It's about time: why we need to consider temporal drift when developing and implementing clinical prediction models

Yanakan Logeswaran[1,2], Dominic Oliver[3,4,5*]

1. Early Psychosis: Interventions and Clinical-Detection (EPIC) Lab, Department of Psychosis Studies, Institute of Psychiatry, Psychology & Neuroscience, King's College London, London, SE5 8AF, UK;
2. Department of Biostatistics and Health Informatics, Institute of Psychiatry, Psychology & Neuroscience, King's College London, London, SE5 8AF, UK;
3. Department of Psychiatry, University of Oxford, Oxford, OX3 7JX, UK;
4. NIHR Oxford Health Biomedical Research Centre, Oxford, OX3 7JX, UK;
5. OPEN Early Detection Service, Oxford Health NHS Foundation Trust, Oxford, UK

Corresponding author:
Dr Dominic Oliver (dominic.oliver@psych.ox.ac.uk)
POWIC Building, Department of Psychiatry, University of Oxford, Warneford Lane, Oxford OX3 7JX, UK.

Word count: 1294/1500
References: 10/10

What is temporal drift?

Temporal drift refers to changes in patient and contextual characteristics over time, leading to differences between the training dataset used for the development of a clinical prediction model (CPM) and future validation and implementation datasets (1). This drift can stem from changes in patient populations, disease prevalence, clinical care (e.g., referral pathways, treatment policies) and information systems amongst other factors, and can lead to deterioration in the accuracy of CPMs and their harmful implementation if not accounted for (2).

A common measure of temporal drift is calibration drift. Calibration measures the agreement between observed and predicted risk (e.g., in 100 people whose predicted risk is 20%, 20 people should go on to develop the outcome of interest if the model is perfectly calibrated). In calibration drift, due to changes in patient and contextual characteristics (as above), observed and predicted event rates diverge over time, leading to inaccurate risk estimation and potentially poor clinical decisions. For example, a high-risk patient may have their risk underestimated by a mis-calibrated CPM and therefore not receive relevant beneficial treatment as a result. Similarly, a low-risk patient may receive over-estimated risk predictions and thus receive unnecessary and potentially harmful treatment. Consequently, it is evident that the responsible implementation of CPMs requires the assessment of such temporal drift and the use of model updating methods to overcome their impact on model performance.

How can we address temporal drift?

In this issue of *Biological Psychiatry: Cognitive Neuroscience and Neuroimaging*, Hartmann et al. (3) assessed the extent of temporal drift in a CPM predicting psychosis onset in individuals at ultra high-risk of psychosis over 23 years, and examined the performance of several model updating methods to address this temporal drift. They developed the CPM on a training dataset from 1995-2014, and assessed both static (no updating; yearly recalibration; continual refitting) and dynamic (Bayesian time-variant) updating methods over a validation dataset from 2016-2018. Static methods are typically reactive, employed at discrete, arbitrary time points (e.g., annually) whereas dynamic updating methods allow CPMs to continuously adapt as data on new individuals becomes available.

They found that, without updating, the CPM exhibited substantial miscalibration over the validation period, primarily with risk underestimation across the entire sample and risk overestimation in high-risk individuals, as well as increasingly poorer discrimination as years progressed further from the training data. The dynamically updated model demonstrated the best discrimination and was the only model to maintain clinical net benefit over the validation period.

These results highlight the importance in assessing temporal drift and that dynamic updating may be a promising approach to mitigate this drift. Considerations do need to be made regarding the more complex digital infrastructure required to implement this method including computational power and a continuous data stream. It is also worth noting that the 95% confidence intervals were relatively wide, demonstrating no statistical superiority of any method, which is likely due to the relatively small sample sizes within each year. Moreover, none of the updating methods were able to maintain the original discrimination performance in later years, which highlights the need for comparing multiple updating methods when developing a CPM and ongoing model surveillance (in addition to model updating) for the safe and optimal use of CPMs post-implementation.

How often should we update?

Regarding the frequency of model updating, a balance needs to be struck between rapid responses to temporal drift and pragmatic clinical considerations. As CPMs often employ risk thresholds for decision-making such as treatment stratification, with highly frequent updating, an individual patient’s predicted risk may abruptly cross a risk threshold (potentially back and forth) when a model updates. This would change the interpretation of the model’s output at each update and treatment provision or mean that two identical patients could receive different treatments if assessed a few weeks apart. This uncertainty could also potentially reduce confidence in eventual treatment decisions.

Drift detection systems (4) may be able to partially overcome this issue; these can dynamically monitor the extent of temporal drift and indicate that model updating is required on detecting significant drift (i.e., exceeding a pre-determined threshold).

This data-driven approach would lead to less frequent model updates and decrease the likelihood of individual patients crossing treatment decision boundaries, therefore reducing the complexity of implementation.

Temporal drift and algorithmic fairness

Whilst the impact of temporal drift on model accuracy has been well-documented, it is important to look at measures beyond this. In particular, algorithmic fairness has become increasingly important in clinical prediction. Algorithmic fairness pertains to the extent to which a CPM performs similarly across demographic subgroups (e.g., ethnicity, gender) and therefore does not systematically disadvantage any subgroup, which otherwise would introduce or maintain existing health disparities.

Temporal drift is likely to differ across subgroups, and can lead to *fairness drift* whereby prediction models that were once fair exhibit substantial reduction in fairness in future implementation (5). Further, a recent study (6) has shown that updating methods do not necessarily overcome temporal fairness drift, and instead can both improve and exacerbate fairness disparities between minority and majority groups.

Altogether, this highlights the need for continual monitoring and assessment of algorithmic fairness to ensure that model updating methods are not sustaining population-level performance at the cost of performance within subgroups of known health inequities, thereby exacerbating disparities.

The impact of model implementation

Another challenge for accurate model updating stems from successful model implementation. Following implementation, a model will inform treatment decisions, changing the relationship between predictors (and the model's predictions) and the outcome, with this disturbance being more pronounced for more effective interventions (7). For example, if an effective preventive treatment is offered to patients considered high risk by a model, the risk within this group will decrease and this will not be accounted for by the model. Consequently, successful model implementation means that model performance as typically captured by metrics such as discrimination and calibration may appear to degrade.

Whilst model updating methods may mitigate such degradation and appear to improve model performance, their use without careful consideration of the aforementioned impact of implementation may lead to harmful predictions and decisions (8). For example, a CPM for psychosis risk may be updated on the latest patient data which indicates a weaker relationship between high-risk individuals and psychosis onset – but this artificial relationship has only resulted from the CPM successfully identifying people at the highest risk and their clinical team subsequently providing them with effective preventive treatment. Such updating may lead to risk underestimation and undertreatment of high-risk individuals. Therefore, if an updating method appears to be accurate using data collected pre-implementation (as in Hartmann et al. (3)), the impact of applying the updating strategy post-implementation needs to be specifically considered.

Evidently, maintaining CPMs in light of implementation-driven temporal drifts is not a trivial task. Alongside model updating, the use of counterfactual causal frameworks may provide a solution. These frameworks can essentially use data to model two divergent scenarios for a given individual. For example, in the real-world, someone was predicted to be at high risk of developing psychosis and provided with preventive treatment. Using data from people with similar characteristics in the dataset, a counterfactual can be created – what likely would have happened if they had not been given treatment. Through this, we may be able to evaluate the impact of risk predictions under different treatment conditions (e.g., pre- and post-implementation) and hence inform decision making, as well as model updating (9,10). However, caution is warranted given the difficulties in estimating causal effects from observational data.

Clinical prediction models are ultimately developed with aspirations of eventual implementation and long-term clinical impact. In order to fulfil these goals, we need to consider temporal drift and suitable updating methods both prior to and following implementation, from the very start of the model development process.

Acknowledgements

YL is supported by the UK Medical Research Council (MR/W006820/1) and King's College London member of the MRC Doctoral Training Partnership in Biomedical Sciences.

Financial disclosures

No potential conflicts of interest or financial disclosures to declare.

References

1. Quiñonero-Candela J, Sugiyama M, Schwaighofer A, Lawrence N (2009): *Dataset Shift in Machine Learning*. Cambridge, MA: MIT Press.
2. Subbaswamy A, Saria S (2020): From development to deployment: dataset shift, causality, and shift-stable models in health AI. *Biostatistics* 21: 345–352.
3. Hartmann S, Dwyer D, Scott I, Wannan CMJ, Nguyen J, Lin A, *et al.* (2025): Dynamic updating of psychosis prediction models in individuals at ultra high-risk of psychosis. *Biological Psychiatry: Cognitive Neuroscience and Neuroimaging*. https://doi.org/10.1016/j.bpsc.2025.03.006
4. Davis SE, Greevy RA, Lasko TA, Walsh CG, Matheny ME (2020): Detection of calibration drift in clinical prediction models to inform model updating. *Journal of Biomedical Informatics* 112: 103611.
5. Deho OB, Bewong M, Kwashie S, Li J, Liu J, Liu L, Joksimovic S (2025): Is it still fair? A comparative evaluation of fairness algorithms through the lens of covariate drift. *Mach Learn* 114: 8.
6. Davis SE, Dorn C, Park DJ, Matheny ME (2025): Emerging algorithmic bias: fairness drift as the next dimension of model maintenance and sustainability. *Journal of the American Medical Informatics Association* ocaf039.

7. Lenert MC, Matheny ME, Walsh CG (2019): Prognostic models will be victims of their own success, unless…. *Journal of the American Medical Informatics Association* 26: 1645–1650.

8. van Amsterdam WAC, van Geloven N, Krijthe JH, Ranganath R, Cinà G (2025): When accurate prediction models yield harmful self-fulfilling prophecies. *Patterns* 6: 101229.

9. Xu Z, Arnold M, Stevens D, Kaptoge S, Pennells L, Sweeting MJ, *et al.* (2021): Prediction of Cardiovascular Disease Risk Accounting for Future Initiation of Statin Treatment. *American Journal of Epidemiology* 190: 2000–2014.

10. Sperrin M, Jenkins D, Martin GP, Peek N (2019): Explicit causal reasoning is needed to prevent prognostic models being victims of their own success. *Journal of the American Medical Informatics Association* 26: 1675–1676.