.

4. Schiff GD, Bates DW. Can electronic clinical documentation help prevent diagnostic errors? *N Engl J Med* 2010; 362 (12): 1066–9.
5. Schiff GD, Volk LA, Volodarskaya M, *et al*. Screening for medication errors using an outlier detection system. *J Am Med Inform Assoc* 2017; 24 (2): 281–7.
6. Challener DW, Prokop LJ, Abu-Saleh O. The proliferation of reports on clinical scoring systems: issues about uptake and clinical utility. *JAMA* 2019; 321 (24): 2405–6.
7. 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): 447–53.
8. Ferryman K, Pitcan M. Fairness in precision medicine. *Data & Society*. 2018. https://datasociety.net/library/fairness-in-precision-medicine/ Accessed November 19, 2019.
9. Gianfrancesco MA, Tamang S, Yazdany J, Schmajuk G. Potential biases in machine learning algorithms using electronic health record data. *JAMA Intern Med* 2018; 178 (11): 1544–7.
10. Riley RD, Ensor J, Snell KI, *et al*. External validation of clinical prediction models using big datasets from e-health records or IPD meta-analysis: opportunities and challenges. *BMJ* 2016; 353: i3140.
11. Bozkurt S, Cahan E, Seneviratne MG, *et al*. Reporting of demographic data, representativeness and transparency in machine learning models using electronic health records. *JAMA Netw Open* 2020; 3 (1): e1919396.
12. Nagendran M, Chen Y, Lovejoy CA, *et al*. Artificial intelligence versus clinicians: systematic review of design, reporting standards, and claims of deep learning studies. *BMJ* 2020; 368: m689.
13. Food and Drug Administration. Proposed regulatory framework for modifications to artificial intelligence/machine learning (AI/ML)-based software as a medical device (SaMD). https://www.fda.gov/medical-devices/software-medical-device-samd/artificial-intelligence-and-machine-learning-software-medical-device Accessed November 19, 2019.
14. Vought RT. Memorandum for the Heads of Executive Departments and Agencies. Guidance for Regulation of Artificial Intelligence Applications. https://www.whitehouse.gov/wp-content/uploads/2020/01/Draft-OMB-Memo-on-Regulation-of-AI-1-7-19.pdf Accessed November 19, 2019.
15. Roberts H, Cowls J, Morley J, Taddeo M, Wang V, Floridi L. The Chinese approach to artificial intelligence: an analysis of policy and regulation. 2019. https://papers.ssrn.com/sol3/papers.cfm?abstract_id=3469784 Accessed November 19, 2019.
16. Cristea IA, Cahan EM, Ioannidis J. Stealth research: lack of peer-reviewed evidence from healthcare unicorns. *Eur J Clin Invest* 2019; 49 (4): e13072.
17. Brazma A, Hingamp P, Quackenbush J, *et al*. Minimum information about a microarray experiment (MIAME)-toward standards for microarray data. *Nat Genet* 2001; 29 (4): 365–71.
18. Moher D, Liberati A, Tetzlaff J, Altman DG; PRISMA Group. Preferred reporting items for systematic reviews and meta-analyses: the PRISMA statement. *Int J Surg* 2010; 8 (5): 336–41.
19. Chan AW, Tetzlaff JM, Altman DG, *et al*. SPIRIT 2013 Statement: defining standard protocol items for clinical trials. *Rev Panam Salud Publica* 2015; 38 (6): 506–14.
20. von Elm E, Altman DG, Egger M, *et al*. The Strengthening the Reporting of Observational Studies in Epidemiology (STROBE) statement: guidelines for reporting observational studies. *Lancet* 2007; 370 (9596): 1453–7.
21. Begg C, Cho M, Eastwood S, *et al*. Improving the quality of reporting of randomized controlled trials. The CONSORT statement. *JAMA* 1996; 276 (8): 637–9.
22. Liu X, Faes L, Calvert MJ, Denniston AK, CONSORT/SPIRIT-AI Extension Group. Extension of the CONSORT and SPIRIT statements. *Lancet* 2019; 394 (10205): 1225.
23. Parthipan A, Banerjee I, Humphreys K, *et al*. Predicting inadequate postoperative pain management in depressed patients: a machine learning approach. *PLoS One* 2019; 14 (2): e0210575.
24. Rajkomar A, Oren E, Chen K, *et al*. Scalable and accurate deep learning with electronic health records. *NPJ Digit Med* 2018; 1 (1): 18.
25. Villarosa L. Why America's black mothers and babies are in a life-or-death crisis. *The New York Times Magazine*. 2018: 11. https://www.nytimes.com/2018/04/11/magazine/black-mothers-babies-death-maternal-mortality.html Accessed November 19, 2019.
26. National Academies of Sciences, Engineering, and Medicine. *Reproducibility and Replicability in Science*. Washington, DC: National Academies Press; 2019.