











Study Design and Research Protocol for diagnostic or prognostic studies in the Age of Artificial Intelligence: A Biostatistician's Perspective

Giulia Barbati⁽¹⁾ , Patrizio Pasqualetti⁽²⁾ , Domenica Matranga⁽³⁾ , Lorenza Scotti⁽⁴⁾,
Matteo Franchi⁽⁵⁻⁶⁾ , Vittorio Simeon⁽⁷⁾ , Simona Signoriello⁽⁷⁾ , Ilaria Gandin⁽¹⁾,
Daniela Pacella⁽⁸⁾ , Annamaria Porreca⁽⁹⁾ , Danila Azzolina⁽¹⁰⁾, Paola Berchiolla⁽¹¹⁾ ,
Simona Villani⁽¹²⁻¹³⁾ 

(1) Biostatistics Unit, Department of Medical Sciences, University of Trieste

(2) Department of Public Health and Infectious Diseases, Sapienza University of Rome

(3) Department of Health Promotion, Mother and Childcare, Internal Medicine and Medical Specialties, University of Palermo

(4) Department of Translational Medicine, University of Piemonte Orientale

(5) National Centre for Healthcare Research and Pharmacoepidemiology, University of Milano-Bicocca, Milan, Italy

(6) Unit of Biostatistics, Epidemiology and Public Health, Department of Statistics and Quantitative Methods, University of Milano-Bicocca, Milan, Italy

(7) Medical Statistics Unit, Department of Mental, Physical Health and Preventive Medicine, University of Campania "Luigi Vanvitelli"

(8) Department of Public Health, University of Naples Federico II

(9) Department of Medical, Oral and Biotechnological Sciences, "G. D'Annunzio" University of Chieti

(10) Department of Medical Sciences, University of Ferrara

(11) Centre for Biostatistics, Epidemiology and Public Health, Department of Clinical and Biological Sciences, University of Torino

(12) Unit of Biostatistics and Clinical Epidemiology, Department of Public Health, Experimental and Forensic Medicine, University of Pavia

(13) Centre for Healthcare Research and Pharmacoepidemiology, University of Pavia

CORRESPONDING AUTHOR: Giulia Barbati, Biostatistics Unit, Department of Medical Sciences, University of Trieste.

E-mail: gbarbati@units.it

SUMMARY

Introduction: As the integration of Artificial Intelligence (AI) in healthcare continues to advance, the need for rigorous study design and research protocols tailored to diagnostic and prognostic studies becomes paramount.

Aim: The primary objective of this work is to highlight the biostatistician's point of view about the key points of the research protocol involving AI.

Methods: Assessing the current state-of-the-art guidelines, we outline the methodological challenges faced by biostatisticians when collaborating on research protocols in the era of AI-driven medical research.

Results: The proposed overview on research protocol involving AI elucidates key considerations in study design, encompassing evaluations of data quality, analysis of biases, methodological approaches, determination of sample size, and validation strategies tailored specifically to AI applications. This position paper underscores the pivotal role of strong statistical frameworks in ensuring the reliability, validity, and applicability of findings derived from AI-based diagnostic and prognostic models. Moreover, the paper seeks to highlight the critical importance of incorporating transparent reporting standards to enhance the reproducibility and clarity of AI-driven studies.

Conclusions: By offering a comprehensive biostatistician's viewpoint, this paper strives to significantly contribute to the methodological progression of diagnostic and prognostic studies in the era of Artificial Intelligence.

Keywords: Artificial Intelligence; Diagnostic and Prognostic studies; Research Protocol; Biostatistics.

DOI: 10.54103/2282-0930/22227

Accepted: 8th February 2024

© 2024 Barbati et al

INTRODUCTION

In July 2023, the European Medicine Agency (EMA) reported that “Artificial intelligence (AI) refers to systems that display intelligent behavior by analyzing their environment and taking actions – with some degree of autonomy – to achieve specific goals” (<https://www.ema.europa.eu/en/use-artificial-intelligence-ai-medicinal-product-lifecycle>). As an application of AI, Machine Learning (ML) enables systems to learn from data without being explicitly programmed [1]. In the following, for the sake of brevity we will refer to “AI” approaches including also ML and DL (Deep Learning, a type of ML based on artificial neural networks).

The increasing availability of digitalized healthcare data and the rapid development of big data analytic methods has made possible the recent exponential increase of applications of AI in healthcare. Three key questions derived: 1) Are AI based studies producing *more accurate* evidence with respect to the “standard” statistical methods? 2) Can we *successfully* use AI approaches to diagnose diseases and predict the prognosis? 3) Will AI *take the place* of a physician in the future even?

With respect to the first point, there is no general answer, since it is strictly related to the specific context of application, the aim of the study and the type of data. As an example, when the performance of neural networks (NN) with respect to logistic regression has been explored using tabular data, with the aim of predicting readmission for all causes in hospital one month after discharge for heart failure [2], authors concluded that the performance of NN and logistic regression models implemented with the LASSO method was the same. Interestingly, in a paper published in 1996 [3] in which the advantages and disadvantages of the application of NN and logistic regression were compared (always referring to tabular data), the author concluded that logistic regression is the best choice if the goal is to explore a possible causal relationship between a dependent variable and independent variables. Otherwise, “neural networks can be particularly useful when the primary objective is the prediction of results and important interactions or complex non-linearities exist in the data set, although these preferences are less clear if a regression modeler can model them using appropriate regression splines and interaction terms”. On the other hand, nowadays with the increasing availability of multi-modal sources of data, AI approaches could be the preferred choice with respect to standard statistical approaches, being able to handle heterogeneous data sources [4].

About the second question, where the main aim is to predict a probability of diagnosis or prognosis, the evidence currently available is probably affected by publication bias. There is a high risk that works using AI with non-positive conclusions may not have been published and therefore what is found is always in favor of successful performances. In addition, in

studies where AI is used with the goal of diagnosing disease, the weak element is often represented by the gold standard or the reference used such as diagnostic guidelines. If there are no established guidelines for the investigated diseases, how accurate the diagnosis from AI can be? This is a relevant issue, as it has been pointed out in an editorial in *Lancet Digital Health* in 2019: “how can an AI model be trained when experts themselves disagree on the correct answer to a question?”: in other words, what is the “ground truth” for AI if physicians did not agree on a diagnosis? [5].

For the third and final question, according to Jiang and coworkers’ reflection on the past, present and future of AI in medicine [6], AI will not take the place of the physicians in the future, although AI could support their decision and in some specific areas replace the clinical evaluations. The same conclusion emerges in an editorial published in the *Radiology Artificial Intelligence* magazine in 2019 [7], where the author concludes that the right question should be whether «radiologists will one day be replaced by those who use AI». In an older review, the conclusion was similar: «There is compelling evidence that medical AI can play a vital role in assisting the clinician to deliver health care efficiently in the 21st century. There is little doubt that these techniques will serve to enhance and complement the ‘medical intelligence’ of the future clinician» [8]. Nowadays the rise of “generative AI” (broadly speaking AI systems that have the ability to generate new content or data that is similar to, but not identical to, existing data) poses additional challenges about the “human role” in the process [9]. Large language models (LLMs) have demonstrated intriguing capabilities in the medical field, but they also exhibit certain limitations [10,11].

As can be seen from the above, the widespread use of AI in medicine has certainly opened a great debate in the medical community. The exponential rise of these new approaches underlines the urgent need to have both guidelines to improve the quality of research involving AI and to better report the evidence from clinical research using AI.

Therefore, the starting point for AI-based studies should be the research protocol as it is when classical statistical methods are employed. A good research protocol must meet some requirements that represent the cornerstones of the research methodology, well beyond the mere estimation of the sample size, which often seems to be the only issue in which the biostatistician should be involved. As also underlined in the EMA draft: “all requirements in the ICH E6 guideline for good clinical practice (GCP) or VICH GL9 Good Clinical Practices (veterinary) would be expected to apply to the use of AI within the context of clinical trials”.

This position paper aims to specifically highlight the biostatistician’s point of view about the key points of the protocol involving AI from the study aim to the sample size and methodological aspects. Starting from reviewing the state of the art for guidelines

currently available, the main points required in a research protocol when AI methods are involved will be discussed.

Guidelines for the use of AI in medical research: state of the art

As it is well known, the most popular and utilized guidelines for reporting the main study types are gathered in the "EQUATOR network", a website aimed at "Enhancing the QUALity and Transparency Of health Research" (<https://www.equator-network.org/>). Considering the scope of this paper, we focused our attention on the most known and applied guidelines, namely CONSORT (CONsolidated Standards of Reporting Trials) for clinical trials, STROBE (STrengthening the Reporting of OBservational studies in Epidemiology) for observational studies, STARD (who deals with STAndards for the Reporting of Diagnostic accuracy studies) and TRIPOD (Transparent Reporting of a multivariable prediction model for Individual Prognosis Or Diagnosis) for diagnostic/prognostic studies, looking for whether and how the corresponding expert boards have taken into account the issue of AI in their activity and in the delivered documents.

Starting with STROBE, the initiative developed and published several extensions after the main issue, some concerning methodological aspects (for respondent-driven sampling: STROBE-RDS, using mendelian randomization: STROBE-MR), others focusing on specific observational studies (Molecular Epidemiology: STROBE-ME; Nutritional Epidemiology: STROBE-nut), still others dedicated to a given pathology (newborn infection: STROBE-NI). Until today, the STROBE group have not published any extension concerning the role and methods of AI in observational studies.

The STARD working group produced a methodological extension to the core guideline in order to assist researchers in the design and reporting of accuracy studies that use Bayesian Latent Class Models (STARD-BLCM) as well as to guide diagnostic research in dementia (STARDdem). Differently from STROBE, the STARD group is promoting the development of STARD-AI extension, with the aim of providing recommendations for reporting "artificial intelligence-centered diagnostic test". Such initiative resulted in a paper published in BMJ Open where they described "the methods that will be used to develop STARD-AI" [12]. Up to now (2023, dec) STARD-AI has not been published.

TRIPOD is becoming the main reference for clinical researchers aiming at developing and validating multivariable diagnostic or prognostic models. At present there is only an extension devoted to clustered data (TRIPOD-Cluster, <https://www.tripod-statement.org/>). Similarly to STARD, a working group started to reflect about specificities of models based on AI and the protocol for developing TRIPOD-AI has been published

[13]. To be noted that such protocol embraces both TRIPOD and PROBAST, the latter dealing more closely with risk of bias in observational studies. Also in this case, up to now (2023, dec) TRIPOD-AI has not been published.

Among initiatives aiming at enhance the quality of clinical studies, CONSORT is the first in order of time and, after the first seminal guideline for reporting parallel group randomized trials, about twenty extensions were delivered addressing specific designs (non-inferiority trials, cluster trials,...) or non-pharmacological interventions (herbal, acupuncture, socio-psychological,...), or others. In this case, CONSORT-AI extension was fully delivered and published on BMJ [14]. This guideline, however, deals with "interventions with an AI component" and does not face methodological and statistical issues. In other terms, as trials of social or psychological interventions need to be described with some crucial peculiarities, also interventions which use AI "need to undergo rigorous, prospective evaluation to demonstrate impact on health outcomes". Accordingly, CONSORT-AI (in parallel with its companion statement for clinical trial protocols the well-known SPIRIT-AI or Standard Protocol Items: Recommendations for Interventional Trials - Artificial Intelligence extension) includes 14 new items which should be taken into consideration when the intervention is AI-based. The CONSORT-AI extension was developed through a staged consensus process, involving a literature review and expert consultation to generate 29 candidate items, which were assessed by an international multi-stakeholder group in a multiple-stage Delphi survey, finally producing the 14 new selected items. It has to be noted that CONSORT-AI does not affect the traditional statistical approach in clinical trials at all and it is focused mainly on specific points: the distinction between inclusion/exclusion criteria at the level of participants and at the level of input data, the onsite and offsite requirements to use AI in the intervention, the management of missing data which present relevant specificities in this context, the human-AI interaction which, if not standardized, could affect the generalizability of the findings.

RESEARCH PROTOCOL SECTIONS

Objectives/Endpoints

As for any kind of study, it is essential firstly to clearly define the objective of the study since it guides all subsequent phases of protocol development from the selection of the study sample to the definition of the data to be collected and the event of interest. Regarding diagnostic/prognostic studies, to define the aim of the study, the following points must be clearly specified: the source population, the predictors and the outcome of interest.

To correctly define the source population, it must be kept in mind which subjects will take advantage of the results of the diagnostic/prognostic model and therefore indicate the characteristics of the selected population (e.g. elderly, patients affected by a specific disease, general population). The identification of the source population is particularly helpful in the definition of inclusion and exclusion criteria (further details are given in the next section). Moreover, the candidate predictors or diagnostic methods that will be evaluated should be listed, for example salivary antibody biomarkers [15], genetic phenotypes [16], cerebrospinal fluid biomarkers, magnetic resonance imaging, functional imaging data [17]. Finally, the outcome should be mentioned as for example the patients' classification based on disease stage (e.g. patients with mild cognitive impairment or Alzheimer's disease) or types (e.g. heart failure subtypes) or patients' prediction of a clinical outcome (e.g. death or recurrence).

Of note, using AI methods to predict the occurrence of clinical outcomes is a methodological issue, therefore, considering the development of AI models as the main objective of the study is misleading as the aim must be of a clinical nature.

Study Design

The study protocol of AI-based studies in addition to the classical description of study settings and list of countries where data will be collected, should include the description of "the onsite and offsite requirements needed to integrate the AI intervention into the trial setting" [18]. This level of detail is requested since AI algorithms are strictly dependent on the environment in which they are developed, which significantly affects their generalizability. It is therefore essential to define the requirements to support both onsite and offsite integration of AI algorithms. The onsite and offsite requirements integrate the information needed i) to ensure the AI system application works within the environment *in which* the AI algorithm has been developed (and here we can talk about reproducibility); and ii) to ensure the AI system application work in a *different* environment than the one in which it has been developed (and here we can talk about replicability).

In a classical research protocol setting, the inclusion and exclusion criteria at the level of participants must be well detailed. In AI research protocol setting, the inclusion/exclusion criteria must be doubled to encompass the input data too. Therefore, the inclusion (or exclusion) criteria regarding data collected on the participants and analyzed through AI approaches should be reported in the protocol. If input data characteristics drive the pre-randomization eligibility, then the inclusion/exclusion criteria for input data should be specified in the protocol. In other words, the minimum requirements for input data must be detailed. For example, the resolution level for imaging data

could be a requirement of data input. It is not enough to report the inclusion/exclusion criteria but also how, when and whom will be evaluated. The risk of selection bias and loss of power is related to inclusion/exclusion criteria in pre-randomization steps or in pre-enrolment. In fact, subjects could meet the inclusion criteria at participant level, but not meet one of the inclusion criteria at input data level so that if the data is unsuitable for the use of the AI system, the participant will be excluded by enrollment. Possibly differential access to the study population is then introduced and the size of the eligible population is reduced, leading to the risk of not having a sufficiently large population from which to select the trial participants.

As concerns the general choice of the study design, conventional experimental or observational designs can be used with AI methodologies. The choice of the appropriate study design usually depends on the main study purpose. For example, if the purpose is diagnostic accuracy, designs include cross-sectional studies, case-control studies, as well as non-randomized or randomized comparative studies. Among the latter, AI techniques may be used for making diagnosis of a specific disease of interest, as compared to the standard diagnostic test. Otherwise, if the study purpose is to develop or validate a prediction model, cohort designs, ideally with prospectively collected longitudinal data, should be employed.

Type and quality of data

In the landscape of applying AI in healthcare, understanding the nuances of data types is paramount to ensuring the reliability and accuracy of outcomes. In a research protocol involving AI it is expected to have a variety of data sources higher than with classical statistical approaches. Data can be broadly categorized into structured, unstructured, and semi-structured formats. Structured data, characterized by a predefined format, includes tabular information commonly found in electronic health records (EHRs). Unstructured data, on the other hand, lacks a predetermined structure and encompasses diverse forms such as narrative clinical notes, medical images, and free-text entries. Semi-structured data falls in between, combining elements of both structured and unstructured data, often seen in documents with defined tags or fields. These various types of health data originate from a multitude of sources, each offering unique insights into patient health. EHRs stand out as a primary source of structured clinical data, capturing essential information from patient demographics to clinical measurements. Biomedical databases collect data from clinical studies and disease registries, forming a foundation for large-scale analytics. Medical imaging platforms store diagnostic images, facilitating collaboration among healthcare professionals. Genomic registries aggregate genetic information, empowering AI applications for personalized

medicine. Wearable devices and sensors provide real-time continuous monitoring data, contributing to a dynamic understanding of patient health over time. However, the accuracy of AI-driven insights hinges on the quality of the underlying data.

Quality control of data is a fundamental step that determines the precision and reliability of the results. This involves rigorous processes such as data cleaning, normalization, and validation to identify and rectify inconsistencies or errors and all these processes should be clearly described in the research protocol.

Data quality parameters can be classified according to whether they concern the outcome or the features (candidate predictors/variables). The first category includes classes overlap, label purity, and class's parity, which can cause an AI classifier to assign an observation to the wrong class. In classification/diagnostic problems, class overlap occurs if subjects from different classes are in close proximity to each other or class boundaries are overlapping with each other. Similarly, label errors or inconsistencies in labels affect the classification task and the decision made during the modelling of the data set. Noise in label assignment can originate from insufficient information, subjectivity, and coding issues. Regarding classes parity, a recent systematic review on data quality in AI models for head and neck cancer [19] suggested that models with good balance in the outcome classes had significantly higher median discrimination than those that did not adjust for classes imbalance.

The second category of data quality parameters include feature relevance, collinearity, data completeness, outlier detection and representativeness. Elimination of features that are either redundant or highly related or irrelevant is highly recommended during training and can be handled through dimensionality reduction techniques. Incomplete, inconsistent, duplicated, or missing data can cause a drastic deterioration in the predictive capacity of the AI model. It is advisable using missing data imputation techniques, choosing between a very simple approach consisting of estimating a value for a feature from observed values (like mean, median, mode or a suitable constant) and then replacing all missing values with the calculated statistic, or more robust missing data imputation techniques based on the maximum-likelihood (frequentist setting) or on the maximum posterior distribution (Bayesian setting) [20].

Finally, to ensure the validity of AI inference, it is crucial that the data accurately represents the characteristics of the target population. Evaluating representativeness involves understanding if the dataset contains all possible instances or if it is a sample of instances extracted from a larger set. In case the dataset is indeed a sample, it is important to ascertain the population size compared to the observed sample and to articulate the degree of representativeness of the sample with respect to the source population. A recent study analyzed the representativeness of U.S. cohorts utilized in training AI models, revealing a

systematic bias in the patient cohorts employed for clinical applications. In fact, seven out of ten of the examined studies relied on cohorts from only three states, while 34 states were not considered at all [21]. The effect of training on cohorts from specific geographical locations and subsequently making inferences on data from different locations could be the worsening of performance and fairness, especially in the presence of unequal geographical distribution [22].

Bias

Systematic biases may occur in every phase of the conduction of diagnostic/prognostic studies from the formulation of the research question to the AI model implementation. It is therefore important to be aware of these biases in the study protocol, to implement adequate strategies to avoid them or mitigate their effects (<https://catalogofbias.org/>). As mentioned before, it is crucial to clearly specify in the protocol inclusion and exclusion criteria of enrollment in the study. Once the source population is defined, it is also extremely important to avoid *sampling bias* that may result in an unrepresentative sample of the initial population. The solution should involve collecting data from randomly selected subjects of all the categories of interest in the reference population, emphasizing the need to precisely define these categories in the protocol. A further issue is in the subsequent data collection and pre-processing phases, where *measurement bias* may be encountered, involving poor precision/accuracy in measuring candidate predictors of outcomes (misclassification), and *exclusion bias*, where features deemed irrelevant are excluded, potentially due to extreme values or missing data. Another concern may be *label bias*, where not all modes of a variable (label) are represented in the collected data, as mentioned in the previous section on data quality.

These biases may lead to the implementation of inaccurate models since relevant predictors may not be included due to the lack of valid information or irrelevant variables selected due to erroneous measurement. Antidotes to these problems include the use of validated tools for defining outcomes and predictors and minimizing missing data. In the protocol definition, the methods used to identify outcomes and features/predictors should be accurately reported as well as the description of any pre-planned stratified analyses, if necessary.

During model development and validation, biases can also emerge from disparities between training and test sets, a crucial point for creating a robust model, and furthermore *confirmation bias* and overfitting, which are both possible and plausible in AI models. Solutions include random allocation of subjects between training and test sets and subsequent internal and external validation of the obtained algorithm. The research protocol should then detail the procedures

for defining training and test sets and describe the algorithm internal or external validation.

In model implementation, the change over time of variables' distribution in the population (covariate shift) or of the strength of the relationship between predictors and outcome (concept drift) may limit the model's predictive ability. Therefore, although no strategies are available to mitigate these biases, monitoring the model's utility over time is crucial to understand whether its use is still appropriate [23].

Methodological approaches

Adopting AI methods does not avoid establishing robust methodological approaches to ensure the reproducibility and validity of findings. To develop and implement AI techniques, a structured approach involving key steps must be used to train and select the final model. In the following, we summarize the steps that should be included in the research protocol in the "Methods" section.

Training Various Models/Algorithms: The initial step involves training multiple models/algorithms on the dataset. Commonly used algorithms (especially on tabular data) include linear discriminant analysis, logistic regression, flexible discriminant analysis, and decision trees. These models are applied to the training dataset, and their performance should be systematically compared on a test set using discrimination measures such as the Receiver Operating Characteristics (ROC) curve with the corresponding Area Under the Curve (AUC) that serves as a summary metric, indicating the classifier's ability to differentiate between positive and negative classes/targets. A higher AUC means superior model performance. In addition to discrimination measures for model evaluation, calibration measures, including Calibration Plots and indices such as the Integrated Calibration Index [24], focus on the reliability of predicted probabilities ensuring that the model's probability estimates reflect the true likelihood of outcomes. These complementary evaluations contribute to a comprehensive assessment of the models/algorithms' performance.

Model Validation Techniques. To evaluate the model's performance variability, k-fold cross-validation is commonly employed. The dataset is divided into K sections or folds, with each fold serving as the testing set at different points. For instance, in 5-fold cross-validation (K=5), the dataset is divided into five folds. During each iteration, one fold is designated as the testing set, and the remaining folds are used for training. This process is iterated until each of the five folds has been used as the testing set. This approach enables an estimation of the model's performance or accuracy, ensuring robustness. Following the initial validation using k-fold cross-validation on the training dataset, the models/algorithms' performance is further assessed using an independent testing dataset.

Selection of Final Model/Algorithm. Post-validation,

i.e. the selection of the final model/algorithm is automated, considering the best cross-validated performance metrics both in terms of discrimination and calibration. The selection process emphasizes also a balance between computational efficiency, robust performance, and the model's transferability. Ultimately, a single model is retained based on its superior performance across these criteria. To enhance reproducibility, transparency is paramount. It is necessary in the study protocol and in the successive study report to fully document the algorithms, model architectures, hyperparameters, and preprocessing steps, in order to facilitate the replication of the process by other researchers.

Validity: whenever possible, the study protocol should provide some details about the evaluation of both internal and external validity, indispensable in ascertaining the relevance and generalizability of findings. Internal validity as detailed above addresses the accuracy and consistency of predictions within a specific dataset, while external validity assesses the applicability of the model's outcomes to diverse patient populations or healthcare settings. To enhance internal validity, as described above, researchers must employ rigorous cross-validation techniques, ensuring that models are not overfitting to peculiarities within the training dataset. Furthermore, incorporating diverse datasets representative of various demographic groups helps minimize bias, enhancing the generalizability of the developed models. External validity, on the other hand, is bolstered by collaboration and data-sharing initiatives across institutions. Multi-center studies and collaborative efforts contribute to a more comprehensive understanding of the diverse factors influencing health outcomes. It is crucial to validate AI models across different healthcare environments to ascertain their utility in varied clinical scenarios.

Sample size

Regarding sample size calculation, we focus on the development of diagnostic and prognostic models, which are the primary applications of AI in healthcare. Regardless of the chosen AI approach, the essential prerequisite for constructing and validating such models is the availability of data with an appropriate sample size. This is crucial to ensure the models' robustness and accuracy in predicting various types of outcomes, whether binary, continuous, or time-to-event, both in terms of calibration and discrimination. Hence, sample size justification is an indispensable section of the research protocols to support reliable and accurate predictive models. Determining an appropriate sample size for a prognostic or diagnostic model typically involves considering the number of predictor variables and the incidence/prevalence of the outcome, with an old rule of the thumb of having at least ten events per predictor [25]. However, this simplistic approach overlooks factors such as the type, magnitude, and

potential values of the predictors, often resulting in poorly fitted models that struggle to generalize to new data. Recent simulation studies suggest that additional considerations are necessary, including the choice of modeling strategy and expected performance on out-of-sample data. Riley and colleagues [26] proposed a more comprehensive approach, incorporating expected model performance, the number of candidate predictors, and the prevalence of the outcome in the target population when calculating the sample size. This kind of approach could be a basis also when an AI algorithm is used instead of a traditional one on tabular data, and work is in progress to generalize the idea to AI tools through simulation approaches (<https://github.com/ewancarr/pmsims-iscb>). Of note, in a recent paper focused on survival prediction models, using real and simulated data, it has been shown that deep neural networks and random forests need at least from 2 to 3 times the sample size calculated according to Riley's method to achieve the performance of the reference [27]. Things become considerably more intricate when non-tabular data, such as unstructured sources like signals, images, or text, are employed in model development, often necessitating the use of Deep Learning algorithms. In this domain, essentially two approaches can be employed. The first is an "a priori" approach, which involves specifying the number of hidden layers in the neural network, the number of neurons within these hidden layers, and determining the minimum number of observations based on the activation function used [28]. This approach, however, has been specifically developed by the authors within a very particular context (discrete choice analysis), utilizing simulations and real data. On the other hand, the second approach relies on post-hoc evaluation, meaning it is applied when (at least part) of the data are already available to the researcher. This method involves empirically evaluating the performance at "small" sample sizes, allowing the extrapolation of performance as a function of the training set size. This is achieved by estimating the learning curve of the algorithm through fitting an inverse power-law function [29,30]. Some extensions are in progress in order to leverage information from publicly available data from related studies to inform the estimation process, to obtain robust estimates of the learning curve at the study planning stage [31].

DISCUSSION AND CONCLUSIONS

When the objective of the study is to predict a probability of diagnosis/prognosis the role of AI approaches is very promising, considering the increasing heterogeneity and complexity of the health data sources. In this context, our recommendations are tailored specifically to cases involving diagnostic and prognostic studies, aligning with the forthcoming TRIPOD-AI guidelines.

When instead the research question is explanatory in nature, we are just now at the very beginning of the potential AI applications in causal discovery and more research in this field is needed [32,33]. This aspect was not addressed in the current paper and is not covered by any other guidelines to our knowledge.

To summarize, we suggest that in research protocol using AI approaches as a first point the data quality control is particularly crucial since therapeutic decisions based on AI analyses can directly affect patient lives. A meticulous approach to ensuring data accuracy not only enhances the credibility of AI applications but also promotes trust among healthcare practitioners and patients. As AI continues to revolutionize healthcare, an unwavering commitment to data quality will be essential in harnessing the full potential of these transformative technologies.

Secondly, addressing biases in AI studies requires meticulous attention at every stage: protocols should transparently report the strategies used to mitigate biases, contributing to the validity and reliability of study results. Finally, open-source code sharing, and comprehensive documentation play pivotal roles in AI applications, enabling the scientific community to validate and build upon existing work. Moreover, the importance of reproducibility extends to clinical settings. Clinicians and healthcare professionals need confidence in the reliability of AI-generated insights for informed decision-making. Transparent methodologies contribute to scientific rigor and foster trust, promoting the responsible adoption of these technologies in real-world healthcare scenarios.

By offering a comprehensive biostatistician's viewpoint, this paper strives to significantly contribute to the methodological progression of diagnostic and prognostic studies in the era of Artificial Intelligence. It underscores scenarios where these methods could provide benefits over conventional approaches and identifies situations in which these approaches might yield biased results. This highlights the importance of a collaborative effort in fostering the development of trustworthy and clinically applicable AI models, with the ultimate goal of bringing substantial improvements in patient outcomes.

REFERENCES

1. Charalambides M, Flohr C, Bahadoran P, Matin RN. New international reporting guidelines for clinical trials evaluating effectiveness of artificial intelligence interventions in dermatology: strengthening the SPIRIT of robust trial reporting. *Br J Dermatol.* 2021 Mar;184(3):381–3.
2. Allam A, Nagy M, Thoma G, Krauthammer M. Neural networks versus Logistic regression for 30 days all-cause readmission prediction. *Sci Rep.* 2019 Jun 26;9(1):9277.

3. Tu JV. Advantages and disadvantages of using artificial neural networks versus logistic regression for predicting medical outcomes. *Journal of Clinical Epidemiology*. 1996 Nov;49(11):1225–31.
4. Kline A, Wang H, Li Y, Dennis S, Hutch M, Xu Z, et al. Multimodal machine learning in precision health: A scoping review. *npj Digit Med*. 2022 Nov 7;5(1):171.
5. Cook TS. Human versus machine in medicine: can scientific literature answer the question? *The Lancet Digital Health*. 2019 Oct;1(6):e246–7.
6. Jiang F, Jiang Y, Zhi H, Dong Y, Li H, Ma S, et al. Artificial intelligence in healthcare: past, present and future. *Stroke Vasc Neurol*. 2017 Dec;2(4):230–43.
7. Langlotz CP. Will Artificial Intelligence Replace Radiologists? *Radiology: Artificial Intelligence*. 2019 May;1(3):e190058.
8. Ramesh A, Kambhampati C, Monson J, Drew P. Artificial intelligence in medicine. *Ann R Coll Surg Engl*. 2004 Sep 1;86(5):334–8.
9. Eriksen AV, Möller S, Ryg J. Use of GPT-4 to Diagnose Complex Clinical Cases. *NEJM AI [Internet]*. 2023 Dec 11 [cited 2023 Dec 12];1(1). Available from: <https://ai.nejm.org/doi/10.1056/Alp2300031>
10. Salvagno M, Taccone FS, Gerli AG. Artificial intelligence hallucinations. *Crit Care*. 2023 May 10;27(1):180.
11. Alkaiissi H, McFarlane SI. Artificial Hallucinations in ChatGPT: Implications in Scientific Writing. *Cureus [Internet]*. 2023 Feb 19 [cited 2024 Jan 27]; Available from: <https://www.cureus.com/articles/138667-artificial-hallucinations-in-chatgpt-implications-in-scientific-writing>
12. Sounderajah V, Ashrafian H, Golub RM, Shetty S, De Fauw J, Hooft L, et al. Developing a reporting guideline for artificial intelligence-centred diagnostic test accuracy studies: the STARD-AI protocol. *BMJ Open*. 2021 Jun;11(6):e047709.
13. Collins GS, Dhiman P, Andaur Navarro CL, Ma J, Hooft L, Reitsma JB, 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 Jul;11(7):e048008.
14. Liu X, Rivera SC, Moher D, Calvert MJ, Denniston AK. Reporting guidelines for clinical trial reports for interventions involving artificial intelligence: the CONSORT-AI Extension. *BMJ*. 2020 Sep 9;m3164.
15. Tseng YJ, Wang YC, Hsueh PC, Wu CC. Development and validation of machine learning-based risk prediction models of oral squamous cell carcinoma using salivary autoantibody biomarkers. *BMC Oral Health*. 2022 Nov 24;22(1):534.
16. Banerjee A, Dashtban A, Chen S, Pasea L, Thygesen JH, Fatemifar G, et al. Identifying subtypes of heart failure from three electronic health record sources with machine learning: an external, prognostic, and genetic validation study. *The Lancet Digital Health*. 2023 Jun;5(6):e370–9.
17. Zhang D, Wang Y, Zhou L, Yuan H, Shen D. Multimodal classification of Alzheimer's disease and mild cognitive impairment. *NeuroImage*. 2011 Apr;55(3):856–67.
18. Cruz Rivera S, Liu X, Chan AW, Denniston AK, Calvert MJ, The SPIRIT-AI and CONSORT-AI Working Group, et al. Guidelines for clinical trial protocols for interventions involving artificial intelligence: the SPIRIT-AI extension. *Nat Med*. 2020 Sep;26(9):1351–63.
19. Adeoye J, Hui L, Su YX. Data-centric artificial intelligence in oncology: a systematic review assessing data quality in machine learning models for head and neck cancer. *J Big Data*. 2023 Mar 4;10(1):28.
20. Aste M, Boninsegna M, Freno A, Trentin E. Techniques for dealing with incomplete data: a tutorial and survey. *Pattern Anal Applic*. 2015 Feb;18(1):1–29.
21. Kaushal A, Altman R, Langlotz C. Geographic Distribution of US Cohorts Used to Train Deep Learning Algorithms. *JAMA*. 2020 Sep 22;324(12):1212.
22. Clemmensen LH, Kjærsgaard RD. Data Representativity for Machine Learning and AI Systems. 2022 [cited 2023 Dec 12]; Available from: <https://arxiv.org/abs/2203.04706>
23. Nazer LH, Zatarah R, Waldrip S, Ke JXC, Moukheiber M, Khanna AK, et al. Bias in artificial intelligence algorithms and recommendations for mitigation. Kalla M, editor. *PLOS Digit Health*. 2023 Jun 22;2(6):e0000278.
24. Austin PC, Steyerberg EW. The Integrated Calibration Index (ICI) and related metrics for quantifying the calibration of logistic regression models. *Statistics in Medicine*. 2019 Sep 20;38(21):4051–65.
25. Peduzzi P, Concato J, Kemper E, Holford TR, Feinstein AR. A simulation study of the number of events per variable in logistic regression analysis. *Journal of Clinical Epidemiology*. 1996 Dec;49(12):1373–9.
26. Riley RD, Snell KI, Ensor J, Burke DL, Harrell Jr FE, Moons KG, et al. Minimum sample size for developing a multivariable prediction model: PART II - binary and time-to-event outcomes. *Statistics in Medicine*. 2019 Mar 30;38(7):1276–96.
27. Infante G, Miceli R, Ambrogi F. Sample size and predictive performance of machine learning methods with survival data: A simulation study. *Statistics in Medicine*. 2023 Nov 10;sim.9931.
28. Alwosheel A, Van Cranenburgh S, Chorus CG. Is your dataset big enough? Sample size requirements when using artificial neural networks for discrete choice analysis. *Journal of Choice Modelling*. 2018 Sep;28:167–82.
29. Viering T, Loog M. The Shape of Learning Curves: A Review. *IEEE Trans Pattern Anal Mach Intell*. 2023 Jun 1;45(6):7799–819.
30. Figueroa RL, Zeng-Treitler Q, Kandula S, Ngo LH. Predicting sample size required for classification performance. *BMC Med Inform Decis Mak*. 2012 Dec;12(1):8.
31. Dayimu A, Simidjievski N, Demiris N, Abraham J. Sample size determination via learning-type curves. 2023 [cited 2023 Dec 12]; Available from: <https://arxiv.org/abs/2303.09575>
32. Chen RJ, Lu MY, Wang J, Williamson DFK, Rodig SJ, Lindeman NI, et al. Pathomic Fusion: An Integrated Framework for Fusing Histopathology and Genomic Features for Cancer Diagnosis and Prognosis. *IEEE*

Trans Med Imaging. 2022 Apr;41(4):757–70.
33. Ghassemi M, Oakden-Rayner L, Beam AL. The false
hope of current approaches to explainable artificial

intelligence in health care. The Lancet Digital Health.
2021 Nov;3(11):e745–50.

