BRITISH JOURNAL OF HOSPITAL MEDICINE

EDITORIAL

How Outcome Prediction Could Aid Clinical Practice

Ashley Kieran Clift[1,*]

[1]Department of Surgery & Cancer, Imperial College London, London, UK
*Correspondence: akc13@imperial.ac.uk (Ashley Kieran Clift)

Abstract

Predictive algorithms have myriad potential clinical decision-making implications from prognostic counselling to improving clinical trial efficiency. Large observational (or "real world") cohorts are a common data source for the development and evaluation of such tools. There is significant optimism regarding the benefits and use cases for risk-based care, but there is a notable disparity between the volume of clinical prediction models published and implementation into healthcare systems that drive and realise patient benefit. Considering the perspective of a clinician or clinical researcher that may encounter clinical predictive algorithms in the near future as a user or developer, this editorial: (1) discusses the ways in which prediction models built using observational data could inform better clinical decisions; (2) summarises the main steps in producing a model with special focus on key appraisal factors; and (3) highlights recent work driving evolution in the ways that we should conceptualise, build and evaluate these tools.

Key words: prediction algorithms; prognosis; machine learning; validation

Submitted: 18 October 2024 Accepted: 18 November 2024

Introduction

Increasing enthusiasm in personalised care renders it incredibly likely that in the coming decade, clinicians will be asked to use a clinical prediction model (CPM), or work in systems where one is implemented.

Risk distributions within patient cohorts are not uniform. Therefore, the ability to reliably predict clinical outcomes could have significant implications (Efthimiou et al, 2024). CPMs leverage accumulated data and cumulative experience to personalise or stratify care based on an individual's characteristics. Uses of CPMs span patient counselling (e.g., mortality or recurrence risk over time), treatment selection (e.g., use of treatment "X" vs. "Y" in lower risk patients), clinical trial recruitment (e.g., identifying higher-risk individuals to adequately power trials), or improving trial efficiency (e.g., adjusting a trial for a prognostic score reduces necessary sample size). They are commonly developed using observational data from large registries or routinely collected electronic healthcare records (EHRs).

For model-informed care, CPMs must be implementable. Factors influencing this include: data quality (must reflect clinical reality), the target clinical decision to be influenced, and model trustworthiness to users and beneficiaries (Collins et al, 2024). Moreover, methodologies used to develop and validate models must be robust (Collins et al, 2024). Numerous systematic reviews have highlighted the

How to cite this article:
Clift AK. How Outcome Prediction Could Aid Clinical Practice. Br J Hosp Med. 2025. https://doi.org/10.12968/hmed.2024.0781

stark disparity between the number of possible CPMs developed in the literature and the number that are either used, or are even recommendable for use. For example, a recent review by Hueting and colleagues (2022) found that of over 900 models developed to predict outcomes in breast cancer, virtually all were at high risk of bias.

The Main Steps in Model Development and Evaluation

Simply put, a model: takes inputs (predictors), processes these in a risk "engine" (e.g., statistical model, artificial intelligence (AI) algorithm), and then produces an output for decision-making. The development and validation of CPMs encompass many facets. Efthimiou et al (2024) provide an excellent recent primer elsewhere that is strongly recommended. Here, we focus on three aspects for clinicians to consider when appraising a model they wish to use or are being asked to integrate within their practice.

Three Key Areas for Clinicians and Clinical Researchers to Consider

Data Quality

The mantra of "garbage in, garbage out" applies to CPM development. Routinely collected observational healthcare datasets offer advantages in terms of scale, coverage, and predictor-outcome relationships, especially when data linkage is employed (Riley et al, 2016). However, there may be data missingness, measurement error, potential misclassification, and the biases around clinical collection and coding.

Commonly used data points like body mass index or ethnicity may not be fully recorded. Missingness can be handled with different approaches, but the underlying mechanism influences strategy validity. Beyond development, consider how missingness is handled during model use. Will the model run automatically in the "backend" of an EHR system? Will it use predicted variable values? How are these predictions calculated? If predictor values are based on clinical codes, e.g., Systematized Nomenclature of Medicine (SNOMED) codes versus free text, what factors influence coding? Individual clinicians may have differing proclivities to code the same phenomenon with the same codes, or to formally code at all.

How do We Know that a Model "Works"?

This boils down to the model evaluation strategy and metrics. The commonest strategy of randomly splitting a dataset into two parts (one each for development and testing) reduces the amount of data available to fit the best model possible, leaving an even smaller set to understand in whom the model works (or not). Other approaches can be more informative (Collins et al, 2024; Steyerberg and Harrell, 2016).

Validation is not simply testing whether a model works on one "unseen" dataset drawn from the same population. Rather, it seeks to estimate how well a model might work when applied to future patients which will be different in time, may be different in location, and there may be temporal trends in predictor distributions and baseline risks. When evaluating a model, consider whether heterogeneity in performance has been assessed (e.g., within clinically relevant sub-groups) and whether the model demonstrates stability.

Many papers report a discrimination metric, e.g., a C-statistic. This evaluates how well a model distinguishes between individuals that had the event or not. This is important, however, models should also be well "calibrated", which pertains to the alignment of predicted risks and observed risks. If a model predicts a 10% risk, will 10% of all with that score experience the event? Is alignment demonstrated across the predicted risks spectrum? Miscalibration could lead to inappropriate decisions by under- or over-estimating risk.

Decision curve analysis is a valuable technique that visually compares the net benefits of different strategies (Chalkou et al, 2023). This can help understand whether or not a model will help make better clinical decisions than "treat everyone the same", and can also be used to compare different models beyond a set of metrics that consider individual aspects of performance.

Trustworthiness

Historic inequities within healthcare access and outcomes seep into real world data and can percolate into CPMs, risking perpetuating bias in future model-based care decisions.

Transparency describes understanding how the calculation was made from the inputs. Explainability describes being able to counsel an individual "why" those inputs lead to the output. This can be complex as it could reflect healthcare system bias, data quality, the effect of adjustment or inclusion of other variables, or complex associations between variables and the outcome. An adjusted coefficient term in a model or a shapley additive explanations (SHAP) value cannot be interpreted as the causal effect except in very specific circumstances (Keogh and Van Geloven, 2024).

Trust is more complex and influenced by transparency and explainability. Consider the following scenario: one patient has been assigned a lower risk score than another. They are identical in all their predictors other than ethnicity, and the risk score informs treatment decisions (Clift et al, 2023). You could be transparent in showing the equations' workings, but could you explain why their ethnicity dictates their risk and renders them less eligible (Obermeyer et al, 2019)? In some scenarios, ethnicity can recognise increased risk and improve equity (Vyas et al, 2020), but others exist where including ethnicity to apportion risk is deemed unacceptable (Citizens' Jury on QCovid, 2022).

Reconsidering How We Model

Validation is not a Rubber Stamp

A single validation study, no matter how robust, won't hold true indefinitely. Models' target populations change over time. Predictor distributions can drift, predictor-outcome associations evolve, measurement approaches change, and new treatments or diagnostics may be introduced. It is unlikely that the performance from the model's original or follow-up validation will hold steadfast in all settings in the future. Exemplified in the coronavirus disease 2019 (COVID-19) pandemic, baseline risks may evolve rapidly and clinical scenarios can shift (e.g., vaccination and novel variants). Van Calster and colleagues (2023) recently boldly stated that "there is no such thing as a prediction model", and there is a push towards having dynamic model systems, where deployed models are monitored and updated as needed. This could include regular re-training. The concept of "machine learning operations (MLOps)" is well established in industry data science, and we should expect healthcare systems to follow suit.

Treatment Drop-in and Causal Prediction

Models predicting long-term outcomes to inform treatment selection typically assume that their outputs correspond to risks in the absence of treatment. However, "treatment drop-in" occurs—people in the training data receive treatment after the prediction is made, but before they have the outcome (Sperrin et al, 2021). This affects prediction validity and interpretation. At baseline, one cannot use future information including treatments that are used later. Further, some models that include terms for medications associated with risk might be used to "simulate" the effect of changing medication status by observing the change in risk score, but this does not represent the true causal effect. Fascinating recent work has focussed on developing methods to integrate causal inference into predictive modelling, thereby understanding the ways we can build and test models that provide reliable predictions under different intervention scenarios—see (Keogh and Van Geloven, 2024; Lin et al, 2024; Sperrin et al, 2021) for excellent summaries.

Risks of Harmful Self-Fulfilling Prophecies

Even well-performing models can be harmful and, in some cases, can cause "self-fulfilling prophecies", particularly where disease aggressiveness correlates with lower intensity of treatment or there is a desire not to "over-treat" (van Amsterdam et al, 2024). Assume a model is developed and used for treatment decisions, which in turn changes outcomes. If a sub-population in the training data experienced poorer care, the "accurate" CPM could output higher risk predictions for this group compared to the "average" individual. The self-fulfilling prophecy here is that population group "A" has historically poor outcomes, the model used to apportion care predicts a worse outcome based on biased historical data, causing reduced treatment access. Model discrimination and calibration don't capture the benefits and harms of treatment changes (van Amsterdam et al, 2024). One should expect new methodological guidance from this burgeoning field in the near future.

Conclusion

Large, observational datasets offer manifold opportunities to develop predictive tools but come with risks. Clinicians are increasingly likely to encounter CPMs in their day-to-day practice. Trust in these models is crucial for users and beneficiaries and will drive or hinder implementation. The science of building and testing CPMs is evolving in real time.

Key Points

- There is a glut of CPMs in the literature and increasing interest regarding predictive analytics, which is compounded by excitement (or "hype") and funding into the field.
- Major obstacles to actionability and implementation of these tools include rife poor methodology, the poor articulation of the exact clinical decisions models seek to inform, and concerns around bias.
- These issues feature strongly in the zeitgeist of increasing attention on artificial intelligence in healthcare, but aren't exclusive to flexible modern methods.
- Complexity is emerging from new thinking around how to accurately model risks under different interventions and combining causal principles with traditional modelling—this should be embraced positively to drive innovation in this field.
- Care providers and receivers (target model users) should be empowered to understand, appraise and critique prediction systems that are developed to aid them.

Availability of Data and Materials

Not applicable.

Author Contributions

AKC was the sole author and was responsible for the design of the work, drafting and revision of content, and approval of the final version to be published. AKC has participated sufficiently in the work and agreed to be accountable for all aspects of the work.

Ethics Approval and Consent to Participate

Not applicable.

Acknowledgement

This editorial is based on an invited presentation given at the European Society of Medical Oncology (ESMO), Barcelona, September 2024.

Funding

This research received no external funding.

Conflict of Interest

The author declares no conflict of interest.

References

Chalkou K, Vickers AJ, Pellegrini F, Manca A, Salanti G. Decision Curve Analysis for Personalized Treatment Choice between Multiple Options. Medical Decision Making. 2023; 43: 337–349. https://doi.org/10.1177/0272023221143058

Citizens' Jury on QCovid: Report on the jury's conclusions and key findings. 2022. Available at: https://www.gov.scot/publications/citizens-jury-qcovid-report-jurys-conclusions-key-findings/ (Accessed: 18 October 2024).

Clift AK, Collins GS, Lord S, Petrou S, Dodwell D, Brady M, et al. Predicting 10-year breast cancer mortality risk in the general female population in England: a model development and validation study. The Lancet. Digital Health. 2023; 5: e571–e581. https://doi.org/10.1016/S2589-7500(23)00113-9

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. BMJ (Clinical Research Ed.). 2024; 384: e074819. https://doi.org/10.1136/bmj-2023-074819

Efthimiou O, Seo M, Chalkou K, Debray T, Egger M, Salanti G. Developing clinical prediction models: a step-by-step guide. BMJ (Clinical Research Ed.). 2024; 386: e078276. https://doi.org/10.1136/bmj-2023-078276

Hueting TA, van Maaren MC, Hendriks MP, Koffijberg H, Siesling S. The majority of 922 prediction models supporting breast cancer decision-making are at high risk of bias. Journal of Clinical Epidemiology. 2022; 152: 238–247. https://doi.org/10.1016/j.jclinepi.2022.10.016

Keogh RH, Van Geloven N. Prediction Under Interventions: Evaluation of Counterfactual Performance Using Longitudinal Observational Data. Epidemiology. 2024; 35: 329–339. https://doi.org/10.1097/EDE.0000000000001713

Lin L, Poppe K, Wood A, Martin GP, Peek N, Sperrin M. Making predictions under interventions: a case study from the PREDICT-CVD cohort in New Zealand primary care. Frontiers in Epidemiology. 2024; 4: 1326306. https://doi.org/10.3389/fepid.2024.1326306

Obermeyer Z, Powers B, Vogeli C, Mullainathan S. Dissecting racial bias in an algorithm used to manage the health of populations. Science. 2019; 366: 447–453. https://doi.org/10.1126/science.aax2342

Riley RD, Ensor J, Snell KIE, Debray TPA, Altman DG, Moons KGM, et al. External validation of clinical prediction models using big datasets from e-health records or IPD meta-analysis: opportunities and challenges. BMJ (Clinical Research Ed.). 2016; 353: i3140. https://doi.org/10.1136/bmj.i3140

Sperrin M, Diaz-Ordaz K, Pajouheshnia R. Invited Commentary: Treatment Drop-in-Making the Case for Causal Prediction. American Journal of Epidemiology. 2021; 190: 2015–2018. https://doi.org/10.1093/aje/kwab030

Steyerberg EW, Harrell FE, Jr. Prediction models need appropriate internal, internal-external, and external validation. Journal of Clinical Epidemiology. 2016; 69: 245–247. https://doi.org/10.1016/j.jclinepi.2015.04.005

van Amsterdam WAC, van Geloven N, Krijthe JH, Ranganath R, Ciná G. When accurate prediction models yield harmful self-fulfilling prophecies. arXiv. 2024. https://doi.org/10.48550/arXiv.2312.01210

Van Calster B, Steyerberg EW, Wynants L, van Smeden M. There is no such thing as a validated prediction model. BMC Medicine. 2023; 21: 70. https://doi.org/10.1186/s12916-023-02779-w

Vyas DA, Eisenstein LG, Jones DS. Hidden in Plain Sight - Reconsidering the Use of Race Correction in Clinical Algorithms. The New England Journal of Medicine. 2020; 383: 874–882. https://doi.org/10.1056/NEJMms2004740