Steps to avoid overuse and misuse of machine learning in clinical research

Machine learning algorithms are a powerful tool in healthcare, but sometimes perform no better than traditional statistical techniques. Steps should be taken to ensure that algorithms are not overused or misused, in order to provide genuine benefit for patients.

Victor Volovici, Nicholas L. Syn, Ari Ercole, Joseph J. Zhao and Nan Liu

The lackluster performance of many machine learning (ML) systems in healthcare has been well documented[1,2]. In healthcare, as in other areas, AI algorithms can even perpetuate human prejudices such as sexism and racism when trained on biased datasets[3].

Given the rapid embracement of artificial intelligence (AI) and ML in clinical research and their accelerating impact, the formulation of guidelines[4,5] such as SPIRIT-AI, CONSORT-AI and, more recently, DECIDE-AI to regulate the use of ML in clinical research have helped to fill a regulatory void.

However, these clinical research guidelines generally concern the use of ML *ex post facto*, after the decision has been made to use an ML technique for a research study. The guidelines do not pose questions about the necessity or appropriateness of the AI or ML technique in the healthcare setting.

Box 1 | Recommendations to avoid overuse and misuse of AI in clinical research

1. Whenever appropriate, (predefined) sensitivity analyses using traditional statistical models should be presented alongside ML models.
2. Protocols should be published and peer reviewed whenever possible, and the choice of model should be stated and substantiated.
3. All model performance parameters should be disclosed and, ideally, the dataset and analysis script should be made public.
4. Publications using ML algorithms should be accompanied by disclaimers about their decision-making process, and their conclusions should be carefully formulated.
5. Researchers should commit to developing interpretable and transparent ML algorithms that can be subjected to checks and balances.
6. Datasets should be inspected for sources of bias and necessary steps taken to address biases.
7. The type of ML technique used should be chosen taking into account the type, size and dimensionality of the available dataset.
8. ML techniques should be avoided when dealing with very small, but readily available, convenience clinical datasets.
9. Clinician–researchers should aim to procure and utilize large, harmonized multicenter or international datasets with high-resolution data, if feasible.
10. A guideline on the choice of statistical approach, whether ML or traditional statistical techniques, would aid clinical researchers and highlight proper choices.

Failure to replicate

At the beginning of the COVID-19 pandemic, before the widespread adoption of reliable point-of-care assays to detect SARS-CoV-2, one highly active area of research involved the development of ML algorithms to estimate the probability of infection. These algorithms based their predictions on various data elements captured in electronic health records, such as chest radiographs.

Despite their promising initial validation results, the success of numerous artificial neural networks trained on chest X-rays were largely not replicated when applied to different hospital settings, in part because the models failed to learn or understand the true underlying pathology of COVID-19. Instead, they exploited shortcuts or spurious associations that reflected biologically meaningless variations in image acquisition, such as laterality markers, patient positioning or differences in radiographic projection[6]. These ML algorithms were not explainable and, while appearing to be at the cutting edge, were inferior to traditional diagnostic techniques such as RT-PCR, obviating their usefulness. More than 200 prediction models were developed for COVID-19, some using ML, and virtually all suffer from poor reporting and high risk of bias[7].

Avoiding overuse

The term 'overuse' refers to the unnecessary adoption of AI or advanced ML techniques where alternative, reliable or superior methodologies already exist. In such cases, the use of AI and ML techniques is not necessarily inappropriate or unsound, but the justification for such research is unclear or artificial: for example, a novel technique may be proposed that delivers no meaningful new answers.

Many clinical studies have employed ML techniques to achieve respectable or impressive performance, as shown by area under the curve (AUC) values between 0.80 and 0.90, or even >0.90 (Box 1). A high AUC is not necessarily a mark of quality, as the ML model might be over-fit (Fig. 1). When a traditional regression technique is applied and compared against ML algorithms, the more sophisticated ML models often offer only marginal accuracy gains, presenting a questionable trade-off between model complexity and accuracy[1,2,8–12]. Even very high AUCs are no guarantees of robustness, as an AUC of 0.99 with an overall event rate of <1% is possible, and would lead to all negative cases being predicted correctly, while the few positive events were not.

There is an important distinction between a statistically significant improvement and a clinically significant improvement in model performance. ML techniques undoubtedly offer powerful

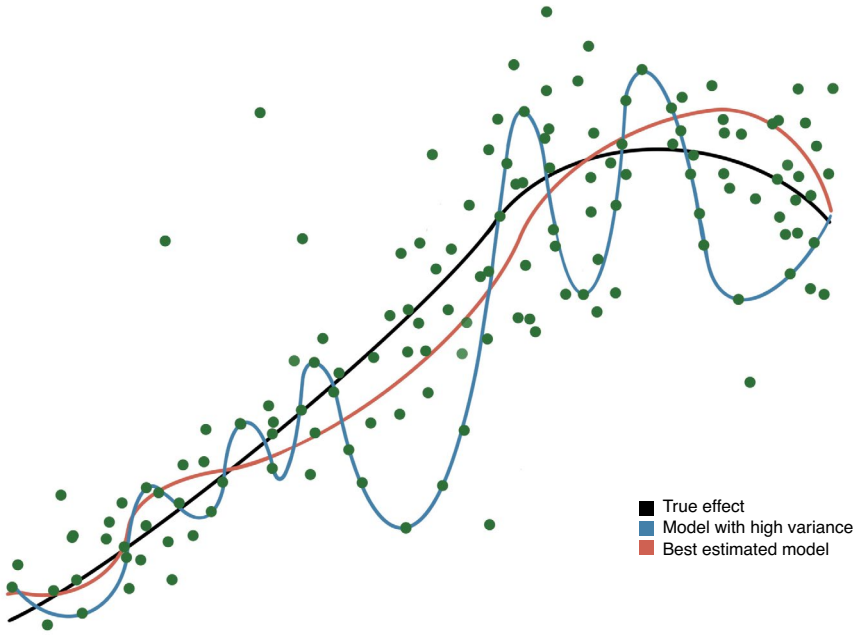

Fig. 1 | Model fitting. Given a dataset with data points (green points) and a true effect (black line), a statistical model aims to estimate the true effect. The red line exemplifies a close estimation, whereas the blue line exemplifies an overfit ML model with over-reliance on outliers. Such a model might seem to provide excellent results for this particular dataset, but fails to perform well in a different (external) dataset.

ways to deal with prediction problems involving data with nonlinear or complex, high-dimensional relationships (Table 1). By contrast, many simple medical prediction problems are inherently linear, with features that are chosen because they are known to be strong predictors, usually on the basis of prior research or mechanistic considerations. In these cases, it is unlikely that ML methods will provide a substantial improvement in discrimination[2]. Unlike in the engineering setting, where any improvement in performance may improve the system as a whole, modest improvements in medical prediction accuracy are unlikely to yield a difference in clinical action.

ML techniques should be evaluated against traditional statistical methodologies before they are deployed. If the objective of a study is to develop a predictive model, ML algorithms should be compared to a predefined set of traditional regression techniques for Brier score (an evaluation metric similar to the mean squared error, used to check the goodness of a predicted probability score), discrimination (or AUC) and calibration. The model should then be externally validated. The analytical methods, and the performance metrics on which they are being compared, should be specified in a prospective study protocol and should go beyond overall performance, discrimination and calibration to also include metrics related to over-fitting.

Conversely, some algorithms are able to say "I don't know" when faced with unfamiliar data[13], an output that is important but often underappreciated, as knowledge that a prediction is highly uncertain may, itself, be clinically actionable.

Rationalize usage

Researchers should start any ML project with clear project goals and an analysis of the advantages that AI, ML or conventional statistical techniques deliver in the specific clinical use case. Unsupervised clustering analyses tend to be well suited for discovering hidden patterns of clustering, for example to propose a novel molecular taxonomy of cancers[14] or define subtypes of a psychiatric disorder[15].

If the objective of a study is to develop a new prognostic nomogram or predictive model, there is little evidence that ML will fare better than traditional statistical models even when dealing with large and highly dimensional datasets[1,2,8–11,16–18]. If the purpose of a study is to infer a causal treatment effect of a given exposure, many well-established traditional statistical techniques, such as structural equation modelling, propensity-score methodology, instrumental variables analysis and regression discontinuity analysis, yield readily interpretable and rigorous estimates of the treatment effect.

Avoiding misuse

In contrast to overuse, the term 'misuse' connotes more egregious usages of ML, ranging from problematic methodology that engenders spurious inferences or predictions, to applications of ML that endeavor to replace the role of physicians in situations which should still require a human input.

Indiscriminately accepting an AI algorithm purely based on its performance, without scrutinizing its internal workings, represents a misuse of ML[19], although it is questionable to what extent every clinician decision is robustly explainable.

Many groups have called for explainable ML or the incorporation of counterfactual reasoning in order to disentangle correlation from causation[20]. Medicine should be based on science, and medical decisions should be substantiated by transparent and logical reasoning that may be subjected to interrogation. The notion of a 'black box' that underpins clinical decision-making is an antithesis to the modern practice of medicine and is increasingly inaccurate, given the growing armamentarium of techniques such as saliency maps and generative adversarial networks that can be used to probe the reasoning made by neural networks.

Researchers should commit to developing ML models that are interpretable, with their reasoning standing up to scrutiny by human experts, and to sharing de-identified data and scripts that would allow external replication and validation. Some researchers might conclude that machines can identify patterns in the data that the human brain cannot discern. Yet, just as an expert should be able to explain their thought patterns on complex topics, so, too, should machines be able to justify the path they took to uncover certain patterns.

Data constraints

Usage of ML in spite of data constraints, such as biased data and small datasets, is another misuse of AI. Training data can be biased and can amplify sexist and racist assumptions[3,21]. Deep learning techniques are known to require large amounts of data, but many publications in the medical literature feature techniques with much smaller sample and feature-set sizes than are typically available in other technological industries. Well-trained ML algorithms may therefore lack access to a complete description of the clinical problem of interest.

Table 1 | Definitions of several key terms in machine learning

Term	Definition
Linear relationship	A relationship between two variables where increasing or decreasing one variable *n* times will cause a corresponding increase or decrease of *n* times in the other variable as well. May be written mathematically as $y = aX + b$, where *y* and *X* are the variables of interest.
Nonlinear relationship	A relationship between two variables that cannot be plotted using a straight line, and where the direct relationship between an independent variable and a dependent variable follows a different pattern, such as $y = aX^3 + bX^2 + cX + d$.
Area under the curve (AUC)	An often-used measure of the discriminative ability of a model obtained from the receiver operating characteristic curve analysis. The AUC provides an aggregate measure of the performance of the model across all possible prediction thresholds and can be interpreted as the probability that the model will rank a random positive example (event) more highly than a random negative example. The AUC may be overly optimistic for imbalanced datasets where one of the outcomes is rare. In this case, the area under the precision-recall curve, which is a curve of positive predictive value against true positive rate, is more appropriate.
Calibration	As important as discrimination, calibration refers to the agreement between observed outcomes and predictions. For example, if a 10% chance of poor outcome is predicted, the observed frequency of actual poor outcome should be approximately 10 out of 100 patients with such a prediction. Poorly calibrated models may over- or underpredict a particular outcome under certain conditions, which may be very important in medical contexts.
Machine learning	The use and development of computer systems that are able to learn and adapt without following explicit instructions, by using algorithms to uncover patterns in data and make inferences.
Deep learning	A type of ML algorithm based on artificial neural networks. These are multiple layers of processing made up of interconnected nodes that are used to extract progressively higher-level features from highly granular data.
Regression techniques	Regression techniques attempt to find a mathematical relationship between measurements of variables *y* (dependent) and $X_1, X_2, \ldots, X_n$, such that the value of variable *y* can be predicted from a measurement of the other variables. The variable *y* usually represents an outcome of interest, such as mortality, poor outcome or length of hospital stay.
Over-fitting	Over-fitting is an error in that occurs when a mathematical function is too closely aligned to a limited set of data points. Such a model may perform well on data it has seen before but will perform poorly when presented with new, unseen data. ML algorithms are prone to over-fitting, as the goal is often to achieve near-perfect prediction in a dataset, which leads them to fail during external validation
External validation	Evaluating model efficiency and goodness of fit in a different dataset to the one it was conceived from.
Goodness of fit	The extent to which the model fits the observed data, usually measuring the discrepancy between predicted values from the model and actual measured values. Multiple parameters are available to test this desiderate.
Nomogram	A graphical representation of a mathematical formula or ML algorithm incorporating several predictors modelled to predict a particular outcome

Meta's Facebook trained its facial recognition software using photos from more than 1 billion users; autonomous automobile developers use billions of miles of road traffic video recordings from hundreds of thousands of individual drivers in order to develop software to recognize road objects; and DeepBlue and AlphaGo learn from millions or billions of played games of chess and Go. In contrast, clinical research studies involving AI generally use thousands or hundreds of radiological and pathological images[22], and surgeon–scientists developing software for surgical phase recognition often work with no more than several dozen surgical videos[23]. These observations underscore the relative poverty of big data in healthcare and the importance of working toward achieving sample sizes like those that have been attained in other industries, as well as the importance of a concerted, international big-data sharing effort for health data.

Human–machine collaboration

The respective functions of humans and algorithms in delivering healthcare are not the same. Algorithms allow clinicians to make the best use of the available data to inform practice, especially when the data have a complex structure or are both large and highly granular.

ML algorithms can complement, but not replace, physicians in most aspects of clinical medicine, from history-taking and physical examination to diagnosis, therapeutic decisions and performing procedures. Clinician–investigators must therefore forge a cohesive framework whereby big data propels a new generation of human–machine collaboration. Even the most sophisticated ML applications are likely to exist as discrete decision-support modules to support specific aspects of patient care, rather than competing against their human counterparts.

Human patients are likely to want human doctors to continue making medical decisions, no matter how well an algorithm can predict outcomes. ML should, therefore, be studied and implemented as an integral part of a complete system of care.

The clinical integration of ML and big data is poised to improve medicine. ML researchers should recognize the limits of their algorithms and models in order to prevent their overuse and misuse, which could otherwise sow distrust and cause patient harm. ❐

Victor Volovici[1]✉, Nicholas L. Syn[2,3], Ari Ercole[4], Joseph J. Zhao[2] and Nan Liu[5]

[1]*Department of Neurosurgery, Erasmus MC University Medical Center, Rotterdam, the Netherlands.* [2]*Yong Loo Lin School of Medicine, National University of Singapore, Singapore, Singapore.* [3]*Department of Surgery, National University Hospital, National University Health System, Singapore, Singapore.* [4]*Cambridge Centre for AI in Medicine, University of Cambridge, Cambridge, UK.* [5]*Programme in Health Services and Systems Research, Duke–NUS Medical School, Singapore, Singapore.*

✉*e-mail: v.volovici@erasmusmc.nl*

Published online: 12 September 2022
https://doi.org/10.1038/s41591-022-01961-6

References

1. Christodoulou, E. et al. *J. Clin. Epidemiol.* **110**, 12–22 (2019).
2. Gravesteijn, B. Y. et al. *J. Clin. Epidemiol.* **122**, 95–107 (2020).
3. Zou, J. et al. *Nature* **559**, 324–326 (2018).
4. Topol, E. J. *Nat. Med.* **26**, 1318–1320 (2020).
5. Vasey, B. et al. *Nat. Med.* **28**, 924–933 (2022).
6. DeGrave, A. J. et al. *Nat. Mach. Intell.* **3**, 610–619 (2021).
7. Wynants, L. et al. *Br. Med. J.* **369**, m1328 (2020).
8. Abràmoff, M. D. et al. *npj Digit. Med.* **1**, 39 (2018).
9. Shin, S. et al. *ESC Heart Fail.* **8**, 106–115 (2021).
10. Cho, S. M. et al. *Can. J. Cardiol.* **37**, 1207–1214 (2021).

11. Uddin, S. et al. *BMC Med. Inform. Decis. Mak.* **19**, 281 (2019).
12. Volovici, V. et al. *J. Neurotrauma* **36**, 3183–3189 (2019).
13. Shashikumar, S. P. et al. *npj Digit. Med.* **4**, 134 (2021).
14. Cancer Genome Atlas Research Network. et al. *Nat. Genet.* **45**, 1113–1120 (2013).
15. American Psychiatric Association. *Diagnostic and Statistical Manual of Mental Disorders* 5th edn (APA, 2013).
16. Futoma, J. et al. *Lancet Digit. Health* **2**, e489–e492 (2020).
17. Pisică, D. et al. *World Neurosurg.* **161**, 230–239.e236 (2022).
18. Marek, S. et al. *Nature* **603**, 654–660 (2022).
19. Rudin, C. *Nat. Mach. Intell.* **1**, 206–215 (2019).
20. Richens, J. G. et al. *Nat. Commun.* **11**, 3923 (2020).
21. Andaur Navarro, C. L. et al. *Br. Med. J.* **375**, n2281 (2021).
22. Stulberg, J. J. et al. *JAMA Surg.* **153**, 586–587 (2018).
23. Twinanda, A. P. et al. *IEEE Trans. Med. Imaging* **36**, 86–97 (2017).

Acknowledgements
We would like to thank M. van Bilsen for the figure and F. Liu for her valuable advice. V.V. wishes to thank D. Volovici for opening up the world of probability, statistics and machine learning.

Author contributions
V.V. conceived the idea, drafted the first manuscript, conceptualized the figure and supervised the work; N. S. substantially revised the manuscript and critically read all versions of the manuscript. A. E., J. J. Z. and N. L. made substantial revisions and approved the final manuscript.

Competing interests
The authors declare no competing interests.