BJPsych The British Journal of Psychiatry (2025)
Page 1 of 5. doi: 10.1192/bjp.2024.258

Feature

# Precision psychiatry: thinking beyond simple prediction models – enhancing causal predictions

Rajeev Krishnadas, Samuel P. Leighton and Peter B. Jones

Making informed clinical decisions based on individualised outcome predictions is the cornerstone of precision psychiatry. Prediction models currently employed in psychiatry rely on algorithms that map a statistical relationship between clinical features (predictors/risk factors) and subsequent clinical outcomes. They rely on associations that overlook the underlying causal structures within the data, including the presence of latent variables, and the evolution of predictors and outcomes over time. As a result, predictions from sparse associative models from routinely collected data are rarely actionable at an individual level. To be actionable, prediction models should address these shortcomings. We provide a brief overview of a general framework for the rationale for implementing causal and actionable predictions using counterfactual explanations to advance predictive modelling studies, which has translational implications. We have included an extensive glossary of terminology used in this paper and the literature (Supplementary Box 1) and provide a concrete example to demonstrate this conceptually, and a reading list for those interested in this field (Supplementary Box 2).

**Keywords**

Precision medicine; big data; causal inference; diagnostic medicine; machine learning methods.

## Illustrative case vignette

Mr A, a 35-year-old individual, was recently diagnosed with a first episode of psychosis. He had a normal body mass index (BMI) despite a family history of diabetes. Dr B initially prescribed risperidone. However, within 12 weeks Mr A was found to be overweight, prompting Dr B to recommend physical activity and switch the antipsychotic to a weight-neutral one as per existing evidence. Mr A experienced rapid weight loss over the next few months, attributing this to the success of the new treatment strategy/intervention. At 6 months, Mr A was found to be emaciated and was diagnosed with ketoacidosis – an unforeseen consequence. His weight loss was a result of undiagnosed/untreated diabetes, rather than the intervention.

While this is not a typical scenario, the vignette underscores the limitation of applying group-level evidence-based guidelines at an individual level. The psychiatrist's interventions for weight gain were based on current available evidence.[1] Nevertheless, the outcome proved counter-intuitive. Similar scenarios are not uncommon in clinical practice. For example, despite the recognised effectiveness of antidepressants, they are not equally effective for everyone. In fact, only a third of the antidepressant-treated population will experience a genuine antidepressant response with the initial treatment.[2] Predicting individualised outcomes is therefore crucial for early and effective clinical decision-making. Considerable funding has been directed towards precision psychiatry methods that have the potential to tailor treatments to an individual's needs, thereby improving outcomes.[3–5]

What is precision psychiatry? Precision psychiatry departs from the conventional, 'one size fits all' approach to treatment decisions. In conventional evidence-based psychiatry, group-level/average treatment effects (ATEs) guide treatment decisions and everyone with a specific condition receives the same treatment. In contrast, precision psychiatry utilises individual patient characteristics to obtain individualised[6–8] treatment effects to predict and thereby maximise individualised treatment benefits. The implicit assumption of outcome prediction is that providing enough forewarning would allow the patient and clinician to make treatment decisions based on individual needs to effect change. This approach can hence help stratify individuals more accurately to existing treatments. The purpose and ultimate goal of predictions and hence the crux of precision psychiatry is to improve individual patient outcomes (decision-making).[9]

Most precision psychiatry approaches use some form of predictive modelling/machine learning techniques to predict individualised outcomes. They use statistical algorithms that learn some relationship (a function) between a predictor and an outcome, from large data-sets (training set). The learned relationship (model) is then evaluated by testing how well it does the prediction task on unseen data (test set).[3,10] This is like asking a child to learn to differentiate cats from dogs using pictures. One would initially show them some example pictures of cats and pictures of dogs. One might even teach them some basic rules to differentiate (supervised learning). Once one thinks the child has learned to differentiate them, one gives them a final test on a brand-new set of pictures. This final test shows how well they perform on examples they have never seen before – generalisation. The statistical algorithm's performance on unseen data (test/validation set) is evaluated using metrics such as 'classification or prediction accuracy' (or some variations).

While research in this field has grown significantly, the practical application of these models in psychiatric decision-making remains limited. This is in part because current statistical models, which map a statistical relationship between predictors and outcomes, often rely on methods that identify correlations (associations) rather than causal relations. Without understanding the underlying causal structure between predictors and outcomes, these models are not actionable and have limited clinical decision-making utility. For instance, consider an 'antipsychotic-induced weight gain' prediction model. Given certain risk factors, the model predicts an 80% probability of weight gain in an individual. Knowing this prediction probability alone does not guide treatment decisions. For example, should the clinician recommend healthy lifestyle, or

https://doi.org/10.1192/bjp.2024.258 Published online by Cambridge University Press

change the antipsychotic? Unfortunately, association models cannot provide true individualised treatment recommendations, as they fail to identify the causal structure between risk factors, and weight gain for the individual. Ultimately, even after getting the prediction probability, the clinician is forced to resort to group-level evidence to make treatment choices.

In the following sections, we will provide a brief overview of causal prediction models. As psychiatrists, we recognise that our expertise lies not in the algorithms or mathematics used in the predictions, but rather in our clinical domain knowledge. An intuitive understanding of causal inference principles and prediction algorithms, and their limitations, can significantly improve our ability to apply our domain knowledge to ensure that such models are actionable.

## Causal inferences are essential for actionable predictions

One of the main limitations of current psychiatric prediction models is their failure to capture causal relationships between predictors and outcomes. Shmueli[11] contests that the problems addressed by 'explanatory/causal' and 'prediction' models are essentially different. Briefly, given two variables, $x$ and $y$, explanatory/causal models aim to find the nature of the relationship between the two variables – why is $x$ related to $y$ (or when mechanisms are concerned – how). Meanwhile, prediction models aim to find the most accurate $y$ (outcome), given a particular value of $x$ (predictor). Prediction models focus on 'associations' (identifying statistical dependencies – is there a relationship between $x$ and $y$?) between variables. They are not modelled to find the 'nature of the relationship' between the variables and hence do not distinguish between causes (potentially modifiable risk factors) and effects. As a result, prediction models remain at the lowest rung of Pearl's causal explanation ladder – 'association' (Supplementary Box 1 available at https://doi.org/10.1192/bjp.2024.258). In addition, we know that association does not mean causation. From here on, we refer to such prediction models as 'association' models (see Supplementary Box 1).[12]

This, of course, leads to some fundamental problems. First, in medicine, actionable predictions are essential for translational purposes. While associative models may reveal a strong association between a risk factor and the disease, without establishing causation, it is unclear whether targeting that risk factor would effectively change outcomes at an individual level. As a concrete and rather hyperbolic example, a predictive algorithm may identify yellow stains on fingers as a strong predictor of lung cancer. However, prescribing a manicure is unlikely to prevent lung cancer because smoking – the 'confounder' – is the cause for both yellow fingers and lung cancer. In other words, to make informed treatment decisions, the predictions should be informed by the underlying causal structure within the data. Second, if a prediction results in a generic intervention recommendation, it misses the point of personalised medicine. For example, if an algorithm predicts an 80% probability of antipsychotic-induced weight gain, should the physician recommend physical activity without knowing if lack of physical activity is causally linked to the person's weight gain? What if the predicted probability is 20%? Considering that one in five people in the population have some cardio-metabolic illness, is it ethical to not recommend physical activity to anyone?

In medicine, randomised controlled trials (RCTs) are traditionally used to address causality by estimating the ATE, which represents the effect of a treatment across a population. By randomising treatment assignment, RCTs eliminate confounding, allowing for reliable causal inference at the population level. However, individual decisions based on RCTs do not guarantee treatment response, because RCTs do not provide insights into individualised treatment effects (ITEs) (for individual treatment effects, see Supplementary Box 1). The ITE reflects the difference in outcome for a specific individual if they received the treatment versus if they did not.[13] Precision medicine focuses on these ITEs, aiming to tailor treatments to individuals based on their unique characteristics. While the ATE is essentially the average of all ITEs, it does not capture the variability in individual responses, known as heterogeneous treatment effect (HTE). In homogeneous populations where individuals share similar characteristics, the ATE may serve as a good proxy for the ITE. However, in most real-world settings, populations are heterogeneous, meaning that treatment effects vary across individuals. This variation highlights the limitation of relying solely on the ATE to predict treatment response, as it does not reveal how much a particular individual will benefit from a given treatment. Returning to the opening clinical vignette, trials assume that Mr A will be Mr Average and that the ATE will suffice; in fact, he was Mr Atypical and assuming the ATE would be relevant was near fatal. Dr B needed an estimate of his ITE. External validation and generalisation tests are often touted as a panacea for all shortcomings of associative prediction models. They, however, do not mitigate the problem of HTEs. While poor generalisation may indicate HTEs, it does not solve the problem. Further investigation and model improvements are necessary to address HTEs. Generalisation procedures do not explain why the model fails to generalise in the presence of HTEs. Specialised approaches are needed to fully understand treatment effects heterogeneity. Lastly, generalisation is not a substitute for causal modelling as the association between yellow fingers and risk of cancer may generalise across all populations, but still does not imply causality.[14]

This then leads on to the fundamental problem of causal estimation.[15] Given that the ITE is the difference in outcome for an individual if they received the treatment compared to if they did not, estimating ITEs inherently involves estimating the 'counterfactual' outcome, in other words, a 'what if' scenario. We can never truly know what an individual's outcome would have been if they had received no treatment or the opposite treatment. This is because once a person receives a particular treatment, we are restricted to observing the outcome (factual) associated with that particular treatment. We do not have the option of giving the person the alternate treatment, and we do not know the potential outcome associated with the alternate treatment (counterfactual). Therefore, in practice, estimating ITEs often relies on strong assumptions and additional data beyond what conventional studies provide. In the absence of homogeneous populations, estimating ITEs often requires advanced statistical methods and additional data on individual characteristics to model the heterogeneity.[15] Recently, advanced statistical models have been used to estimate ITEs from ATEs by combining them within regression models, along with other relevant covariates/interactions. Other examples are causal forests, doubly robust estimation (which uses an outcome regression model and propensity score weighting as predictors) and targeted maximum likelihood estimation (TMLE).[8,15,16]

Where RCTs are not plausible, or do not represent real-world situations, causal estimates can also be estimated from observational studies. In observational studies, confounders are particularly problematic because treatments or exposures are not randomly assigned. In particular, unmeasured confounders – for example, in an observational study investigating antipsychotic-induced weight gain – genetic factors that influence drug metabolism or appetite regulation are not routinely measured in clinic. They also require several key assumptions that are crucial for valid causal inference, including exchangeability, positivity, consistency and no interference (see Supplementary Box 1). ATEs could then be estimated

https://doi.org/10.1192/bjp.2024.258 Published online by Cambridge University Press

using techniques such as propensity score matching, inverse probability weighting (IPW) and Mendelian randomisation methods (instrumental variables).[15] More recently, advanced counterfactual machine learning approaches such as generative adversarial networks (GANs) have provided a potential solution to estimating ITEs from observational studies. These approaches estimate causal effects by simulating 'what would have happened' under different conditions. This requires additional generation of counterfactual outcomes: the potential outcomes under the treatment patients did not receive using simulations.[8,15,16] By simulating 'what if' scenarios, they enable counterfactual estimation, helping validate ITEs. Simulation methods such as Monte Carlo simulation can robustly estimate ITEs where the true potential outcomes are unknown and provide a flexible and intuitive way to account for HTEs, allowing predictions based on their covariates. Target trial emulation (TTE) is another powerful technique for simulating the results of a hypothetical RCT based on observational data[15] (see Supplementary Box 1). All methods have their limitations, and one way to achieve a reasonable estimate of the true causal inference is by triangulating across the different methods.

Within a causal prediction framework, the parameters estimated from the above ITE estimation methods (a representation of the causal model) can then be used to predict outcomes in individuals, helping decision-making processes (see Supplementary Box 2 for a simple example). Such causal prediction models provide actionable insights, allowing us to answer critical questions such as, 'What predictors should we manipulate to bring about a desired change in this individual?'. Or 'What is the minimum change that should be made in a particular risk factor that would bring about the desired change in this individual?'. They enable personalised recommendations and a true precision medicine approach.[8] Unfortunately, this field is still in its infancy and causal machine learning research has primarily focused on evaluating methods through simulations on synthetic data-sets, ignoring the complexity of real-world disease dynamics. Application of innovative causal machine learning techniques in clinical contexts can be an important first step towards generating valuable insights.[16,17]

## Statistical models with few predictors that ignore domain knowledge can affect causal inference

Causal predictions require expert domain knowledge, with a sufficient number of variables that can account for the causal relationship. Association models with sparse predictors – derived from routinely collected data – avoid potential over-fitting in the context of small sample size and offer parsimony of explainability and implementation efficiency.[9,10] However, such sparse models may struggle to learn reliable associations between predictors and outcomes, resulting in unstable outcomes with minor distribution changes in unseen data (like a child trying to differentiate a cat from a dog just by looking at its eyes and ears).[12,13,18,19] Sparse models might learn by 'memorisation' of noise and irrelevant patterns and miss the underlying true relationship.[14,20] They overlook component heterogeneity, are more susceptible to the influence of confounders and preclude ITE estimation. In practical terms, inclusion of predictors should be guided by theoretical knowledge of the underlying causal structure. Dahabreh and Hernán[21] propose including a sufficient number of predictors for the expected outcome to avoid model misspecification. However, adding non-treatment modifier covariates that differ between two populations can lead to variance inflation.[22] Bias in variable selection can be minimised through automated confounder adjustment, causal discovery algorithms (including directed acyclic graphs) and regularisation. Ignoring domain knowledge for want of sparsity results in models that fail to account for the intricate relationships between predictors.[23] For instance, predicting a patient's cardiovascular risk based solely on their serum cholesterol, BMI and age, ignoring their complex interactions (mediation/confounding/moderation – causal effects), may result in inaccurate prediction (see Fig. 1). This prevents us from making any causal inferences, limiting actionability. Finally, causal models should account for the emergence of newly identified risk factors. For instance, the introduction of second-generation antipsychotics led to the emphasis on cardiometabolic side-effects, which has a direct effect on physical morbidity.

## Causal relationships between predictors and outcomes evolve temporally – neglecting this temporal evolution can lead to suboptimal predictions

Many prediction problems involve predictors and outcomes whose causal relationships change over time. The evolution of an illness over time can reflect its natural course, any emergent risk factors, the volatility and seasonality, as well as the effect of any interventions. To make accurate predictions, models must consider the temporal dependencies and dynamics within the data. Here, it is crucial to consider the impact of past epochs on the current presentation (different layer/unit of analysis) and future predictions. In addition, there is often what is considered a circular argument – where the outcome at one point in time may be the best predictor of the outcome at another point in time. For instance, premorbid function is often the best predictor of functioning a year after a psychotic episode. The simple aphorism that the past is the best predictor of future behaviour is often lost within the sophistication provided by algorithms (L. Palaniyappan, personal communication, 2023).

Current prediction models are often 'victims of their own success'. The more effective the model and interventions are at improving outcomes, the faster a model's performance will degrade. This is because interventions based on the model predictions disrupt the underlying association between the predictors and the outcome. For example, with QRISK version 3 (University of Nottingham and EMIS; https://www.qrisk.org/), a specific combination of predictors will be associated with a lower 10-year cardiovascular risk if the patient has been prescribed a statin. In addition, over time, population demographics, prevalence of disease and clinical practice and provisions change. Consequently, predictions that ignore dynamics of causal relations become outdated and inaccurate – they undergo 'concept drift' or 'artificial intelligence ageing'.[24] Another concern is 'calibration drift'. In any particular model, calibration – that is, the level of agreement between the predicted outcome and the actual observed outcome – deteriorates over time.[14,23–25] Since treatment decisions are recommended based on specific probability thresholds, calibration drift can lead to over- or under-treatment over time, requiring model updating and recalibration.[25,26] As well as being costly and cumbersome, model updating leads to sudden changes in risk, which does not actually reflect any actual change in the underlying distribution of outcomes. Further, a refitted model may identify different groups of patients (with different risk factors) as high risk, leading to inconsistent treatment decisions and potentially denying care to the original high-risk patients. This would lead to an iterative cycle of different groups of patients not receiving treatment. Dynamic causal prediction models could capture temporal dependencies and evolutions of predictors and outcomes, leading to more accurate predictions. Techniques such as marginal structural models (MSMs) with inverse probability of treatment weighting (IPTW) and G-computation handle time-varying confounders by adjusting for their

https://doi.org/10.1192/bjp.2024.258 Published online by Cambridge University Press

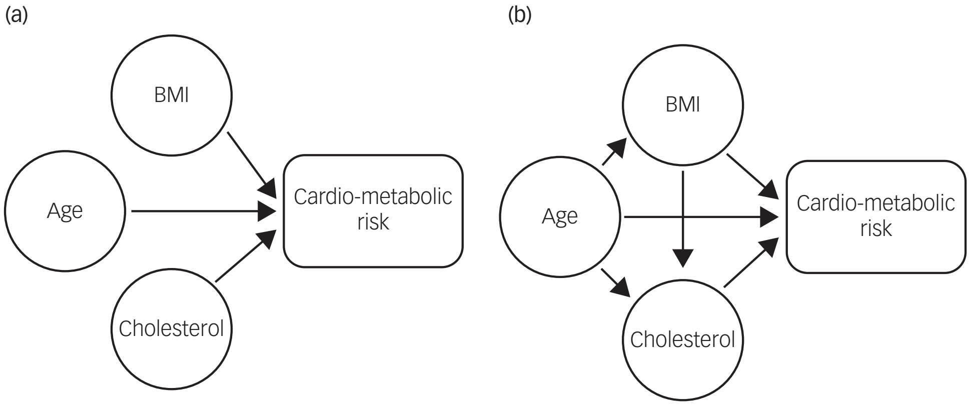

**Fig. 1** (a) Conventional model, which does not take into consideration causal interactions. (b) Potential causal structure between the variables. BMI, body mass index.

evolving impact on treatments and outcomes.[15] Temporal causal forests and dynamic Bayesian networks (DBNs) provide flexible approaches to estimating treatment effects by incorporating time-series dependencies.[16] Linkage of electronic health records across diverse sources enables mapping patient journeys longitudinally with repeated measures of care in larger populations than would be feasible in traditional studies.

## Capitalising on latent constructs that represent causal mechanisms

In psychiatry, causal relationships are rarely simple. and often involves latent constructs. Most current predictive models do not accommodate for the presence of potential latent constructs. Latent constructs are hidden variables (that may potentially be used as causal predictors) that cannot be overtly measured, but only inferred. Quantifying these constructs (and their relationships) calls for applying advanced theoretical, statistical and sophisticated feature engineering techniques. These include methods that utilise computational approaches that ultimately unveil new features that represent latent constructs. A recent example of such an effort is the Research Domain Criteria (RDoC) framework for mental illnesses that seeks to establish cognitive and behavioural domains that furnish reliable and valid predictors.[27] Techniques such as generative embedding have emerged within the cognitive neuroscience framework to address such issues. Such techniques bridge the gap between traditional machine learning and mechanistic understanding of mental illness.[28] Generative embedding uses generative models (see Supplementary Box 1) that mimic the brain's neuro-computational signature of the disorder within predictive models[a]. For example, Queirazza et al[30] employed a computational model that captures the cognitive and neural mechanisms of decision-making to predict response to cognitive–behavioural therapy (CBT) in depression. They found that the functional magnetic resonance imaging (fMRI) signals representing these mechanisms in the brain predicted an individual's response to CBT. Such techniques can help identify future personalised treatment options and bespoke treatment allocation. By analysing the extracted features and their response to simulated therapy, researchers could predict which intervention best addresses a patient's unique cognitive bias. In theory, generative embedding can enable the model to make more accurate actionable predictions by capturing mechanistically[1] relevant hidden variables that simpler models might overlook.

In conclusion, simple, associative models that fail to capture the causal relationship between predictors and outcomes may be futile at best and actively harmful at worse. We often impose model simplicity because of the challenge of collecting relevant predictors, risking unrealistic and unactionable solutions to the problem at hand (idealisation)[b]. While parsimony can be elegant, even the most complex model is a simpler representation of the actual relationship – akin to the most sophisticated map of a city that still lacks the intricate details of the actual landscape (abstraction). Variable selection using exhaustive domain knowledge involving clinicians is crucial in developing and improving causal prediction that helps actionable decisions. In addition to being actionable, causal models that are explicitly modelled to estimate and use ITEs ensure explainability. Data/decision transparency and data protection rules warrant that the decision processes are laid out for the patient. Black box models often avoid such explanations, and cannot help decision-making about an intervention at an individual level. As a result, there is now an emphasis on ethical and explainable machine learning within the medical field (see Lane and Broome[31] and Joyce et al[32] for a detailed exposition). The growing availability of clinical data from electronic healthcare records affords exciting new opportunities for pragmatic, cost-effective research to be conducted in an entirely naturalistic clinical setting. Incorporating causal and actionable predictive models using counterfactual explanations within such a framework can enable us to make informed decisions about interventions at an individual level, helping us truly implement evidence-based precision psychiatry (and medicine) as it was meant to be.

**Rajeev Krishnadas**, Department of Psychiatry, University of Cambridge, Cambridge, UK; **Samuel P. Leighton**, School of Health and Wellbeing, University of Glasgow, Glasgow, UK; **Peter B. Jones**, Department of Psychiatry, University of Cambridge, Cambridge, UK

**Correspondence:** Rajeev Krishnadas. Email: rk758@cam.ac.uk

First received 3 Apr 2024, revised 9 Oct 2024, accepted 31 Oct 2024

## Supplementary material

Supplementary material is available online at https://doi.org/10.1192/bjp.2024.258.

[a] See Chirimuuta[29] for a critique on computations as mechanisms.

[b] Like the physicist who developed a solution to increase milk productivity that only worked on spherical cows in a vacuum.

https://doi.org/10.1192/bjp.2024.258 Published online by Cambridge University Press

## Data availability

Data availability is not applicable to this article as no new data were created or analysed in this study.

## Acknowledgements

We would like to thank Professor Lena Palaniyappan and Dr Filippo Queirazza for their comments on an initial draft of the manuscript. We would like to thank Dr Alex Barradell, Assistant Psychologist, for his comments and helping with proof reading the manuscript.

## Author contributions

R.K. conceived the article. All authors contributed to writing the manuscript.

## Funding

S.P.L. was funded by a clinical academic PhD fellowship from the Chief Scientist Office, Scotland (CAF/19/04).

## Declaration of interest

S.P.L. does not have any conflict of interest pertaining to this manuscript. P.B.J. is the founder of Cambridge Adaptive Testing. R.K. is an editorial board member of the *British Journal of Psychiatry*. He did not take part in the review or decision-making process of this paper.

## Transparency declaration

This is an honest, accurate and transparent account of the study being reported. This is an opinion piece.

## References

1 Siskind D, Gallagher E, Winckel K, Hollingworth S, Kisely S, Firth J, et al. Does switching antipsychotics ameliorate weight gain in patients with severe mental illness? a systematic review and meta-analysis. *Schizophr Bull* 2021; **47**: 948–58.

2 Furukawa TA, Cipriani A, Atkinson LZ, Leucht S, Ogawa Y, Takeshima N, et al. Placebo response rates in antidepressant trials: a systematic review of published and unpublished double-blind randomised controlled studies. *Lancet Psychiatry* 2016; **3**: 1059–66.

3 Dwyer D, Krishnadas R. Five points to consider when reading a translational machine-learning paper. *Br J Psychiatry* 2022; **220**: 169–71.

4 Kambeitz-Ilankovic L, Koutsouleris N, Upthegrove R. The potential of precision psychiatry: what is in reach? *Br J Psychiatry* 2022; **220**: 175–8.

5 Meehan AJ, Lewis SJ, Fazel S, Fusar-Poli P, Steyerberg EW, Stahl D, et al. Clinical prediction models in psychiatry: a systematic review of two decades of progress and challenges. *Mol Psychiatry* 2022; **27**: 2700–8.

6 Rubin DB. Estimating causal effects of treatments in randomized and non-randomized studies. *J Educ Psychol* 1974; **66**: 688–701.

7 Imbens GW, Rubin DB. *Causal Inference for Statistics, Social, and Biomedical Sciences: An Introduction*. Cambridge University Press, 2015.

8 Bica I, Alaa AM, Lambert C, Van Der Schaar M. From real-world patient data to individualized treatment effects using machine learning: current and future methods to address underlying challenges. *Clin Pharma Ther* 2021; **109**: 87–100.

9 Coutts F, Koutsouleris N, McGuire P. Psychotic disorders as a framework for precision psychiatry. *Nat Rev Neurol* 2023; **19**(4): 221–34.

10 Kanwar MK, Kilic A, Mehra MR. Machine learning, artificial intelligence and mechanical circulatory support: a primer for clinicians. *J Heart Lung Transpl* 2021; **40**: 414–25.

11 Shmueli G. To explain or to predict? *Stat Sci* 2010; **25**: 289–310.

12 Pearl J, Glymour M, Jewell NP. *Causal Inference in Statistics: A Primer*. Wiley, 2016.

13 Molak A, Jaokar A. *Causal Inference and Discovery in Python: Unlock the Secrets of Modern Causal Machine Learning with DoWhy, EconML, PyTorch and More*. Packt Publishing Limited, 2023.

14 Collins GS, Dhiman P, Ma J, Schlussel MM, Archer L, Van Calster B, et al. Evaluation of clinical prediction models (part 1): from development to external validation. *Br Med J* 2024; **384**: e074819.

15 Hernan MA, Robins JM. *Causal Inference: What if* 1st ed. Taylor and Francis, 2023.

16 Feuerriegel S, Frauen D, Melnychuk V, Schweisthal J, Hess K, Curth A, et al. Causal machine learning for predicting treatment outcomes. *Nat Med* 2024; **30**: 958–68.

17 Lin L, Sperrin M, Jenkins DA, Martin GP, Peek N. A scoping review of causal methods enabling predictions under hypothetical interventions. *Diagn Progn Res* 2021; **5**: 3.

18 James G, Witten D, Hastie T, Tibshirani R, Taylor JE. *An introduction to Statistical Learning: With Applications in Python*. Springer, 2023.

19 Aliferis C, Simon G. Overfitting, underfitting and general model overconfidence and under-performance pitfalls and best practices in machine learning and AI. In *Artificial Intelligence and Machine Learning in Health Care and Medical Sciences* (eds GJ Simon, C Aliferis): 477–524. Springer, 2024.

20 Tänzer M, Ruder S, Rei M. Memorisation versus generalisation in pre-trained language models. *Arxiv* [Preprint] 2021. Available from: https://doi.org/10.48550/ARXIV.2105.00828.

21 Dahabreh IJ, Hernán MA. Extending inferences from a randomized trial to a target population. *Eur J Epidemiol* 2019; **34**: 719–22.

22 Colnet B, Josse J, Varoquaux G, Scornet E. Reweighting the RCT for generalization: finite sample error and variable selection. *Arxiv* [Preprint] 2024. Available from: http://arxiv.org/abs/2208.07614.

23 Deng C, Ji X, Rainey C, Zhang J, Lu W. Integrating machine learning with human knowledge. *iScience* 2020; **23**: 101656.

24 Vela D, Sharp A, Zhang R, Nguyen T, Hoang A, Pianykh OS. Temporal quality degradation in AI models. *Sci Rep* 2022; **12**: 11654.

25 Van Calster B, McLernon DJ, van Smeden M, Wynants L, Steyerberg EW. Calibration: the Achilles heel of predictive analytics. *BMC Med* 2019; **17**: 230.

26 Davis SE, Greevy RA, Lasko TA, Walsh CG, Matheny ME. Detection of calibration drift in clinical prediction models to inform model updating. *J Biomed Inf* 2020; **112**: 103611.

27 Insel T, Cuthbert B, Garvey M, Heinssen R, Pine DS, Quinn K, et al. Research domain criteria (RDoC): toward a new classification framework for research on mental disorders. *AJP* 2010; **167**: 748–51.

28 Illingworth B, Sandhu T, Smith E, Lage C, McCoy B, Lawson R. Transdiagnostic approaches to precision medicine: a computational psychiatry primer. *Authorea* [Preprint] 2022. Available from: https://doi.org/10.22541/au.166672691.16173006/v2.

29 Chirimuuta M. *The Brain Abstracted: Simplification in the History and Philosophy of Neuroscience*. MIT Press, 2024.

30 Queirazza F, Fouragnan E, Steele JD, Cavanagh J, Philiastides MG. Neural correlates of weighted reward prediction error during reinforcement learning classify response to cognitive behavioral therapy in depression. *Sci Adv* 2019; **5**: eaav4962.

31 Lane N, Broome M. Towards personalised predictive psychiatry in clinical practice: an ethical perspective. *Br J Psychiatry* 2022; **220**: 172–4.

32 Joyce DW, Kormilitzin A, Smith KA, Cipriani A. Explainable artificial intelligence for mental health through transparency and interpretability for understandability. *npj Digit Med* 2023; **6**: 6.

https://doi.org/10.1192/bjp.2024.258 Published online by Cambridge University Press