Importance of sample size on the quality and utility of AI-based prediction models for healthcare

Richard D Riley, Joie Ensor, Kym I E Snell, Lucinda Archer, Rebecca Whittle, Paula Dhiman, Joseph Alderman, Xiaoxuan Liu, Laura Kirton, Jay Manson-Whitton, Maarten van Smeden, Karel G Moons, Krishnarajah Nirantharakumar, Jean-Baptiste Cazier, Alastair K Denniston, Ben Van Calster, Gary S Collins

Rigorous study design and analytical standards are required to generate reliable findings in healthcare from artificial intelligence (AI) research. One crucial but often overlooked aspect is the determination of appropriate sample sizes for studies developing AI-based prediction models for individual diagnosis or prognosis. Specifically, the number of participants and outcome events required in datasets for model training and evaluation remains inadequately addressed. Most AI studies do not provide a rationale for their chosen sample sizes and frequently rely on datasets that are inadequate for training or evaluating a clinical prediction model. Among the ten principles of Good Machine Learning Practice established by the US Food and Drug Administration, the UK Medicines and Healthcare products Regulatory Agency, and Health Canada, guidance on sample size is directly relevant to at least three principles. To reinforce this recommendation, we outline seven reasons why inadequate sample size negatively affects model training, evaluation, and performance. Using a range of examples, we illustrate these issues and discuss the potentially harmful consequences for patient care and clinical adoption. Additionally, we address challenges associated with increasing sample sizes in AI research and highlight existing approaches and software for calculating the minimum sample sizes required for model training and evaluation.

Introduction

With the growing interest in artificial intelligence (AI) for healthcare research, researchers should promote and implement rigorous standards in study design and analysis. One important design aspect requiring particular attention is sample size determination in AI studies. In contrast to other healthcare research studies, such as randomised trials, AI-based studies seldom include a justification for sample size selection.[1–8] This omission is observed in studies focused on training (developing) or testing (evaluating) AI-based models, which apply statistical or machine learning techniques to inform the diagnosis or prognosis of diseases in individuals. For example, a review of studies on machine learning-based models showed that 125 of 152 studies published in general medical journals did not include a sample size justification.[9,10] A similar pattern was observed in oncology-related articles, in which 57 of 62 models did not include sample size justification.[11] Additionally, a review of 606 COVID-19 prognostic models revealed that 67% were trained or tested using an inadequate sample size.[8]

The inability to justify sample size persists despite recommendations in the Transparent Reporting of a Multivariable Prediction Model for Individual Prognosis or Diagnosis (TRIPOD) guideline, published in 2015, which explicitly instructs authors to explain how the study size was arrived at.[12,13] Further, of the ten principles of Good Machine Learning Practice jointly developed by the US Food and Drug Administration, the UK Medicines and Healthcare products Regulatory Agency, and Health Canada, sample size guidance is directly relevant to three, including the recommendation for a "sample of adequate size in the clinical study and training and test datasets, so that results can be reasonably generalised to the population of interest. This is important to manage any bias, promote appropriate and generalisable performance across the intended patient population, assess usability, and identify circumstances where the model may underperform."

To reinforce these recommendations, in this Viewpoint, our group (which includes methodologists, clinicians, and patient representatives) explains why sample size estimation is a key design consideration for AI-based healthcare studies. We outline seven reasons why inadequate sample size has an effect on model training, evaluation, and performance. Using a range of examples, we illustrate these issues and discuss the potential consequences for patient care and clinical adoption. We also address barriers to achieving adequate sample sizes in AI healthcare research.

Effect and consequences of small sample sizes

Datasets not representative of the target population and setting

Even when a dataset constitutes a random sample of participants from the target population and setting, full representativeness remains unlikely when the sample size is small. Specifically, the case-mix variation (distribution of predictor values) might be narrow, and participants with specific characteristics could be missed, which restricts attempts to improve and assess model generalisability across relevant settings and subgroups.[14] Such limitations also inflate the problem of having insufficient information from under-represented groups and reduce opportunities to tailor or examine models (eg, based on ethnicity) for such individuals,[15,16] particularly when checking or improving model fairness.[17]

Lancet Digit Health 2025; 7: 100857

Published Online June 2, 2025 https://doi.org/10.1016/j.landig.2025.01.013

Department of Applied Health Sciences, School of Health Sciences, College of Medicine and Health (Prof R D Riley PhD, J Ensor PhD, K I E Snell PhD, L Archer PhD, R Whittle PhD, Prof K Nirantharakumar MD, Prof A K Denniston PhD), **Institute of Inflammation and Ageing** (J Alderman MBChB), **and Cancer Research UK Clinical Trials Unit** (L Kirton MSc), **University of Birmingham, Birmingham, UK; National Institute for Health and Care Research (NIHR) Birmingham Biomedical Research Centre, Birmingham, UK** (Prof R D Riley, J Ensor, K I E Snell, L Archer, R Whittle, J Alderman, X Liu PhD, Prof K Nirantharakumar, Prof A K Denniston); **Centre for Statistics in Medicine, Nuffield Department of Orthopaedics, Rheumatology and Musculoskeletal Sciences, University of Oxford, Oxford, UK** (P Dhiman PhD, Prof G S Collins PhD); **Peterhouse, Trumpington Street, University of Cambridge, Cambridge, UK** (J Manson-Whitton); **Julius Center for Health Sciences and Primary Care, University Medical Centre Utrecht, Utrecht University, Utrecht, Netherlands** (M van Smeden PhD, Prof K G Moons PhD); **Francis Crick Institute, London, UK** (Prof J-B Cazier PhD); **Department of Development and Regeneration, and Leuven Unit for Health Technology Assessment Research (LUHTAR), KU Leuven, Leuven, Belgium** (Prof B Van Calster PhD)

Correspondence to:
Prof Richard D Riley, Department of Applied Health Sciences, School of Health Sciences, College of Medicine and Health, University of Birmingham, Birmingham B15 2TT, UK
r.d.riley@bham.ac.uk

For more on **Good Machine Learning Practice guiding principles**, see http://www.fda.gov/media/153486/download

Even apparently large datasets could have small effective sample sizes for specific key groups, especially when outcome events or specific predictor values are rare. Thomassen and colleagues[18] used a large training dataset of over 23 000 participants and showed that the effective sample size dropped below 30 for some individuals defined by a particular set of predictor values. Thus, smaller datasets exacerbate the issue.

Consequently, inadequate representation in a dataset (including insufficient diversity and inclusivity) might negatively influence model performance and face validity, leading to scant endorsement by patients, clinicians, and regulators, especially from under-represented groups.

Large uncertainty in predictor effects and their importance

Small training samples produce instability in selected model predictors,—ie, different training samples of equal size result in different selected predictors and substantial changes in how a particular predictor impacts predictions. Attempts to meaningfully explain such unstable models, therefore, become futile,[19] as parameter estimates (eg, intercept and predictor effect estimates), predictor selection strategies (eg, lasso, recursive feature elimination), and post-hoc explanation methods (eg, Locally Interpretable Model-agnostic Explanations [LIME] and Shapley values [SHAP]) likewise become unstable and potentially misleading.[20–23] Compared with regression approaches, instability tends to be greater for other AI-based methods, as these usually allow greater complexity by default and thus require larger training sample sizes.[24,25]

Thus, instability in selected predictors and their contributions to predictions weakens confidence in models among regulators or stakeholders considering implementation.

Large uncertainty in model predictions

Large model instability owing to small training datasets yields considerable uncertainty in individual predictions.[26] To illustrate this, consider the variability across 1000 predictions observed for each of nine individuals, derived from 1000 models, each trained with a different sample of equal size from the same population (figure 1). When the sample size remains small (eg, n=50 or 100 in this example), the prediction (estimated risk) might range from 0 to 1, regardless of their true (large sample) risk. In contrast, when the sample size is large (eg, n=5000), uncertainty around predictions narrows significantly, resulting in higher reliability.

Consequently, models with highly uncertain predictions are rendered unreliable; in these situations, using a single-point estimate of risk might lead to inaccurate risk communication with patients and inappropriate clinical decisions. For example, wide uncertainty intervals of the predicted risk for an individual (eg, your risk of a stroke in the next year is somewhere between 1% and 99%) render such predictions meaningless for guiding clinical decision making.

Lower discrimination performance

The predictive performance of models trained on new data with small sample sizes is typically lower than those trained on larger sample sizes. Smaller sample sizes reduce the ability to distinguish noise (unexplainable error or aleatoric uncertainty) from signal (genuine predictor–outcome relationships), which leads to increased model-based error (epistemic uncertainty) and lower predictive performance compared with using larger datasets. Figure 2 illustrates how smaller sample sizes for model training degrade model performance.[20] It shows model performance (assessed in large external data) for models trained using six different approaches using varying training sample sizes. Notably, using smaller rather than larger training datasets reduces the c-statistic (a measure of discrimination between participants with and without outcome events; also known as the area under the receiver operating characteristic curve) by approximately 0·05 (median of 0·65 *vs* 0·70 in large samples), whereas R^2 (explained variation) is halved (median of 0·075 *vs* 0·15 in large samples). These findings align with a review of simulation studies evaluating methods for developing prediction models,[27] as seven of ten studies reported improved performance across methods with increased training sample size.

Thus, lowering a model's discrimination performance reduces the potential to distinguish between individuals with and without the outcome event, and weakens its potential clinical utility (eg, to discern those who do and do not require treatment).

Miscalibrated predictions

Small sample sizes during model training increase the likelihood of generating poorly calibrated predictions in new data (ie, estimated risks might not align with observed risks in new data). This issue is presented in figure 2, which shows that small training datasets introduce substantial variability in calibration slope (ideal value of 1) and calibration-in-the-large (ideal value of 0) when models are evaluated using large external data from the same population. For example, with a sample size of 100 participants, penalised regression methods such as the lasso produce calibration slopes ranging from approximately 0 (over-prediction of risks) to 4 (under-prediction of risks).

After model training, instability in calibration can be assessed using a calibration instability plot, which helps to visualise the variation in calibration curves across, for example, 200 bootstrap models applied to the original data. Calibration instability was observed for a random forest model trained on 752 participants, in which substantial instability in individual-level predictions (figure 3A) corresponds with wide variability in calibration curves (figure 3B). This variability highlights a major concern that predictions will be poorly calibrated in new data. Additional sources of miscalibration could arise when a model is deployed in practice due to shifts in outcome incidence, case-mix, predictor effects, or predictor measurement methods.

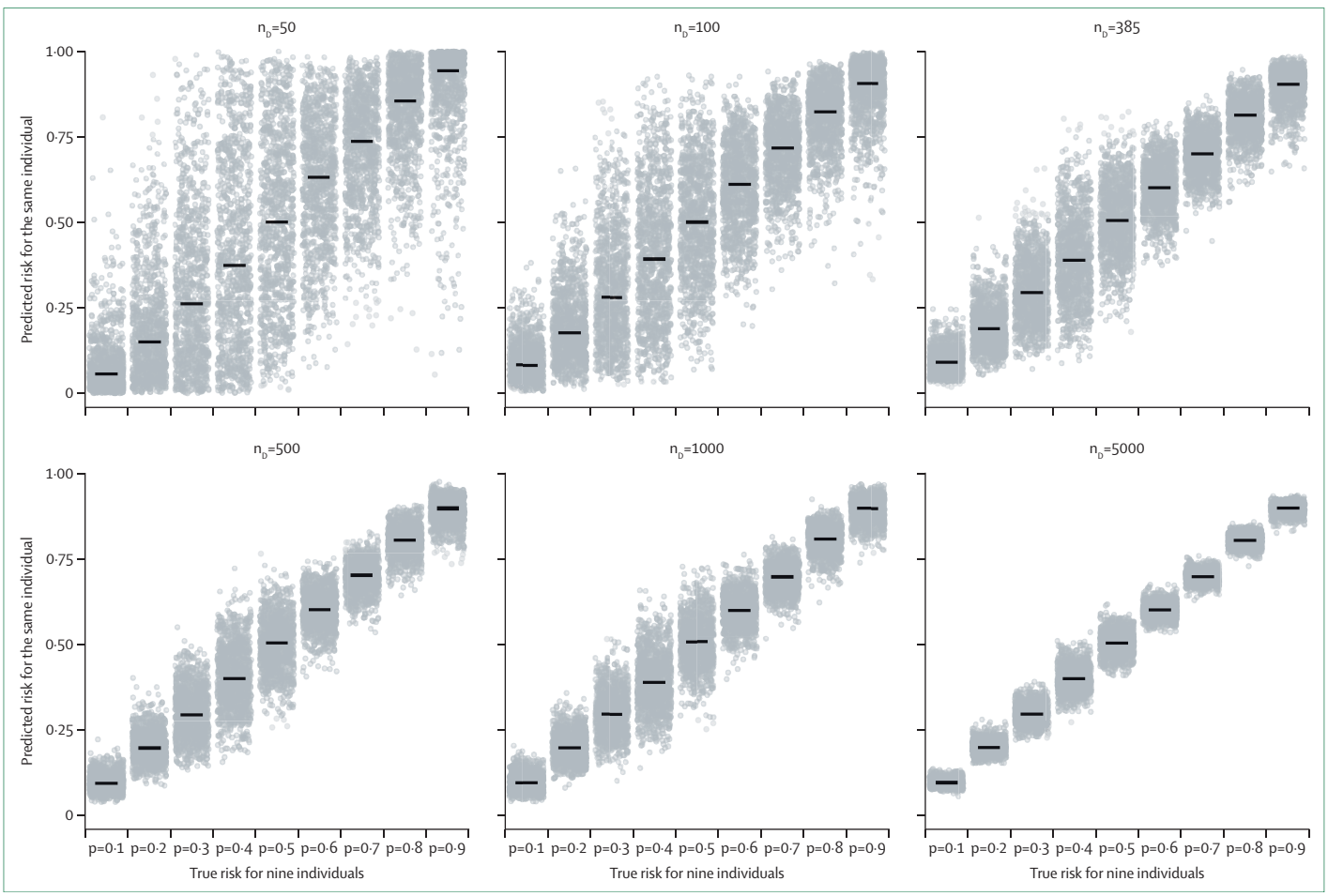

Figure 1: **Instability of estimated risks across 1000 prediction models for nine individuals with true (large-sample) risks ranging from 0·1 to 0·9, based on training sample sizes (n_D) of 50, 100, 385, 500, 1000, and 5000 participants**

Each model was developed using logistic regression with a lasso penalty, fitted to a different random sample drawn from a population with an overall risk of 0·5, using one genuine predictor (X∼N(0,4)) and 10 noise variables (Z1, …, Z10∼N(0,1)). Reproduced from Riley and Collins,[26] with permission under a CC-BY license.

When a model provides predictions that are poorly calibrated for individuals in the target population or setting, health professionals are unlikely to rely on those models for patient care and decision making.

Lower clinical utility and increased potential for incorrect clinical decisions

The poor predictive performance associated with small training sample sizes (described earlier) directly impacts clinical utility and might even cause harm. As the c-statistic decreases, the clinical utility of the model (as measured by net benefit[28]) also declines across potential risk thresholds that might be used for clinical decision making (appendix p 1).[29] Poorly calibrated models will lead to a reduction in net benefit and can lead to suboptimal decisions, such as an unnecessary prescription or withholding of treatment.[29] For example, at a risk threshold of 0·2, two miscalibrated models (ie, those that are systematically overestimating risk) yield lower net benefit than the treat-all approach (appendix p 1). In contrast, the well calibrated model provides a higher net benefit than the treat-all approach and thus would improve clinical utility by increasing benefits relative to harms for some patients.

Thus, smaller sample sizes for model training reduces clinical utility of a model, and thereby weakens the overall benefit of a model in decision making and increases the likelihood of it guiding incorrect clinical decisions compared with those made in current practice, making the model unsuitable for implementation.

Large uncertainty in test and validation performance

Model evaluation using data other than those used for model training requires a sufficiently large sample size to provide precise estimates of model performance (including calibration, discrimination, and clinical utility).[30] However, many evaluations observed in the literature used sample sizes that are considerably small,[31–33] resulting in performance estimates with wide or implausible confidence intervals and misleading claims about model reliability or superiority over alternative models.

See Online for appendix

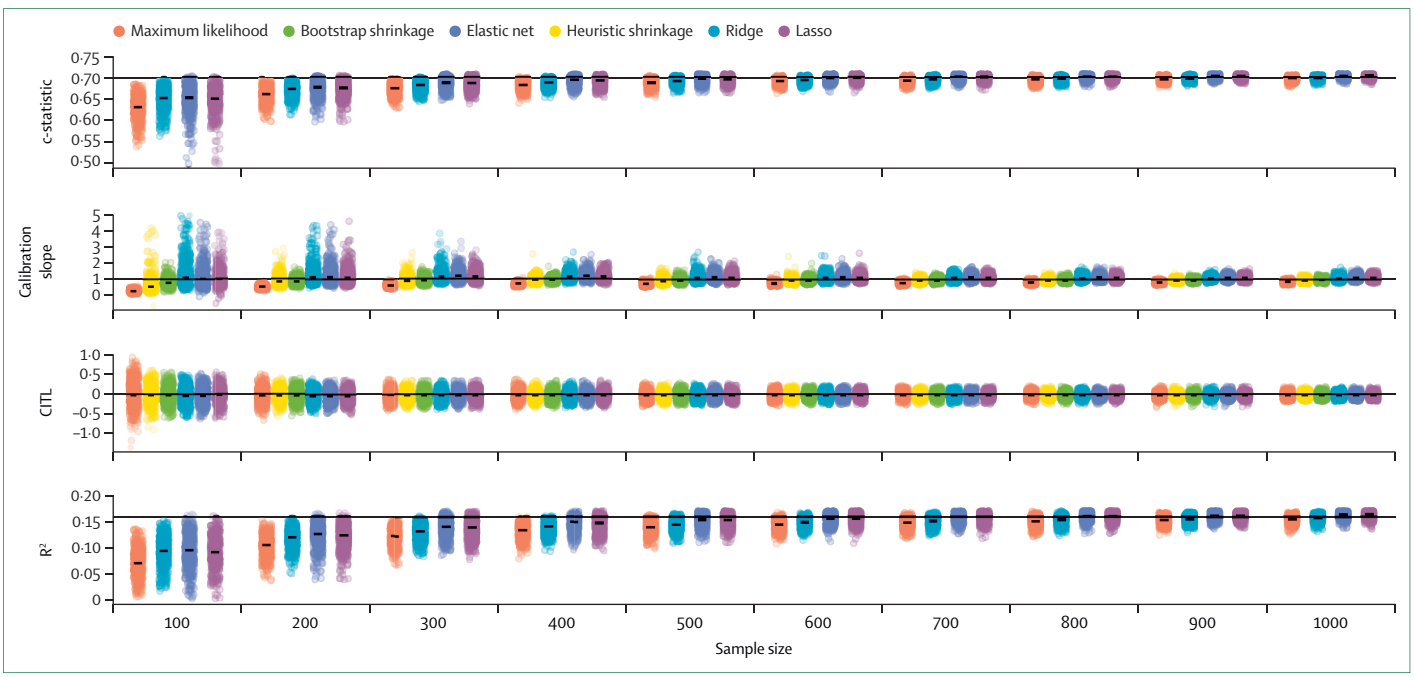

Figure 2: **Median values (short horizontal lines) and scatter plots showing variability in the predictive performance of developed models within a large validation dataset across different model training approaches and training dataset sizes**

For each sample size, 500 datasets were simulated, and for each dataset, a model was developed using the same method. Predictive performance was then tested in a large external dataset. Long horizontal lines represent performance values that would be observed had the model been developed in an extremely large sample (with almost zero epistemic uncertainty). Horizontal spread within each sample size grouping reflects random jitter for visualisation. The c-statistic is not shown for heuristic or bootstrap shrinkage, as these methods do not change the ranking of predictions compared with that using maximum likelihood estimation, and thus the c-statistic is the same. R^2 is the Nagelkerke R^2 value (which denotes the proportion of explained variation), and CITL is calibration-in-the-large. Reproduced from Riley et al,[20] with permission under a CC-BY license.

For example, a study on the development of a prediction model using XGBoost with 48 predictors claimed the model can predict death risk accurately in patients with COVID-19.[34] However, the validation dataset included only 279 participants with just seven deaths, and reported a c-statistic of 1 with an implausible 95% CI of "1·000 to 1·000", without assessing calibration. In another study, authors claimed good calibration power despite having a validation cohort with a sparse sample size of only 59 participants, of whom 19 developed the event of interest. The authors also deemed the calibration to be favourable despite substantial miscalibration evident in the calibration plot (appendix p 2).[35]

As a consequence, insufficient evidence regarding model performance should preclude endorsement by regulators or stakeholders. Small sample sizes increase the likelihood of exaggerated claims (eg, our model has been 'validated'), which potentially leads to premature or erroneous adoption of models in clinical practice.

Improving education and addressing potential barriers

Education and training are needed to emphasise the role of sample size in AI research for healthcare. In non-healthcare settings (eg, Google searches or Netflix recommendations), incorrect model predictions generally have minimal consequences at the individual level. In healthcare, such errors can result in inappropriate clinical decisions for individuals such as inappropriate treatments, hospital admissions, or monitoring frequencies, which might compromise patient outcomes, reduce quality of life, and worsen healthcare disparities.

Thus, training courses in AI and data science for healthcare research should promote a research culture that integrates statistical expertise and acknowledges the importance of robust study design, including protocol development and sample size considerations. A common but erroneous claim[20,21,36] in AI research is that sample size calculations are unnecessary or that modern methods circumvent sample size constraints. For example, a study on predicting cervical cancer screening uptake emphasised that sample size calculations were irrelevant for their purpose because classification and regression tree (CART) analysis generates non-parametric, predictive models; therefore, traditional statistical power analyses are not applicable.[37] Although traditional statistical power calculations are indeed not relevant for prediction modelling (as they estimate sample sizes based on hypothesis tests rather than model development requirements), the chosen sample size should still be justified to ensure sufficient data were used for training reliable models and precisely evaluating their performance in the intended clinical setting. Established methods exist for such calculations, whether collecting prospective data or using fixed datasets,[30,38–46] including approaches specifically

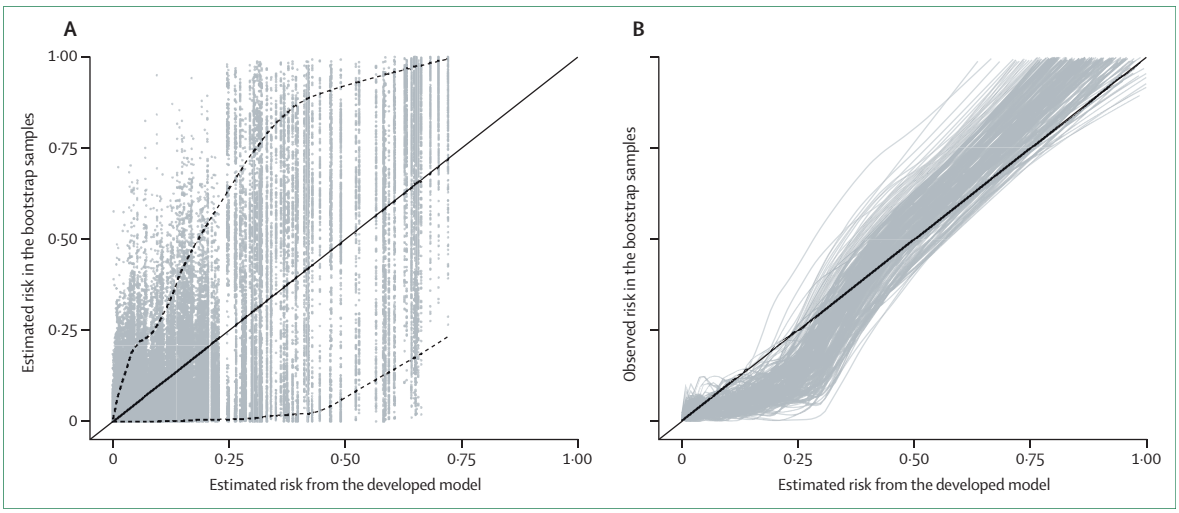

Figure 3: **Instability in predictions for a model developed using a random forest approach with seven predictors in a training dataset of 752 participants**
The model was optimised using 500 trees, with tuning of the minimum node size, and the number of candidate variables for splitting at each node, implemented via the tuneRanger package in R, with the c-statistic as the performance measure. (A) Prediction instability plot displaying the variability in predictions for each individual across 200 bootstrap models. (B) Calibration instability plot illustrating calibration curves for 200 bootstrap models applied to the original training dataset. Reproduced from the supplementary material of Riley and Collins,[26] with permission under a CC-BY license.

designed to improve model fairness (health equity) across relevant subgroups,[47] such as subgroups defined by under-represented characteristics.

For model training, at a minimum the sample size should be sufficient to estimate the overall (population) risk with precision, ensuring that the mean predicted risk aligns with the mean observed risk in the target population.[38–40] Datasets unable to meet this criterion should not be used for training models intended for individual risk prediction. To improve model stability and ensure well calibrated predictions in new data, the sample size should also help to minimise overfitting by maintaining an appropriate ratio of participants and outcome events to candidate predictor parameters of interest (and vice-versa). The *pmsampsize* module in R and Stata provides tools for these calculations. Although these methods were derived from regression theory, they serve as a foundation for other AI-based approaches, especially as models trained using non-regression approaches tend to require larger sample sizes to achieve similar stability and calibration.[24] Nevertheless, further research is required to refine sample size calculations for specific AI methodologies, including neural networks, multimodal models, and Bayesian approaches that incorporate existing knowledge.

For model evaluation (testing), the *pmvalsampsize* module in R and Stata can be used to calculate the required sample size for precise performance estimates. These calculations apply to both regression and non-regression machine learning models,[30,43–46] provided the model outputs an estimated risk (or continuous value[43]). A common practice in AI research[6] involves splitting datasets into training and testing sets (eg, 70% for training and 30% for testing). However, this approach is often inefficient,[48–50] as it reduces the sample size available for training, leading to model instability and poor calibration while also increasing imprecision in evaluation performance estimates. When feasible (eg, in regression models), using the full dataset for training and applying resampling techniques (eg, bootstrapping, cross-validation) for performance evaluation is preferable.[48] When complex models (such as deep learning methods) prohibit resampling, a split-sample approach might be necessary. In larger datasets (eg, those derived from individual participant data meta-analyses,[51] or federated learning approaches[52]), internal–external cross-validation can be used to examine performance and generalisability across clusters defined by subgroups.[14,50]

Many AI research studies rely on existing datasets that are readily available rather than recruiting new participants, as done in randomised trials or prospective cohort studies. Using an existing dataset provides a convenience sample, but researchers should not proceed before considering whether the sample size is adequate. When the existing dataset is considerably small, abstaining from using it[53,54] and instead seeking additional data is often a more appropriate course of action. For example, strategies such as combining individual participant data from multiple studies[51] or obtaining data from electronic health records in which predictors and outcomes are routinely measured could help to increase sample size. However, access to large datasets does not necessarily equate to high-quality data.

Funders should also exercise caution when supporting studies with insufficient sample sizes, as such studies might result in unreliable models, contributing to misallocated research resources or adverse effects on patients if implemented (as described in the earlier section). However, an ideal sample size cannot be expected in every setting. Obtaining a large number of participants or outcome events

For more on **pmsampsize in Stata and R**, see https://ideas.repec.org/c/boc/bocode/s458569.html and https://cran.r-project.org/web/packages/pmsampsize/index.html

For more on **pmvalsampsize in Stata and R**, see https://ideas.repec.org/c/boc/bocode/s459226.html and https://cran.r-universe.dev/pmvalsampsize

is especially challenging in rare disease research; nevertheless, prediction models are still needed in these clinical areas. Under these constraints, researchers should acknowledge increased levels of prediction uncertainty and instability or focus on developing models with only a few key predictors. A model might still provide clinical value despite moderate instability in individual-level predictions, depending on the decision-making context. For example, when prediction uncertainty occurs primarily at high-risk thresholds (eg, an uncertainty interval ranging from 0·3 to 1), the clinical impact might be minimal when decision thresholds are much lower (eg, 0·05 to 0·1). Nevertheless, transparency on the uncertainty of prediction and model instability is essential. Techniques such as instability plots and effective sample size estimates help to quantify these issues.[18,26]

Finally, increasing the sample size does not necessarily address all challenges related to model development and data quality. For example, models should account for biases inherent in current healthcare systems,[55] ensure appropriate predictor and outcome definitions (eg, standardised measurement methods and timeframes), and represent all relevant populations and settings of interest for model deployment. Even models trained on large datasets from a single population might require periodic updates (eg, due to calibration drifts) and local recalibration for use in other populations.

Conclusion

High study design standards are essential in AI research to train clinical prediction models that positively influence patients and healthcare. A key aspect of this process involves selecting appropriate sample sizes for model training and evaluation to ensure that predictions and model performance estimates are sufficiently precise to guide clinical decision making and patient–doctor discussions. Public discussions during the Standards for Data Diversity, Inclusivity, and Generalisability (STANDING Together) project highlighted that uncertainty in model performance or performance disparities between groups should not be an acceptable reason for patients receiving suboptimal care or exacerbating disparate health outcomes.[56] These considerations align with the TRIPOD+AI guideline, which includes the following reporting item: "Explain how the study size was arrived at (separately for development and evaluation), and justify that the study size was sufficient to answer the research question. Include details of any sample size calculation."

In conclusion, future AI studies in healthcare must address sample size considerations, to improve the quality of prediction model research and help ensure the development and identification of reliable models that benefit patients and clinical decision making.

For more on the **Standards for Data Diversity, Inclusivity, and Generalisability (STANDING Together) project**, see https://www.datadiversity.org/draft-standards

Contributors

RDR conceptualised the paper and drafted an initial list of items. All authors discussed this over three meetings between October, 2023, and March, 2024, providing input from methodological, clinical, and patient perspectives. RDR drafted the article, including examples and figures, with GSC and JE. Other authors (KIES, LA, RW, PD, JA, XL, LK, JM-W, MvS, KGM, KN, J-BC, AKD, and BVC) then provided feedback on the seven reasons, examples, and structure, which led to changes and additional examples and text being included. RDR and GSC then revised the manuscript for submission, and all authors approved the submitted version. After receiving reviewer comments, RDR revised the article, and all authors approved the final version.

Declaration of interests

RDR receives royalties for textbooks on Prognosis Research and Individual Participant Data Meta-Analysis; declares grants from the Engineering and Physical Sciences Research Council (EPSRC) and National Institute for Health and Care Research (NIHR) Birmingham Biomedical Research Centre (BRC); and has received an NIHR Senior Investigator Award. LA reports consulting fees from the *BMJ*. KN reports consultancy fees from Boehringer Ingelheim and CEGEDIM and roles with Network for Improving Critical Care Systems and Training, a charity, and OpenClinical, a Social Enterprise, and Dexter AI Ltd. JA has roles with The Alan Turing Institute. GSC reports an NIHR Senior Investigator Award. All other authors declare no competing interests.

Acknowledgments

RDR, KIES, LA, RW, GSC, JE, PD, J-BC, KN, and AKD are supported by an EPSRC grant for Artificial intelligence innovation to accelerate health research (EP/Y018516/1). RDR, KIES, LA, RW, JE, KN, and AKD are supported by the NIHR Birmingham BRC at the University Hospitals Birmingham NHS Foundation Trust and the University of Birmingham. GSC is supported by Cancer Research UK (CRUK; C49297/A27294). PD is supported by CRUK (PRCPJT-Nov21\100021). LK is supported by an NIHR Doctoral Fellowship (NIHR303331) and a core funding grant awarded to the Cancer Research UK Clinical Trials Unit by CRUK (CTUQQR-Dec22/100006). BVC is supported by Research Foundation–Flanders (FWO; G097322N) and Internal Funds KU Leuven (C24M/20/064). JE reports grants from the EPSRC and NIHR Birmingham BRC. KIES reports grants from the EPSRC, NIHR, and Medical Research Council (MRC). LA reports grants from the MRC, EPSRC and NIHR Birmingham BRC. RW reports grants from the MRC, CRUK, and NIHR Blood and Transplant Research Unit in Data Driven Transfusion Practice. KN reports grants from the NIHR, UK Research and Innovation/MRC, Kennedy Trust for Rheumatology Research, Health Data Research UK, Wellcome Trust, European Regional Development Fund, Institute for Global Innovation, Boehringer Ingelheim, Action Against Macular Degeneration Charity, Midlands Neuroscience Teaching and Development Funds, South Asian Health Foundation, Vifor Pharma, College of Policing, and CSL Behring. JA reports grants from the EPSRC, MRC, NIHR, and University of Birmingham QR fund. RDR, AKD, and GSC are NIHR Senior Investigators. The views expressed are those of the author(s) and not necessarily those of the NHS, NIHR, or Department of Health and Social Care.

References

1 Dhiman P, Ma J, Andaur Navarro CL, et al. Methodological conduct of prognostic prediction models developed using machine learning in oncology: a systematic review. *BMC Med Res Methodol* 2022; **22:** 101.

2 Dhiman P, Ma J, Navarro CA, et al. Reporting of prognostic clinical prediction models based on machine learning methods in oncology needs to be improved. *J Clin Epidemiol* 2021; **138:** 60–72.

3 Andaur Navarro CL, Damen JAA, Takada T, et al. Completeness of reporting of clinical prediction models developed using supervised machine learning: a systematic review. *BMC Med Res Methodol* 2022; **22:** 12.

4 Andaur Navarro CL, Damen JAA, Takada T, et al. Systematic review finds "spin" practices and poor reporting standards in studies on machine learning-based prediction models. *J Clin Epidemiol* 2023; **158:** 99–110.

5 Dhiman P, Ma J, Andaur Navarro CL, et al. Overinterpretation of findings in machine learning prediction model studies in oncology: a systematic review. *J Clin Epidemiol* 2023; **157:** 120–33.

6 Dhiman P, Ma J, Qi C, et al. Sample size requirements are not being considered in studies developing prediction models for binary outcomes: a systematic review. *BMC Med Res Methodol* 2023; **23:** 188.

7 Collins SD, Peek N, Riley RD, Martin GP. Sample sizes of prediction model studies in prostate cancer were rarely justified and often insufficient. *J Clin Epidemiol* 2021; **133:** 53–60.

8 Wynants L, Van Calster B, Collins GS, et al. Prediction models for diagnosis and prognosis of Covid-19: systematic review and critical appraisal. *BMJ* 2020; **369:** m1328.

9 Andaur Navarro CL, Damen JAA, Takada T, et al. Risk of bias in studies on prediction models developed using supervised machine learning techniques: systematic review. *BMJ* 2021; **375:** n2281.

10 Andaur Navarro CL, Damen JAA, van Smeden M, et al. Systematic review identifies the design and methodological conduct of studies on machine learning-based prediction models. *J Clin Epidemiol* 2023; **154:** 8–22.

11 Dhiman P, Ma J, Andaur Navarro CL, et al. Risk of bias of prognostic models developed using machine learning: a systematic review in oncology. *Diagn Progn Res* 2022; **6:** 13.

12 Collins GS, Reitsma JB, Altman DG, Moons KGM. Transparent Reporting of a multivariable prediction model for Individual Prognosis or Diagnosis (TRIPOD): the TRIPOD Statement. *Ann Intern Med* 2015; **162:** 55–63.

13 Moons KGM, Altman DG, Reitsma JB, et al. Transparent Reporting of a multivariable prediction model for Individual Prognosis or Diagnosis (TRIPOD): explanation and elaboration. *Ann Intern Med* 2015; **162:** W1–73.

14 Riley RD, Ensor J, Snell KIE, et al. External validation of clinical prediction models using big datasets from e-health records or IPD meta-analysis: opportunities and challenges. *BMJ* 2016; **353:** i3140.

15 Chen IY, Pierson E, Rose S, Joshi S, Ferryman K, Ghassemi M. Ethical machine learning in healthcare. *Annu Rev Biomed Data Sci* 2021; **4:** 123–44.

16 Seyyed-Kalantari L, Zhang H, McDermott MBA, Chen IY, Ghassemi M. Underdiagnosis bias of artificial intelligence algorithms applied to chest radiographs in under-served patient populations. *Nat Med* 2021; **27:** 2176–82.

17 Grote T, Keeling G. On algorithmic fairness in medical practice. *Camb Q Healthc Ethics* 2022; **31:** 83–94.

18 Thomassen D, Cessie Sl, van Houwelingen H, Steyerberg E. Effective sample size: a measure of individual uncertainty in predictions. *Stat Med* 2024; **43:** 1384–96.

19 Ghassemi M, Oakden-Rayner L, Beam AL. The false hope of current approaches to explainable artificial intelligence in health care. *Lancet Digit Health* 2021; **3:** e745–50.

20 Riley RD, Snell KIE, Martin GP, et al. Penalization and shrinkage methods produced unreliable clinical prediction models especially when sample size was small. *J Clin Epidemiol* 2021; **132:** 88–96.

21 Van Calster B, van Smeden M, De Cock B, Steyerberg EW. Regression shrinkage methods for clinical prediction models do not guarantee improved performance: simulation study. *Stat Methods Med Res* 2020; **29:** 3166–78.

22 Sauerbrei W, Boulesteix AL, Binder H. Stability investigations of multivariable regression models derived from low- and high-dimensional data. *J Biopharm Stat* 2011; **21:** 1206–31.

23 Sauerbrei W, Buchholz A, Boulesteix AL, Binder H. On stability issues in deriving multivariable regression models. *Biom J* 2015; **57:** 531–55.

24 van der Ploeg T, Austin PC, Steyerberg EW. Modern modelling techniques are data hungry: a simulation study for predicting dichotomous endpoints. *BMC Med Res Methodol* 2014; **14:** 137.

25 Infante G, Miceli R, Ambrogi F. Sample size and predictive performance of machine learning methods with survival data: A simulation study. *Stat Med* 2023; **42:** 5657–75.

26 Riley RD, Collins GS. Stability of clinical prediction models developed using statistical or machine learning methods. *Biom J* 2023; **65:** e2200302.

27 Smith H, Sweeting M, Morris T, Crowther MJ. A scoping methodological review of simulation studies comparing statistical and machine learning approaches to risk prediction for time-to-event data. *Diagn Progn Res* 2022; **6:** 10.

28 Vickers AJ, Van Calster B, Steyerberg EW. Net benefit approaches to the evaluation of prediction models, molecular markers, and diagnostic tests. *BMJ* 2016; **352:** i6.

29 Van Calster B, Vickers AJ. Calibration of risk prediction models: impact on decision-analytic performance. *Med Decis Making* 2015; **35:** 162–69.

30 Riley RD, Snell KIE, Archer L, et al. Evaluation of clinical prediction models (part 3): calculating the sample size required for an external validation study. *BMJ* 2024; **384:** e074821.

31 Collins GS, de Groot JA, Dutton S, et al. External validation of multivariable prediction models: a systematic review of methodological conduct and reporting. *BMC Med Res Methodol* 2014; **14:** 40.

32 Groot OQ, Bindels BJJ, Ogink PT, et al. Availability and reporting quality of external validations of machine-learning prediction models with orthopedic surgical outcomes: a systematic review. *Acta Orthop* 2021; **92:** 385–93.

33 Peek N, Arts DG, Bosman RJ, van der Voort PH, de Keizer NF. External validation of prognostic models for critically ill patients required substantial sample sizes. *J Clin Epidemiol* 2007; **60:** 491–501.

34 Guan X, Zhang B, Fu M, et al. Clinical and inflammatory features based machine learning model for fatal risk prediction of hospitalized COVID-19 patients: results from a retrospective cohort study. *Ann Med* 2021; **53:** 257–66.

35 Han S, Liu Y, Li X, et al. Development and validation of a preoperative nomogram for predicting benign and malignant gallbladder polypoid lesions. *Front Oncol* 2022; **12:** 800449.

36 Šinkovec H, Heinze G, Blagus R, Geroldinger A. To tune or not to tune, a case study of ridge logistic regression in small or sparse datasets. *BMC Med Res Methodol* 2021; **21:** 199.

37 Greene MZ, Hughes TL, Hanlon A, Huang L, Sommers MS, Meghani SH. Predicting cervical cancer screening among sexual minority women using Classification and Regression Tree analysis. *Prev Med Rep* 2019; **13:** 153–59.

38 Riley RD, Ensor J, Snell KIE, et al. Calculating the sample size required for developing a clinical prediction model. *BMJ* 2020; **368:** m441.

39 Riley RD, Snell KI, Ensor J, et al. Minimum sample size for developing a multivariable prediction model: part II – binary and time-to-event outcomes. *Stat Med* 2019; **38:** 1276–96.

40 Riley RD, Snell KIE, Ensor J, et al. Minimum sample size for developing a multivariable prediction model: part I – Continuous outcomes. *Stat Med* 2019; **38:** 1262–75.

41 Christodoulou E, van Smeden M, Edlinger M, et al. Adaptive sample size determination for the development of clinical prediction models. *Diagn Progn Res* 2021; **5:** 6.

42 Pavlou M, Qu C, Omar RZ, et al. Estimation of required sample size for external validation of risk models for binary outcomes. *Stat Methods Med Res* 2021; **30:** 2187–206.

43 Archer L, Snell KIE, Ensor J, Hudda MT, Collins GS, Riley RD. Minimum sample size for external validation of a clinical prediction model with a continuous outcome. *Stat Med* 2021; **40:** 133–46.

44 Snell KIE, Archer L, Ensor J, et al. External validation of clinical prediction models: simulation-based sample size calculations were more reliable than rules-of-thumb. *J Clin Epidemiol* 2021; **135:** 79–89.

45 Riley RD, Collins GS, Ensor J, et al. Minimum sample size calculations for external validation of a clinical prediction model with a time-to-event outcome. *Stat Med* 2022; **41:** 1280–95.

46 Riley RD, Debray TPA, Collins GS, et al. Minimum sample size for external validation of a clinical prediction model with a binary outcome. *Stat Med* 2021; **40:** 4230–51.

47 Riley RD, Collins GS, Whittle R, et al. Sample size for developing a prediction model with a binary outcome: targeting precise individual risk estimates to improve clinical decisions and fairness. *arXiv* 2025; published online January 25 (version 2). https://doi.org/10.48550/arXiv.2407.09293 (preprint).

48 Steyerberg EW, Harrell FEJ, Borsboom GJ, Eijkemans MJ, Vergouwe Y, Habbemaa JDF. Internal validation of predictive models: efficiency of some procedures for logistic regression analysis. *J Clin Epidemiol* 2001; **54:** 774–81.

49 Collins GS, Dhiman P, Ma J, et al. Evaluation of clinical prediction models (part 1): from development to external validation. *BMJ* 2024; **384:** e074819.

50 Steyerberg EW, Harrell FE Jr. Prediction models need appropriate internal, internal-external, and external validation. *J Clin Epidemiol* 2016; **69:** 245–47.

51 Riley RD, Tierney JF, Stewart LA, eds. Individual participant data meta-analysis: A handbook for healthcare research. John Wiley & Sons, 2021.

52 Li BS, Cai T, Duan R. Targeting underrepresented populations in precision medicine: a federated transfer learning approach. *Ann Appl Stat* 2023; **17:** 2970–92.

53 Myers PD, Ng K, Severson K, et al. Identifying unreliable predictions in clinical risk models. *NPJ Digit Med* 2020; **3:** 8.

54 Kompa B, Snoek J, Beam AL. Second opinion needed: communicating uncertainty in medical machine learning. *NPJ Digit Med* 2021; **4:** 4.

55 Nazer LH, Zatarah R, Waldrip S, et al. Bias in artificial intelligence algorithms and recommendations for mitigation. *PLoS Digit Health* 2023; **2:** e0000278.

56 Alderman JE, Palmer J, Laws E, et al. Tackling algorithmic bias and promoting transparency in health datasets: the STANDING Together consensus recommendations. *Lancet Digit Health* 2025; **7:** e64–88.