

AHA SPECIAL REPORT

The Transparent Reporting and Evaluation of Artificial Intelligence Models (TRUE-AIM) Report Card: A Quality Assurance Initiative of the American Heart Association Journals to Promote High-Quality, Reproducible Artificial Intelligence Research

Márton Kolossváry¹ MD, PhD; Eric E. Smith² MD, MPH; Mathew J. Reeves³ PhD; Anita Layton⁴ PhD; Nicholas L. Mills⁵ MD, PhD; Manish Motwani, MBChB, PhD; Shinya Goto⁶ MD, PhD; Neil Atkins⁷ MA

Artificial intelligence (AI) aims to emulate human perception and reasoning to create intelligent agents that can achieve and potentially surpass human capacities.¹ The flexibility of AI techniques allows for rapid implementation of the newest advancements, potentially transforming medical practice.² However, AI models are fine-tuned to their development data set, and therefore contain knowledge representations of training population characteristics. Implementation of AI models to new populations with different characteristic distributions, differences in machines used to generate the inputs to the models, and changes in treatment protocols and disease classifications, all may contribute to model drift and poorer performance. Resulting medical errors can lead to patient harm, and implicit biases coded in the training of the models can perpetuate existing disparities in medical care.

The editorial board members of the American Heart Association (AHA) journals are committed to transparency, open-science principles, and quality assurance in the publication of AI-based research articles, with the goal of achieving high-quality, reproducible research that results in generalizable, inclusive models while minimizing the potential for harm.^{3,4} Because many AI health tools, including those with regulatory approval, are not clinically validated,⁵ medical journal editors have an

obligation to scrutinize every AI-based research article for quality, reproducibility, and adherence to a minimum set of agreed standards, regardless of whether the data are intended as the basis for regulatory approval or clinical adaptation.^{6–8}

Over the past 5 years, a wide array of AI-related guidance documents and checklists have been published or are under development.^{9,10} Broadly speaking, these documents cover the development, evaluation, and implementation of AI-related models and tools with the goal of improving transparency, reproducibility, and fairness. Whereas the central mission of all these documents is to improve the quality and reproducibility of published research, some reports present checklists specifically designed to improve the structure and content of submitted manuscripts that are organized around a specific type of study design (eg, clinical trials [CONSORT-AI (Consolidated Standards of Reporting Trials–Artificial Intelligence)],¹¹ prediction modeling [TRIPOD+AI (Transparent Reporting of a Multivariable Prediction Model for Individual Prognosis or Diagnosis + Artificial Intelligence)],¹² or diagnostic studies [STARD-AI (Standards for Reporting Diagnostic Accuracy–Artificial Intelligence)]¹³). Other guidance documents take on specific technical issues, such as ethics and fairness (eg, FUTURE-AI [Fairness, Universality, Traceability, Usability, Robustness,

Key Words: artificial intelligence ■ deep learning ■ machine learning ■ research

Correspondence to: Márton Kolossváry, MD, PhD, Gottsegen National Cardiovascular Center, 29 Haller u. Budapest 1096, Hungary. Email marton.kolossvary@gokvi.hu
For Sources of Funding and Disclosures, see page e322.

© 2025 American Heart Association, Inc.

Circulation is available at www.ahajournals.org/journal/circ

Nonstandard Abbreviations and Acronyms

AHA	American Heart Association
AI	artificial intelligence
TRUE-AIM	Transparent Reporting and Evaluation of Artificial Intelligence Models

and Explainability–Artificial Intelligence]¹⁴), data and code sharing (eg, MI-CLAIM [Minimum Information About Clinical Artificial Intelligence Modeling],¹⁵ MINIMAR [Minimum Information for Medical AI Reporting]¹⁶), model evaluation metrics (eg, CLAIM [Checklist for Artificial Intelligence in Medical Imaging],¹⁷ MI-CLAIM,¹⁵ TRIPOD+AI¹²), quality scoring and risk of bias assessments (eg, PROBAST+AI [Prediction Model Risk of Bias Assessment Tool + Artificial Intelligence],¹⁸ METRICS [Methodological Radiomics Score]¹⁹), and presentation details (TRIPOD+AI¹²). Given the diverse range of goals and objectives of these reports, typically there is no single guidance document or checklist that is clearly superior; often more than one document is relevant depending on the specific purpose and content of the manuscript.

In response to the rapid increase in AI-related submissions to the AHA journals, and given the inherent complexity of AI development and its broad range of clinical applications, we developed the TRUE-AIM (Transparent Reporting and Evaluation of Artificial Intelligence Models) report card through a series of consensus-building discussions. The primary aim of the TRUE-AIM report card is to provide a framework that ensures standardized reporting of AI-related methodologic aspects of a study that may affect the validity and generalizability of the results. Providing standard responses to most of the questions of the report card promotes standardization, reproducibility, and comparability within AI research. Furthermore, similar to the Model Facts label initiative, which aims to provide clinicians with standardized, easy-to-understand information regarding the clinical adaptation of specific AI models,²⁰ we designed the TRUE-AIM report card to provide a comprehensive overview of relevant information concerning methodologic aspects of the conducted AI research. Rather than offering a simple checklist with page numbers where relevant information can be found, authors are required to provide responses using predefined single-choice answers (ie, yes/partially/no/not applicable) or short answers, and a comments section where further elaboration can be provided. Our aim with the TRUE-AIM report card is to provide reviewers and readers an understandable, standardized summary of applied methodologies to aid in the evaluation and interpretation of AI research articles to promote transparent, reproducible, high-quality AI research. In the long term, further aspects, such as continuous learning, must become an integral part of these systems. Indeed, this is the foundation of all research: this initiative is

fundamentally about ensuring that this process starts correctly with the initial training and deployment and that the initial models are as robust as possible.

We build on Mitchell's²¹ definition of machine learning (ML) to define AI for the purposes of this article. We consider a technologic solution to be AI if, for a specific type of class (eg, classification of heart failure categories) and a given performance measure (eg, accuracy), the AI model's performance, as measured by the given performance measure, improves with experience (eg, additional data) and model parameters can be changed to further improve model performance. Authors of clinical, preclinical, or basic science research papers submitted to AHA journals that involve the training of any AI solution are required to submit a completed TRUE-AIM report card as part of the first revision. Authors are encouraged to fill out the report card at the time of the initial submission and follow the principles outlined in this article when preparing the original submission. Compliance with the TRUE-AIM initiative and report card will be confirmed by the journal's technical reviewers, and, where necessary, the editors and original reviewers. Any other relevant checklists corresponding to the study design and specific methods (eg, CONSORT [Consolidated Standards of Reporting Trials], STROBE [Strengthening the Reporting of Observational Studies in Epidemiology]) should also be submitted as part of the first revision.

The report card is divided into 7 sections. Each section covers a specific area of model development and evaluation. Each section has specific questions with corresponding descriptions. Authors are required to select the most appropriate response or provide the requested information and add additional context in the comments sections if needed. Authors are also required to provide the page number where the given topic is discussed in the submitted clean manuscript file. The final TRUE-AIM report card will be published as part of the supplemental material accompanying the article.

The first sections—Study design and Data—cover topics regarding study design, data acquisition, and software. These sections must be completed for all research articles. The other 6 sections ([A] Data generation and augmentation, [B] Preprocessing, [C] Feature generation, [D] Model development, [E] Model evaluation, and [F] Model interpretation) should be completed if any of these were performed in the analysis, regardless of whether AI solutions were used for that specific step. We discuss each topic in detail in the following sections. Question identification numbers are presented in square brackets to help readers identify sections discussing specific topics. These sections may provide further guidance on the information requested in the manuscript. For each item, authors should provide sufficient information in the text and supplemental materials to reproduce the findings.

The TRUE-AIM report card is presented in the Table. A summary of the report card and the development cycle of AI research is provided in Figure 1.

Table. The TRUE-AIM Report Card

Sections and identification numbers	Topic	Description	Response	Comment	Page no.
Study design					
1	Primary objective	What is the primary objective of the current paper?			
2	Primary outcome	What is the primary outcome variable of the current project?			
3	Software	What software and corresponding packages (with version numbers) were used for given parts of the analysis?			
4	Random seed	Was the random seed number set to a specific value before all analyses?	Yes/partially/no/not applicable		
Data training set					
101	Training set acquisition	How was data acquired for the training set?	Prospectively/retrospectively/not applicable		
102	Training set sites	How many different sites contributed to the training set?			
103	Training set characteristics	What was the sample size; proportion of race, sex, and ethnicity groups; and distribution of outcomes in the training data?			
104	Training set overlap	Are multiple instances of a sample present in the training data?	Yes/partially/no/not applicable		
105	Training set availability	Are all deidentified training data (eg, tabulated, images) acquired for study purposes of the training set needed to replicate the results available?	Publicly available/available upon request/no/not applicable		
Validation set (only answer 201 to 207 if the answer to question 200 is "yes" or "partially")					
200	Validation set	Was a validation set (separate from the test set) used during the development?	Yes/partially/no/not applicable		
201	Validation set acquisition	How were data acquired for the validation set?	Prospectively/retrospectively/not applicable		
202	Validation set sites	How many different sites contributed to the validation set?			
203	Validation set characteristics	What was the sample size; proportion of race, sex, and ethnicity groups; and distribution of outcomes in the validation data?			
204	Validation set overlap	Are multiple instances of a sample present in the validation data?	Yes/partially/no/not applicable		
205	Validation set availability	Are all deidentified validation data (eg, tabulated, images) acquired for study purposes of the validation set needed to replicate the results available?	Publicly available/available upon request/no/not applicable		
206	Validation and training set overlap	Were there instances of the same sample in both in the training and validation set?	Yes/partially/no/not applicable		
207	Validation and training set separation	How were the training and validation set separated from each other?	Spatially/temporally/random split/other		
Test set (only answer 301 to 308 if the answer to question 300 is "yes" or "partially")					
300	Test set	Does the analysis have a separate test set?	Yes/partially/no/not applicable		
301	Test set acquisition	How were data acquired for the test set?	Prospectively/retrospectively/not applicable		
302	Test set sites	How many different sites contributed to the test set?			
303	Test set characteristics	What was the sample size; proportion of race, sex, and ethnicity groups; and distribution of outcomes in the test data?			
304	Test set overlap	Are multiple instances of a sample present in the test data?	Yes/partially/no/not applicable		
305	Test set availability	Are all deidentified test data (eg, tabulated, images) acquired for study purposes of the test set needed to replicate the results available?	Publicly available/available upon request/no/not applicable		

(Continued)

Table. Continued

Sections and identification numbers	Topic	Description	Response	Comment	Page no.
306	Test and training set overlap	Were there instances of the same sample in both in the training and test set?	Yes/partially/no/not applicable		
307	Test and training set separation	How were the training and test set separated from each other?	Spatially/temporally/random split/other		
308	Test set type	Are the results from the test set considered internal or external evaluation?	Internal evaluation/external evaluation/both		
Data characteristics					
400	Missing data	What is the type of missingness in the training, validation, and test data?	Missing completely at random/missing at random/missing not at random/no missing data/not applicable		
401	Strategy for missing data	How was missingness handled in the analysis?	Imputation/exclusion of instance or feature/statistical modeling/other/not applicable		
402	Class imbalance	Were the data imbalanced requiring resampling or other measures?	Yes/partially/no/not applicable		
403	Strategy for class imbalance	How was imbalance handled?			
(A) Data generation and augmentation (only answer A1 to A4 if the answer to question A is "yes using AI," "yes without AI," or "partially")					
A	Data generation and augmentation	Were new instances created, or data augmented?	Yes using AI/yes without AI/partially/No		
A1	Generated data availability	Are all newly created instances by the generative model available?	Publicly available/available upon request/no/not applicable		
A2	Bias mitigation	Were measures taken to determine whether the generated data set was representative and free from bias?	Yes/partially/no/not applicable		
A3	Train-time augmentation	Was data augmentation done during any part of model development?	Yes/partially/no/not applicable		
A4	Code availability	Is all code used for data generation or augmentation purposes available?	Publicly available/available upon request/no/not applicable		
(B) Preprocessing (only answer B1 to B9 if the answer to question B is "yes using AI," "yes without AI," or "partially")					
B	Preprocessing	Were preprocessing steps done before analysis?	Yes using AI/yes without AI/partially/no		
B1	Feature scaling	What type of feature scaling was applied before analysis?			
B2	Image preprocessing	What preprocessing steps were done before image analysis?			
B3	Collinearity assessment	Were actions taken to address multicollinearity of input features?	Yes/partially/no/not applicable		
B4	Feature selection	Were features filtered before evaluation?	Supervised/unsupervised/other/not applicable		
B5	Dimensionality reduction	Were dimensionality reduction techniques used to reduce the number of features?	Supervised/unsupervised/other/not applicable		
B6	Hyperparameters	How were the hyperparameters of the preprocessing pipeline optimized?	Search among possible values/default values/preselected/not applicable		
B7	Reproducibility	Was reproducibility of the preprocessing pipeline evaluated?	Yes/partially/no/not applicable		
B8	Preprocessing data set	Were instances used to evaluate model performance part of the data set used to develop the preprocessing pipeline?	Yes/partially/no/not applicable		
B9	Code availability	Is all code used for preprocessing purposes available?	Publicly available/available upon request/no/not applicable		
(C) Feature generation (only answer C1 to C6 if the answer to question C is "yes using AI," "yes without AI," or "partially")					
C	Feature generation	Were new features created?	Yes using AI/yes without AI/partially/no		
C1	Methodology	What method was used for generating new features?	Supervised/unsupervised/other/not applicable		

(Continued)

Table. Continued

Sections and identification numbers	Topic	Description	Response	Comment	Page no.
C2	Optimization	On which data was the feature generation conducted?	Separate training set/cross-validation/whole data set/other/not applicable		
C3	Hyperparameters	How were the hyperparameters of the feature generation models optimized?	Search among possible values/default values/preselected/not applicable		
C4	Reproducibility	Was reproducibility of new generated features evaluated?	Yes/partially/no/not applicable		
C5	Feature generation data set overlap	Were instances used to evaluate model performance part of the data set used for feature development?	Yes/partially/no/not applicable		
C6	Code availability	Is all code used for feature generation purposes available?	Publicly available/available upon request/no/not applicable		
(D) Model development (only answer D1 to D11 if the answer to question D is “yes using AI,” “yes without AI,” or “partially”)					
D	Model development	Were statistical models created?	Yes using AI/yes without AI/partially/no		
D1	Model initiation	Was a pretrained model used to initiate values of the model?	Yes/partially/no/not applicable		
D2	Methodology	What method was used for model building?	Supervised/unsupervised/other/not applicable		
D3	Optimization	On which data were the model measures (including cutoff values) optimized?	Validation set/cross-validation/whole data set/other/not applicable		
D4	Hyperparameters	How were the hyperparameters of the models optimized?	Search among possible values/default values/preselected/not applicable		
D5	Test-time augmentation and ensemble techniques	Were methods used to further improve the prediction performance?	Yes/partially/no/not applicable		
D6	Overfitting	How was potential overfitting addressed?			
D7	Bias mitigation	How were potential model biases handled?			
D8	Calibration	Were models calibrated?	Yes/partially/no/not applicable		
D9	Model building data set	Were instances used to evaluate model performance part of the data set used to develop the models?	Yes/partially/no/not applicable		
D10	Code availability	Is all code used for model building purposes available?	Publicly available/available upon request/no/not applicable		
D11	Model availability	Is the fully trained model available?	Publicly available/available upon request/no/not applicable		
(E) Model evaluation (only answer E1 to E8 if the answer to question E is “yes using AI,” “yes without AI,” or “partially”)					
E	Model evaluation	Were model performances evaluated and compared?	Yes using AI/yes without AI/partially/no		
E1	Evaluation	On what data were the models evaluated?	Separate test set/cross-validation/whole data set/other/not applicable		
E2	Performance metrics	What performance metrics were used to evaluate model performance?			
E3	Model comparisons	What statistical tests were used to compare models?			
E4	Performance uncertainty	Was the uncertainty of model performance evaluated?	Yes/partially/no/not applicable		
E5	Subgroup analysis	Were model biases evaluated (performance reported in different age, sex, race, and ethnicity groups)?	Yes/partially/no/not applicable		
E6	Superiority of generated features	Were models including newly generated features compared with simpler models including clinical or quantitative features readily available?	Yes/partially/no/not applicable		
E7	Superiority of machine learning models	Was the performance of machine learning models compared with simpler statistical models (eg, regression)?	Yes/partially/no/not applicable		
E8	Code availability	Is all code used for model evaluation purposes available?	Publicly available/available upon request/no/not applicable		

(Continued)

Table. Continued

Sections and identification numbers	Topic	Description	Response	Comment	Page no.
(F) Model interpretation (only answer F1 to F4 if the answer to question F is "yes using AI," "yes without AI," or "partially")					
F	Model interpretation	Were models used to interpret results?	Yes using AI/yes without AI/partially/no		
F1	Methodology	What methods were used to help interpretation of model reasoning?			
F2	Hyperparameters	How were the hyperparameters of the models optimized?	Search among possible values/default values/preselected/not applicable		
F3	Multiple methods	Were multiple methods used to interpret the results of the models?	Yes/partially/no/not applicable		
F4	Code availability	Is all code used for model interpretation purposes available?	Publicly available/available upon request/no/not applicable		

Please choose the most appropriate answer in the response column (delete the other responses). Please provide further explanation or context in the comment column, if needed. Provide the page number where the given claim is discussed in the submitted clean manuscript file. TRUE-AIM indicates Transparent Reporting and Evaluation of Artificial Intelligence Models.

GENERAL CONSIDERATIONS FOR SCIENTIFIC ARTICLES USING AI

We refer to instances as a single occurrence of a sample or individual. Characteristics of an instance (eg, age, sex, extracted imaging markers, other variables) are referred to as features. In case of tabulated data, instances most commonly correspond to the rows of a data set (eg, patients), and the features to the columns (ie, variables). Instances may have further attributes that are not stored in tabulated format, such as images or signal outputs. The term "model" refers to a trained object that is used to process the input data and produce the predicted outcome; the term "model architecture" refers to the type and characteristics of the modeling algorithm, learning measures, and technique.

Conventional statistical methods, such as ordinary least squares regression, have a single optimal solution that can be calculated explicitly and always yield the same model for the same data. However, AI models may have several good alternative solutions, and these cannot be explicitly identified. The model needs to undergo a training process that involves the tuning of model weights and measures to minimize the difference between the model prediction and actual outcome. However, this iterative training process may result in overfitting of the model to the data that were used for development. Therefore, methodologic conventions have been developed in the AI research to provide the best real-world estimates of model performance. In AHA journals, authors are required to use the training/validation/test data set classification scheme, where training and validation data sets may be referred to as the development data set, whereas the test data set is used only after development to evaluate the model. Authors are encouraged to have all 3 data sets to allow optimal model training and evaluation. Furthermore, authors are required to take precautions to assure that there is no data leakage between the data sets (eg, the different instances of the same individual are part of the

development and test data). In all cases, the training, validation, and test sets should reflect the characteristics of the population where the model would be applied.

"Training data" refers to any data used to develop a model, through training, to predict a particular outcome. "Validation data" refers to data used to monitor model performance during development. The intention of the validation is to select the most optimal model by simulating performance on a data set that was not used for training. However, results from validation data should not be considered as validation or evaluation of the model. Because the validation set is used during optimization and tuning, its results may still reflect overfitting, and therefore such results should only be presented as supplemental material. Validation data are often referred to as "test data" in descriptions of ML processes with the concept of a training–test split. However, any data used for testing during the model development is considered validation data in the current initiative. "Test data" refers to data used to evaluate model performance after the final model is developed; these data played no part in any of the development steps. The performance of the final model should only be evaluated once on the test data. No further changes should be made to the model to further optimize the results on the test data. To conceptualize test data, consider a new system that has just been installed using the developed model. The test data consist of data from the first patients used to evaluate the performance of the model. Test data comprise almost a mini-study in their own right and are distinct from development. "Internal test data" refers to test data acquired under similar circumstances as the training data (eg, from the same hospital); "external test data" refers to test data acquired from another data source (eg, a different hospital) under different circumstances (eg, with different machinery). Terms such as "derivation," "validation," and other general terms should be avoided because they may cause confusion. Summary of data set characteristics and model development and evaluation strategies are provided in Figure 2.

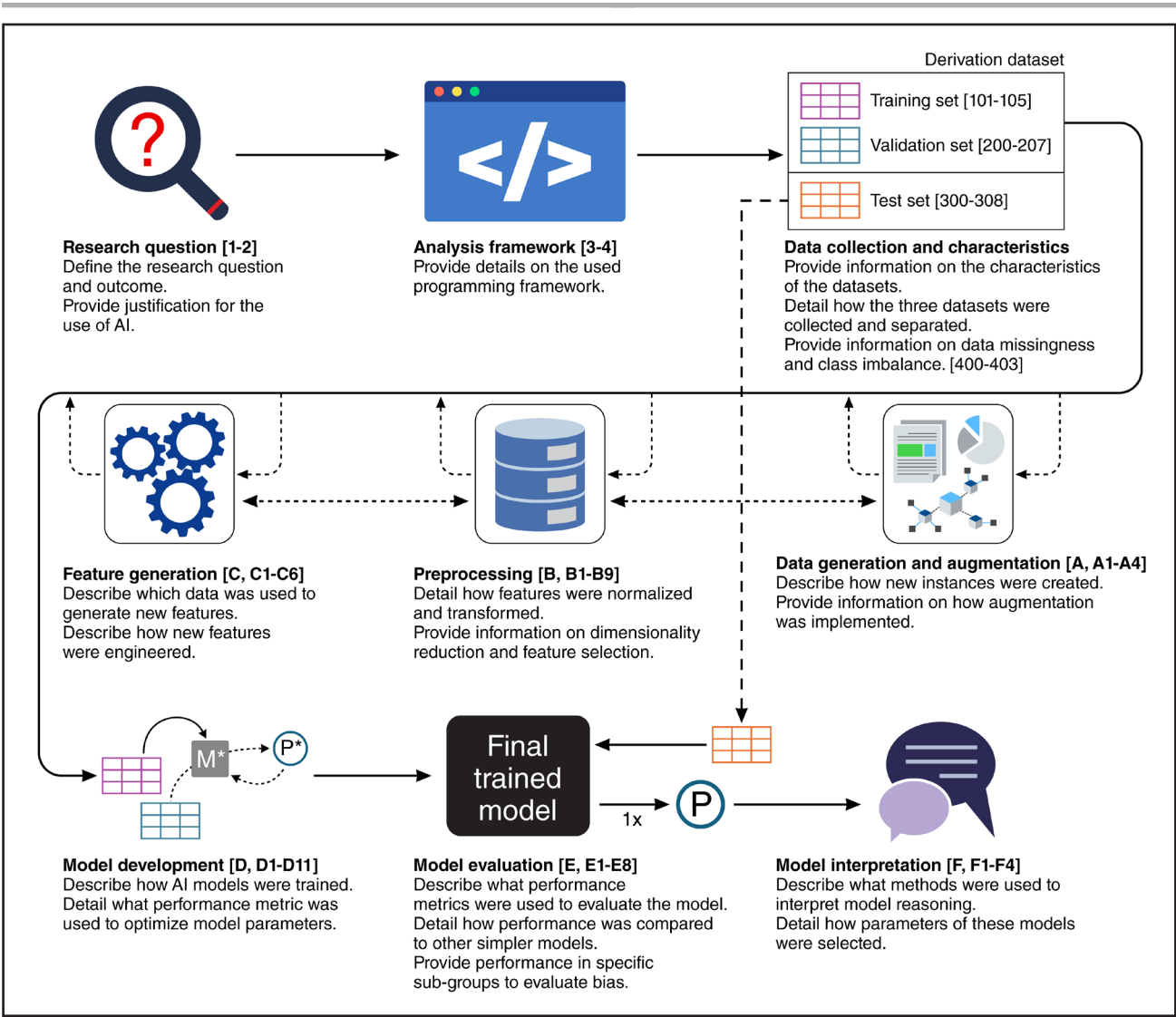


Figure 1. Overview of the TRUE-AIM framework. The TRUE-AIM (Transparent Reporting and Evaluation of Artificial Intelligence Models) report card organizes information into 7 sections. The first study design section is split into 3 pictograms (top row). Question identification numbers corresponding to given topics are provided in square brackets. General considerations regarding study design and data collection need to be filled out for all items (question identification values starting with numbers); other sections need to be filled out if that given process was done in the analysis (question identification values starting with A through F). Dashed arrows indicate optional elements in the artificial intelligence (AI) development process. These steps may be repeated or occur in a different order than that provided. The framework presents an evaluation strategy using separate training, validation, and test sets. The final performance of the model should be evaluated only once after the development process. M* indicates models during the training process; P*, performance metric of trained model on the validation set; and P, performance of the final trained model.

THE TRUE-AIM REPORT CARD

Study Design

Study design covers aspects of data acquisition and study design.

The authors should state the primary objective [1] and outcome [2] of their study in the abstract, main text, and TRUE-AIM report card. The software [3] environment (eg, Python version 3.13.5) and packages used with the corresponding version numbers (eg, PyTorch version 2.7.1) should be reported in the Methods or supplemental material. Authors are encouraged to use and share their

virtual environments to promote replication of findings. To promote reproducible model building and evaluation, authors are requested to set the random seed [4] number to a specific value before all analyses in all software packages used, leave it unchanged throughout the project, and report the number in the Methods, supplemental material, or provided computer code.

Authors are required to report on how the training data were acquired [101], including the number of sites [102] and the corresponding number of individuals or samples; the proportion by sex, race, and ethnicity; and the distribution of outcomes in the training data [103]

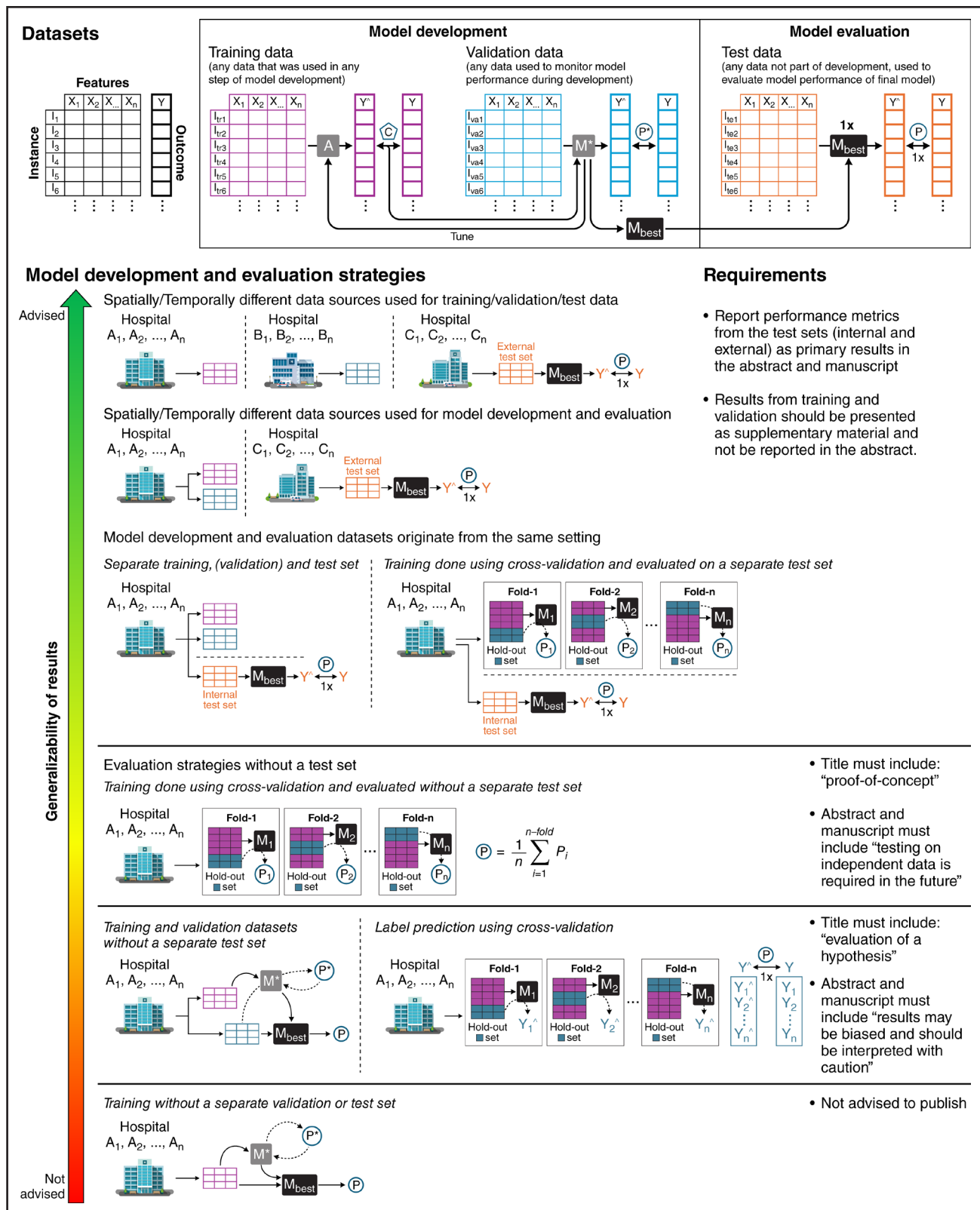


Figure 2 Continued. exist. The best practice is to have multiple different sources for each data set. In cases where data are limited, a single data source can be used, in which case the instances are assigned to the different data sets (eg, randomly). In data-scarce circumstances, cross-validation may also be used for model development and then tested on a separate test set. Even though not advised, evaluation strategies exist without separate test sets. In case of model development using cross-validation, the central tendency (eg, average) and spread (eg, standard deviation) may be used to describe projected performance. A less optimal solution is only having training and a single validation data set or obtaining predictions by cross-validation and evaluating prediction performance. Publishing analyses without a validation or test set is inadvisable. A indicates model architecture; C, cost function; I, instance; M*, models during the training process; M_{best}, best fitted model; P*, performance metric of trained model on the validation set; P, performance of the final model; X, feature; Y, outcome; and Y[^], outcome prediction.

in the Methods. A sample may have multiple instances in hierarchical (eg, segmental values of myocardial blood flow) or repeated-measure designs (eg, ejection fraction at baseline and follow-up) within individuals, or individuals may be clustered within regional areas or health care providers, which requires special attention during model building and evaluation to account for inherent clustering in the data. Authors should state whether multiple instances of the same sample or individual are included in the training data [104], and, if so, the reason. Authors are asked to state whether all deidentified training data (including all data needed to reproduce any reported analysis) are available [105]. If authors used a validation set to tune their models [200], they must respond to similar questions [201–205], and state whether there is any overlap [206] between instances in the training and validation sets. If not, then how the training and validation sets were separated [207] should be stated. Similar questions are required if a test set [300] was used for model evaluation [301–307], further requiring the statement of whether the test set is considered internal or external [308]. A data set can only be called a test set if there is no instance within the data set that was also part of the training or validation sets.

Regarding data characteristics, the manuscript should state whether there are any missing data in any data set; if there are, the assumed type of missingness [400] must be reported. How missingness was handled [401] must be reported in detail in the Methods, as well as corresponding information on why the given strategy was chosen and how potential hyperparameters (arguments) of these functions were chosen. In order to ensure that the outcomes are balanced sufficiently, the distribution of outcomes, (including any subdivisions of those outcomes, especially where outcome variations are due to different causes and thereby affecting prediction), should be described. The reasoning justifying whether this distribution is considered to be balanced or imbalanced should be provided [402]. If the data are imbalanced, depending on the degree of the imbalance and the algorithm used, balancing may or may not be required; if balancing was not used, explain the reasoning behind the choice; if balancing was required, describe the techniques used, the reasoning behind the selection, why other options were rejected, and evaluation differences between options tried [403].

Data Generation and Augmentation

All methods of creating new instances of data are considered data generation techniques. These model architectures include diffusion models and transformers.^{22,23} Data augmentation refers to methods that aim to increase the variance of the input data through modification.

Authors should state whether newly created instances of the generative model are available publicly [A1]. Furthermore, authors are encouraged to pursue methods to mitigate potential biases during data generation [A2], as these models may also identify and therefore re-create hidden features in the new data that may result in bias.²⁴ Train-time data augmentation is a commonly used method in ML to increase data heterogeneity, resulting in more generalizable models. The authors are requested to publish sufficient details regarding data augmentation (eg, image rotation ranges) performed during the training process to support the reproducibility of results [A3]. Furthermore, authors are encouraged to publish their software code used for data generation and augmentations to support reproducible research [A4].

Preprocessing

All processes before feature generation and model development are considered preprocessing. However, preprocessing steps can occur in other parts of the analysis pipeline, and should also be considered here. These steps should ideally form a pipeline, essentially sequenced into a single function that ensures that data are preprocessed consistently. Authors are encouraged to use methods throughout their development pipeline that can be implemented to new data.

Authors should report in detail whether and how data transformations and scaling (eg, Box-Cox transformation) were performed on any data before analytic steps [B1]. In case of images, the hardware used for acquisition (eg, manufacturer, model), the image acquisition protocol if appropriate (eg, type of ECG gating), and any processing steps used (eg, reconstruction kernels and algorithms) should be described. The image format, resizing (including filters or algorithms), padding, cropping, noise reduction, normalization, color spaces, color balancing, and any other steps considered during preprocessing [B2] must be reported. Whether any collinearity assessment was done [B3] and whether feature selection [B4] or dimensionality reduction [B5] techniques were used to create a more parsimonious data set should be mentioned. Many

preprocessing steps may use functions whose arguments produce different results. Therefore, authors are requested to state how hyperparameter tuning was done in the preprocessing pipeline [B6]. Hyperparameter tuning methods include grid search, random search, and Bayesian optimization, among others. If default values were chosen or based on a priori assumptions, this should be stated in the Methods. Authors are asked to state whether the robustness of the preprocessing pipeline was evaluated through reproducibility analyses [B7] and whether data used for model evaluation were also part of the data set used to develop the preprocessing pipeline [B8]. Sample-based transformations (such as standardization or principal component analysis) must be based only on the training set, with the generated measures (such as mean, standard deviation, and weights) becoming part of the preprocessing for the model and the same values used for the validation data set and, when finalized, the test data set. In cross-validation, they are generated for each fold separately. These steps guarantee that there is no information leakage between the data sets. Authors are encouraged to make their preprocessing code publicly available [B9].

Feature Generation

Feature generation is considered as any process that involves the creation of new features from gathered data. Methods for dimension reduction (eg, principal component analysis) are discussed in Preprocessing (section B); methods to create risk scores summarizing several features are discussed in Model development (section D). Feature generative techniques may include methods characterizing some quality of the input data (eg, radiomic features of images) or transforming text or image inputs to numeric values (eg, autoencoders) or creating labels using clustering (eg, k-means clustering).

Authors are asked to state in the Methods the name of the method used [C1] and from which data set these were derived [C2]. Furthermore, if the creation of these features depends on arguments of these functions (hyperparameters), how these were optimized and chosen [C3] must be detailed. Authors are encouraged to conduct reproducibility analyses regarding their new features [C4], investigating potential sources of bias regarding the robustness of the features (eg, effect of segmentation variance on radiomic feature stability). Authors are asked to state whether instances used to evaluate the performance of the model were included in the feature development process [C5]. Authors are encouraged to make their feature generation code publicly available to stimulate future research [C6].

Model Development

Model development refers to methods involving the creation of models providing an output that is directly used

to evaluate model performance or the effect size of the features. Authors should provide sufficient detail for independent centers to develop similar models on other data sets.

Authors should state whether pretrained models were used and further fine-tuned during model development [D1], or whether model architecture measures were initiated at random. The name and type of the model architecture used should be stated in the Methods (eg, random forest regression) [D2], and the rationale of choosing the given model architecture and its measures (eg, the number of hidden layers in a neural network) explained. Further information on how these models were developed and on what data they were optimized [D3] is required. Strategies may include using a separate validation set to find the most optimal model during training. However, having separate training, validation, and test data requires a great deal of data, which may not be available at all times. Therefore, instead of having separate training and validation data, cross-validation may be used for model development in situations with limited data while allowing for the test data set to have been set aside earlier. Cross-validation refers to techniques in which all steps before model evaluation are performed on a subset of the data (eg, 80% in case of 5-fold cross-validation), and predictions on the remaining data (hold-out set) provide assessment of the fold's specific model performance. This step is then repeated until each instance is part of the hold-out set once. Authors are asked to state what steps were done during training and how hyperparameters were optimized during the model development phase [D4].

Several strategies exist for splitting the data (eg, N-fold versus Leave-p-out, stratified versus unstratified, grouped versus ungrouped), which needs to be justified and stated in the Methods. The cross-validation procedure may be repeated using different data splits (repeated cross-validation) to find the optimal model with the highest performance. The selection process of the final model needs to be detailed in the manuscript and should be completed before evaluation on the test data.

Several strategies exist to further increase the performance of models. These may include modifying the input during evaluation (eg, test-time augmentation) or creating several models and combining their predictions (ensemble methods: bagging, boosting, stacking) [D5]. Sufficient information must be provided in the Methods or supplemental material to reproduce the results. Authors are also requested to state whether any strategies were used during the model development process to counter overfitting [D6] and minimize biases [D7] (eg, sex, race) that may be inherently present in the data set. Authors are encouraged to calibrate (eg, Platt scaling) their models [D8], so that the predicted probabilities are in line with actual probabilities in the data set.

The authors are requested to state explicitly whether instances used to evaluate the performance of the model were included in the model development process [D9]. Authors are encouraged to make their code [D10] and the fully trained models [D11] publicly available.

Model Evaluation

Model evaluation refers to any process involving the testing of model performance or the evaluation of the effect size of the features.

The data set on which the evaluation was done and information for choosing the given strategy in detail [E1] must be provided. Evaluation of a model should be done using a separate test set that only includes instances that were not part of any step before model evaluation. Even in cases of cross-validation, the final model performance should be evaluated on a separate test set. In case of a separate test set, performance metrics from the test sets (internal and external) should be reported as primary results in the abstract and manuscript. Results from training and validation should be presented as supplementary material and not reported in the abstract. Even though not advised, if there is no separate test set, and cross-validation was used for model development, the central tendency (mean or median) and the spread (standard deviation or 2.5% and 97.5% percentile values) of the models' performances should be reported in the abstract and manuscript. Furthermore, the title of the article must include "proof-of-concept." In the abstract and main text, "testing on independent data is required in the future" should be stated. A less optimal solution is only having training and a single validation data set or obtaining prediction by cross-validation and evaluating prediction performance. In this case, the title of the paper must include "evaluation of a hypothesis." The abstract and main text also needs to state that "results may be biased and should be interpreted with caution." Publishing articles in which a model is created on a data set and then evaluated on the same data is inadvisable (Figure 2).

There cannot be any data leakage between model development and evaluation in any evaluation strategy (eg, the same individual is present in the development and test data set, or an optimal measures cutoff is identified using the Youden index and tested on the individuals who were also used to estimate this value). Necessary precautions should be taken to avoid any circumstances where data or information may leak from any developmental step to model evaluation. Final model performance should only be evaluated once on the test set at the end of the research project. No further modifications to improve of the performance of the model on the test set are allowed.

Authors are requested to state what performance metrics [E2] were used to evaluate model performance

(eg, accuracy, area under the curve) and what statistical tests were used to compare these among different models [E3]. Authors are encouraged to compare models regarding efficiency and complexity and requested to state whether measures were taken to estimate performance uncertainty [E4] (eg, estimating confidence intervals using statistical models, estimating performance variance using bootstrapping). Because generalizability of models is especially important in medical fields, authors are required to state whether subanalyses were done during evaluation to assess whether different performance metrics (eg, sensitivity, specificity) are similar in different cohorts (eg, by age, by sex, among patients of different races or ethnicities) to better understand potential biases present in the model [E5].

If newly generated features are used in the manuscript, the authors are encouraged to compare their models including these features with nested models including clinical or quantitative features readily available in clinical practice [E6] to emphasize the additive value of their findings. If ML models were used for model development, authors are encouraged to compare their results with those using simpler statistical models, such as linear or logistic regression [E7], to determine whether ML models were superior to more conventional methods. Authors are strongly encouraged to make all their code used for model evaluation publicly available to support reproducible research [E8].

Model Interpretation

Model interpretation techniques are those that attempt to infer the importance of the different input features to help understand which data contribute the most to the predictions of the given ML model. The use of explainability methods to make AI models less of a "black box" (eg, local interpretable model-agnostic explanations, Shapley additive explanations, gradient-weighted class activation mapping) is encouraged.

Authors are requested to state which methods were used to help with the interpretation of the models [F1] as well as the methods used to set the potential arguments for these models (eg, the scoring method and number of repeats for permutation-based importance metrics) [F2]. The authors are encouraged to use several different methods, because different models may provide different results [F3].²⁵ Authors are also encouraged to make their code used for model interpretation publicly available [F4].

CONCLUSIONS

The TRUE-AIM report card aims to increase the transparency of AI research and promote reproducible research that will result in better AI applications in the medical field. By providing an at-a-glance standardized presentation regarding the most important methodologic

aspects of the research, the report card is expected to increase the transparency of reporting of AI results in medicine, making the latest AI results more accessible. In all AHA journals, a completed TRUE-AIM report card will be published as supplementary material for all articles that involve the training of any AI solution.

ARTICLE INFORMATION

Received February 17, 2025; accepted September 2, 2025.

Affiliations

Gottsegen National Cardiovascular Center, Budapest, Hungary (M.K.). Physiological Controls Research Center, University Research and Innovation Center, Óbuda University, Budapest, Hungary (M.K.). Calgary Stroke Program, Department of Clinical Neurosciences, University of Calgary, Alberta, Canada (E.E.S.). Department of Epidemiology and Biostatistics, Michigan State University, East Lansing (M.J.R.). Department of Applied Mathematics, Department of Biology, Cheriton School of Computer Science, School of Pharmacology, University of Waterloo, Ontario, Canada (A.L.). BHF Centre for Cardiovascular Science, University of Edinburgh, UK (N.L.M.). Manchester University NHS Foundation Trust, UK (M.M.). Department of Medicine, Tokai University School of Medicine, Japan (S.G.). Meddaval Ltd, Dublin, Ireland (N.A.).

Sources of Funding

Dr Kolossváry was supported by the János Bolyai Research Scholarship of the Hungarian Academy of Sciences and by the 2025-2.1.1-EKÖP-2025-00019-FK-127 University Research Scholarship Program of the Ministry of Culture and Innovation from the National Research, Development, and Innovation Fund of Hungary. Project PD147269 and Excellence 151118 have been implemented with the support provided by the Ministry of Culture and Innovation of Hungary from the National Research, Development and Innovation Fund, financed under the PD_23 and Excellence_24 funding scheme. Dr Mills is supported by a Chair Award, Programme Grant, and Research Excellence Award (grants CH/F/21/90010, RG/20/10/34966, and RE/24/130012) from the British Heart Foundation. Dr Layton was supported by the Canada 150 Research Chair program, the National Sciences and Engineering Research Council of Canada Discovery Grant (grant RGPIN- 2025-03958), and the Canada Institutes of Health Research Project Grant (grant TNC-174963). Dr Goto acknowledges funding support from the Strategic Program for Innovative Research Field 1 for Super-computational Life Science, MEXT/JSPS KAKENHI Grant-in-Aid (grants 19H03661), AMED (grants A368TS and A447TR), the Fukuda Memorial Foundation, a grant from the Nakatani Foundation for Advancement of Measuring Technologies in Biomedical Engineering, Suzuken Memorial Foundation, and the Cross-ministerial Strategic Innovation Promotion Program on "Integrated Health Care System" (grant JPJ012425). Dr Reeves and N. Atkins report no disclosures.

Disclosures

Dr Kolossváry receives consultation fees from Elucid Bioimaging Inc. Dr Goto received modest personal fees from Amgen, Merck Sharp & Dohme (MSD), Janssen Pharma, and Antos Therapeutics, and received quality fees from the American Heart Association as an Associate Editor for *Circulation* and from Duke University and Harvard University as a steering committee member for clinical trials. The other authors have no conflicts of interest to disclose.

REFERENCES

- Russell S, Norvig P. *Artificial Intelligence: A Modern Approach*. Pearson; 2016.
- Rajpurkar P, Chen E, Banerjee O, Topol EJ. AI in health and medicine. *Nat Med*. 2022;28:31–38. doi: 10.1038/s41591-021-01614-0
- Baker M. 1,500 scientists lift the lid on reproducibility. *Nature*. 2016;533:452–454. doi: 10.1038/533452a
- Lenharo M. The testing of AI in medicine is a mess: here's how it should be done. *Nature*. 2024;632:722–724. doi: 10.1038/d41586-024-02675-0
- Chouffani El Fassi S, Abdullah A, Fang Y, Natarajan S, Masroor AB, Kayali N, Prakash S, Henderson GE. Not all AI health tools with regulatory authorization are clinically validated. *Nat Med*. 2024;30:2718–2720. doi: 10.1038/s41591-024-03203-3
- Geis JR, Brady AP, Wu CC, Spencer J, Ranschaert E, Jaremko JL, Langer SG, Kitts AB, Birch J, Shields WF, et al. Ethics of artificial intelligence in radiology: summary of the Joint European and North American Multisociety Statement. *Can Assoc Radiol J*. 2019;70:329–334. doi: 10.1016/j.carj.2019.08.010
- Lyell D, Wang Y, Coiera E, Magrabi F. More than algorithms: an analysis of safety events involving ML-enabled medical devices reported to the FDA. *J Am Med Inform Assoc*. 2023;30:1227–1236. doi: 10.1093/jamia/ocad065
- Wu E, Wu K, Daneshjou R, Ouyang D, Ho DE, Zou J. How medical AI devices are evaluated: limitations and recommendations from an analysis of FDA approvals. *Nat Med*. 2021;27:582–584. doi: 10.1038/s41591-021-01312-x
- Klontzas ME, Gatti AA, Tejani AS, Kahn CE Jr. AI reporting guidelines: how to select the best one for your research. *Radiol Artif Intell*. 2023;5:e230055. doi: 10.1148/ryai.230055
- Kolbinger FR, Veldhuizen GP, Zhu J, Truhn D, Kather JN. Reporting guidelines in medical artificial intelligence: a systematic review and meta-analysis. *Commun Med (Lond)*. 2024;4:71. doi: 10.1038/s43856-024-00492-0
- Liu X, Cruz Rivera S, Moher D, Calvert MJ, Denniston AK, Spirit AI; SPIRIT-AI and CONSORT-AI Working Group. Reporting guidelines for clinical trial reports for interventions involving artificial intelligence: the CONSORT-AI extension. *Nat Med*. 2020;26:1364–1374. doi: 10.1038/s41591-020-1034-x
- Collins GS, Moons KGM, Dhiman P, Riley RD, Beam AL, Van Calster B, Ghassemi M, Liu X, Reitsma JB, van Smeden M, et al. TRIPOD-AI statement: updated guidance for reporting clinical prediction models that use regression or machine learning methods. *BMJ*. 2024;385:e078378. doi: 10.1136/bmj-2023-078378
- Sounderajah V, Ashrafian H, Golub RM, Shetty S, De Fauw J, Hooft L, Moons K, Collins G, Moher D, Bossuyt PM, et al; STARD-AI Steering Committee. Developing a reporting guideline for artificial intelligence-centred diagnostic test accuracy studies: the STARD-AI protocol. *BMJ Open*. 2021;11:e047709. doi: 10.1136/bmjopen-2020-047709
- Lekadir K, Osuala R, Gallin C, et al. FUTURE-AI: guiding principles and consensus recommendations for trustworthy artificial intelligence in medical imaging. *arXiv*. Preprint posted online September 20, 2021; revised July 22, 2024. doi: 10.48550/arXiv.2109.09658
- Norgeot B, Quer G, Beaulieu-Jones BK, Torkamani A, Dias R, Gianfrancesco M, Arnaout R, Kohane IS, Saria S, Topol E, et al. Minimum information about clinical artificial intelligence modeling: the MI-CLAIM checklist. *Nat Med*. 2020;26:1320–1324. doi: 10.1038/s41591-020-1041-y
- Hernandez-Boussard T, Bozkurt S, Ioannidis JPA, Shah NH. MINIMAR (Minimal Information for Medical AI Reporting): developing reporting standards for artificial intelligence in health care. *J Am Med Inform Assoc*. 2020;27:2011–2015. doi: 10.1093/jamia/ocaa088
- Mongan J, Moy L, Kahn CE Jr. Checklist for Artificial Intelligence in Medical Imaging (CLAIM): a guide for authors and reviewers. *Radiol Artif Intell*. 2020;2:e200029. doi: 10.1148/ryai.2020200029
- Collins GS, Dhiman P, Andaur Navarro CL, Ma J, Hooft L, Reitsma JB, Logullo P, Beam AL, Peng L, Van Calster B, 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:e048008. doi: 10.1136/bmjopen-2020-048008
- Kocak B, Akinci D'Antonoli T, Mercaldo N, Alberich-Bayarri A, Baessler B, Ambrosini I, Andreychenko AE, Bakas S, Beets-Tan RG, Bresslem K, et al. Methodological Radiomics Score (METRICS): a quality scoring tool for radiomics research endorsed by EuSoMI. *Insights Imaging*. 2024;15:8.
- 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;3:41. doi: 10.1038/s41746-020-0253-3
- Mitchell TM. *Machine Learning*. McGraw-Hill; 1997.
- Koohi-Moghadam M, Bae KT. Generative AI in medical imaging: applications, challenges, and ethics. *J Med Syst*. 2023;47:94. doi: 10.1007/s10916-023-01987-4
- Kim W. Seeing the unseen: advancing generative AI research in radiology. *Radiology*. 2024;311:e240935. doi: 10.1148/radiol.240935
- Gichoya JW, Banerjee I, Bhimreddy AR, Burns JL, Celi LA, Chen LC, Correa R, Dullerud N, Ghassemi M, Huang SC, et al. AI recognition of patient race in medical imaging: a modelling study. *Lancet Digit Health*. 2022;4:e406–e414. doi: 10.1016/S2589-7500(22)00063-2
- Adebayo J, Gilmer J, Muelly M, Goodfellow I, Hardt M, Kim B. Sanity checks for saliency maps. *arXiv*. Posted online 2018. doi: 10.48550/arXiv.1810.03292