BJPsych
The British Journal of Psychiatry (2022)
220, 169–171. doi: 10.1192/bjp.2022.29

Editorial

Five points to consider when reading a translational machine-learning paper

Dominic Dwyer and Rajeev Krishnadas

Summary

Machine-learning techniques are used in this BJPsych special issue on precision medicine in attempts to create statistical models that make clinically relevant predictions for individual patients. In this primer, we outline five key points that are helpful for a new reader to consider in order to engage with the field and evaluate the literature. These points include the consideration of why we are interested in new statistical approaches, how they may produce individualised predictions, what caveats need to be kept in-mind and why the interest and engagment of clinicians and clinical researchers is critical to successful model development and implementation. We hope that the following primer will provide shared understanding to encourage dialogue between clinical and methodological fields.

Keywords

Precision medicine; machine learning; predictive models; classification models; precision psychiatry.

Copyright and usage

Dominic Dwyer is a research fellow focussed on enhancing clinical care with machine learning in Melbourne, Australia (Orygen & The Centre for Youth Mental Health, Melbourne University) and holds an honorary appointment at Ludwig Maximilian University in Munich. Rajeev Krishnadas is a Consultant Psychiatrist and NHS Research Fellow at NHS Greater Glasgow and Clyde.

Background

For the purposes of this editorial, we define machine learning as a field where statistical models automatically improve using computer algorithms that respond to repeated experience with data. We separate this approach from more traditional approaches where a static method is used once to derive an outcome (for example a *t*-test), despite the recognition that the two fields share methods and are interconnected.

Computational power

The first point is to acknowledge why we are interested in new statistical approaches. Psychiatric research predominantly aims to guide decisions based on group averages for a majority of individuals (e.g. from a clinical trial) or provide insight regarding associations (e.g. from cohort studies). These approaches are essential, but there is an increasing recognition that obtaining a specific diagnostic-, prognostic- or treatment-response likelihood for an individual patient will help clinicians to make personalised decisions regarding care.

The machine-learning paradigm has achieved such predictions for individual examples or cases in other fields (such as speech, text or image recognition) by using statistics to solve practical problems with computers. This shift in culture, together with large advances in computational power, cast old statistical techniques in a new light and opened the door to advanced methods that are now seamlessly integrated into daily life (such as those employed by Google, Amazon, Netflix or Apple). As such, we are interested in this new statistical approach because we hope that such a pragmatic approach will fast-track personalised psychiatric treatment by providing additional tools to clinicians, clients and their families.[1]

Pattern detection

The second point is related to limitations within the existing psychiatric research culture. Traditional psychiatric methods restrict statistical choices and rely on assumptions to facilitate inferences to a population beyond the sample. Researchers design studies with such restrictions in mind, analyse data and ultimately make decisions that influence guidelines, inform our understanding of illness and identify new therapies. In general, the majority of such statistical models are either not designed to be used on individuals or, if they have, they have not been powerful enough to obtain a clinically translatable prediction that is currently used.[1]

The machine-learning field partly grew from the idea that to facilitate prediction at the level of a single observation (for example an individual) we need to permit more statistical freedom, relax assumptions and entertain exploratory approaches that allow computers to learn from often multilayered and multidimensional data (for example from the clinical, brain or genetic sources as seen in this issue). The power of this freedom to find new predictive patterns in multidimensional data is largely why machine learning has replaced traditional statistical and computer programming approaches in multiple corporate and scientific domains.[2]

Overfitting risk

A danger of more statistical freedom, however, is that it comes with an increased risk of finding results that are only accurate in a single sample and cannot be more widely applied in other contexts. This is known as ‘overfitting’ where idiosyncratic attributes of a sample (such as random noise) are modelled instead of identifying patterns that generalise to new cases and contexts. Thus, the third main point is that this overfitting risk is thought to be enhanced in machine-learning contexts and this needs to be kept in-mind at the current time. However, it is also important to recognise that the machine-learning field has popularised and extended statistical methods that test and optimise the ability of the algorithms to generalise to new cases, samples, sites, countries or continents.

At an initial level, most methods that assess generalisability rely on data resampling schemes that simulate the application of

https://doi.org/10.1192/bjp.2022.29 Published online by Cambridge University Press

algorithms to new data in order to obtain accuracy estimates;[1] for example, the commonest is to use cross-validation where a subsample of individuals is put aside, algorithms learn patterns in the remaining sample, the models are applied to the held-out subsample to determine their accuracy, and the process is repeated. In addition to these simulations, many articles in this special issue use forms of 'external validation' where the statistical algorithms are tested in completely new data-sets – for example from different studies or geographic locations. Such techniques are not unique to machine-learning contexts, but are more important in the field because of the risk of overfitting.

Representativeness of samples

A related fourth point to consider regards the representativeness of the sample that determines the scope of generalisability claims and potential sources of bias. Representativeness of the sample can first be assessed by considering clinical knowledge regarding the degree to which the results from the sample can support the conclusions of the study. For example, when making strong translational claims it is important for samples to be representative of real-world clinical environments rather than highly controlled scientific designs or methods.

Questions regarding bias can also be derived from clinical experience and relate to such factors as site, study, country, demographics or clinical differences. Assessing whether biases have been addressed is important for translational claims and can be tested with innovative resampling schemes (such as leave-group-out cross-validation[1]) in addition to the gold standard use of diverse external validation samples. Without assessments of bias there is the potential that the statistical models may not perform accurately based on such individual factors as race, ethnicity or gender – where machine-learning recommendations have been shown in other fields to be less accurate because the algorithms have predominantly learned decision rules from dominant majority groups. The integration of clinical knowledge into the design of machine learning tools is thus especially important in order to increase the representativeness of the samples and consider potential biases.

Real-world utility and implementation

The final point from a clinical perspective is to consider the real-world clinical utility and implementation of machine-learning tools, which are areas where the engagement with the wider research community is especially important. The usefulness of a statistical prediction is only as good as the ability for it to improve care to a degree that justifies the cost (and risk) of its implementation. Such questions can first be addressed by considering the potential of a tool to improve the status quo of clinical routines related to diagnoses, prognoses and treatment selection by assessing common quantitative metrics used in predictive contexts (such as accuracy, positive predictive value or area under the curve; see the Appendix). Increased confidence in the potential clinical utility can also be generated with additional assessments; for example, comparing machine-learning predictions with those made by clinicians in the same study, using net-benefit analyses to quantify the balance between the benefit (for example accurately predicting an illness) with potential harms (for example unnecessary testing), or by using decision curve and calibration analyses.[3]

Even if a tool is deemed to be sufficiently generalisable, the biases are known, it is better than existing clinical tools and has a clinical benefit, the final component of assessing whether the tool could be used is whether it could be practically implemented. Recent work in general medical fields has highlighted the unexpected difficulties with implementing highly promising tools into hospital settings,[4] which emphasises the need for ongoing input from clinical teams around how the most promising tools may actually work in real life. Towards these translational ends, some studies now provide web- or app-based platforms to test the capacity to deploy machine-learning algorithms (such as www.proniapredictors.eu). Additionally, providing algorithms is increasingly important for enhancing transparency through open science principles that are critical across the clinical sciences to facilitate understanding, replication and collaboration.

Conclusions

When combined, the five points to consider when reading a machine-learning paper were designed to provide important context for the papers in the following special issue and to engage a clinical audience. Moving forward, this clinical engagement will be critical for the field to progress and we hope that the special issue will encourage further dialogue towards a clinical future that includes the ability to tailor treatment approaches to individuals in real-time based on machine learning models. To further facilitate such a dialogue we have provided a glossary of terms (Appendix) that can be used as a reference and also a supplementary figure to aid understanding about analytic pipelines (see Supplementary Materials available at https://doi.org/10.1192/bjp.2022.29). We also invite interested readers to engage with other review papers in psychiatry.[1,5]

Dominic Dwyer, Department of Psychiatry and Psychotherapy, Ludwig Maximilian University, Germany; Orygen, Melbourne, Australia; and The Centre for Youth Mental Health, University of Melbourne, Australia; **Rajeev Krishnadas**, NHS Greater Glasgow and Clyde, University of Glasgow, UK

Correspondence: Rajeev Krishnadas. Email: Rajeev.Krishnadas@glasgow.ac.uk

First received 15 Nov 2021, final revision 11 Jan 2022, accepted 20 Jan 2022

Supplementary material

To view supplementary material for this article, please visit http://dx.doi.org/10.1192/bjp.2022.29

Author contribution

D.D. and R.K. wrote the article and provided the Table and Figure.

Funding

R.K.'s research is funded by Research and development, National Health Service - Greater Glasgow and Clyde, Chief scientist office Scotland and Medical research council, UK. D.D.'s research is funded by a National Alliance for Research on Schizophrenia & Depression Young Investigator Grant (no. 30196).

Declaration of interest

R.K. and D.D. do not have any conflicts of interest pertaining to this article.

https://doi.org/10.1192/bjp.2022.29 Published online by Cambridge University Press

Appendix

Glossary of terms

Terms

Accuracy: the proportion of correctly predicted cases in reference to all cases.

Algorithm: a sequence of statistical, mathematical or programmatic rules usually conducted by a computer to achieve a goal; for example an algorithm for multiplication or to predict disease.

Area under the curve (AUC): usually refers to the area under the receiver operating characteristic (ROC) curve. It is a value that measures the overall performance of a classifier within the range (0.5–1.0), where 0.5 represents the performance of a random classifier and the maximum value would correspond to a perfect classifier.

Cross-validation: an internal validation resampling technique used to empirically assess the accuracy and potential generalisability of statistical models, usually for a specific outcome.

Features: data (such as variables) that are used and modified to classify or predict an output.

Function: a mathematical relationship between two variables *x* and *y*. For example if the function that maps *y* to *x* is a 'square root function' $f(x) = \sqrt{x}$, then given $x = 16$; $y = \sqrt{x} = 4$. Mathematical functions used in supervised learning algorithms/optimisation are usually more complex.

Generalisability: algorithm performance on new data that can be assessed with internal validity (such as using cross-validation techniques) or external validity (such as validating the models on data from a different study, time period or geographic location). Also includes the assessment of model bias towards certain dominant groups (for example Western European groups).

Label/output: the predictive target used in supervised learning that is assigned to each case, such as diagnoses or prognostic outcomes.

Model: usually, the set of features and their parameters (weights) that maps features to outputs.

Negative predictive value: given a negative test, the probability of not actually having the disease/outcome.

Optimisation: a mathematical function or algorithmic technique used to find the highest performing parameters given a criterion (such as accuracy).

Parameter: the weight given to a feature in a model.

Positive predictive value: given a positive test, the probability of actually having the disease/outcome.

Receiving operating characteristic (ROC) curve: a graph used to evaluate the performance of classifiers. ROC plots show the sensitivity/specificity trade-off of a classifier for all possible thresholds.

Reinforcement learning: algorithms are used to learn by interacting with the environment using reward and penalties to perform a task.

Sensitivity: the proportion of affected cases with a positive test result in reference to all affected cases.

Specificity: the proportion of non-affected cases with a negative test result in reference to all non-affected cases.

Supervised learning: predicting an outcome using known target labels (for example a diagnosis or prognosis).

Testing: the application of trained algorithms without modification to held-out data that has not been used in the creation of the models.

Training: a statistical procedure that involves fitting a model to a data-set by modifying parameters (such as for prediction).

Unsupervised learning: the algorithm is not provided with any pre-assigned labels or scores and is used to find homogeneous subgroups of individual cases within a set of features.

References

1 Dwyer DB, Falkai P, Koutsouleris N. Machine learning approaches for clinical psychology and psychiatry. *Annu Rev Clin Psychol* 2018; **14**: 91–118.

2 Topol EJ. High-performance medicine: the convergence of human and artificial intelligence. *Nat Med* 2019; **25**: 44–56.

3 Leighton SP, Krishnadas R, Upthegrove R, Marwaha S, Steyerberg EW, Gkoutos GV, et al. Development and validation of a nonremission risk prediction model in first-episode psychosis: an analysis of 2 longitudinal studies. *Schizophr Bull Open* 2021; **2**: sgab041.

4 He J, Baxter SL, Xu J, Xu J, Zhou X, Zhang K. The practical implementation of artificial intelligence technologies in medicine. *Nat Med* 2019; **25**: 30–6.

5 Chekroud AM, Bondar J, Delgadillo J, Doherty G, Wasil A, Fokkema M, et al. The promise of machine learning in predicting treatment outcomes in psychiatry. *World Psychiatry* 2021; **20**: 154–70.

EXTRA CONTENT ONLINE

https://doi.org/10.1192/bjp.2022.29 Published online by Cambridge University Press