DYNAMIC MODELLING METHODS FOR CLINICAL PREDICTION MODEL UPDATING AND MONITORING: IDENTIFICATION AND COMPARISON OF DYNAMIC MODELS FOR HEALTHCARE USE

A thesis submitted to The University of Manchester for the degree of

Doctor of Philosophy

In the Faculty of Biology, Medicine and Health

2022

David A. Jenkins

School of Health Sciences
Division of Informatics, Imaging and Data Science

Blank page

Table of contents

Final word count (including footnotes, endnotes, references): 37616

List of tables

List of figures

Abstract

There is growing interest in the use of clinical prediction models (CPMs) to aid decision making across healthcare. CPMs provide risk estimates for the presence of disease, or future outcomes, given current information about a patient. The pipeline of getting a CPM into clinical practice involves, (i) model development where a dataset is used to estimate the model parameters, (ii) model validation, where the predictive performance of the model is evaluated, (iii) impact assessment, where the clinical impact of the CPM is evaluated, and then finally (iv) model implementations into practice. It is commonly the case that once a model has been implemented, the model coefficients/parameters remain fixed, or at best are updated at arbitrary time points. However, healthcare and patient populations experience changes in terms of processes and case-mix, respectively, which means the covariate-outcome associations of the CPM also need to change, which is not reflected in most CPMs to-date. This results in the accuracy of the CPMs diminishing over time. This is known as calibration drift and is one of the major pitfalls of CPMs to date. Dynamic prediction models are a possible solution as the model parameters are not fixed and they attempt to acknowledge/model the temporal nature of the data.

This thesis explores the challenges of CPMs in the presence of calibration drift. The aims of the thesis are to (a) provide a comprehensive understanding of the methodology and challenges with dynamic modelling, (b) compare the predictive performance of the different models, and (c) to develop a method to address the problem of arbitrary updating.

Chapter 2 identifies existing methods used for dynamic prediction modelling through a review of the literature and highlights the current methodological challenges in dynamic prediction modelling. Chapter 3 discusses potential solutions to overcome the challenges described in chapter 2, leading to the suggestion of dynamic prediction systems, a way to continuously update and monitor a model over time. Following on from these chapters, chapters 4 and 5 compare the methods identified in chapter 2 through a simulation study and real-world data examples in cardiovascular disease. Despite the issues identified in chapters 2 and 3, chapters 4 and 5 found dynamic models perform as well as or better than non-dynamic models, which are currently the norm in the field. Building on this, chapter 6 develops a solution to one of the major challenges in predictive modelling, continuous monitoring and feedback, and illustrates the novel approach through simulation.

Generally, this thesis has the potential to improve performance and monitoring of prediction models, especially in presence of performance drift, by moving away from the current CPM framework and methods towards the proposed dynamic prediction systems. Practically, the thesis has used traditional and novel methodology to further the field of CPM development and validation.

Declaration

No portion of the work referred to in the thesis has been submitted in support of an application for another degree or qualification of this or any other university or other institute of learning.

Copyright Statement

Acknowledgments

Firstly, I express great appreciation to my supervisors Professor Niels Peek, Dr Glen Martin, Dr Matthew Sperrin and Dr Benjamin Brown, for their advice and encouragement throughout. Their guidance has been influential in my PhD and as a researcher and medical statistician. I could not have wished for more supportive, knowledgeable, and approachable PhD supervisors. Particular thanks for giving me insight in the sphere of prognostic research with an enthusiasm I only wish to match. I look forward to many future collaborations.

My personal thanks are extended to the Connected Health Cities Northern Health Science Alliance project and the NIHR Greater Manchester Safety and Translational Research Centre for funding my PhD and giving me the opportunity to carry out this research.

I would also like to thank all of the researchers, staff and students within the Centre for Health informatics who I have had the pleasure to work alongside. Particular thanks to Camilla Sammut-Powell and Rose Sisk for the discussions in the office, help with coding, answering my questions and listening to me waffle.

Last but certainly not least, I would like to offer special thanks to my family and friends for their wholehearted love and support. Particular thanks to my mum and dad, Sharon and Anthony, who always believed in me and have showed me strength, resilience and work ethic that I wish to emulate. You have always believed in me and given me the belief and pushes I needed that have led me to where I am today. Your love and support has been extremely precious to me (as always) and has unquestionably contributed to the completion of my PhD. To my late grandma who was always so proud and supportive, even though she did not understand what I was doing but listened anyway. Special thanks to my wife Hayley who has always supported and been there for me. You were with me when I started this journey but so much has happened and we are now married and have a wonderful son. I am so glad you have been right next to me on this roller coaster. From listening to me talk about my work, looking after our son and being such a great mother and through the highs and lows you have been amazing, so thank you. That leads me to my son, Samuel. You do not know but you always cheer me up and lighten up my day. You have bought me and your mum so much joy and have given me the motivation to complete my PhD. It was far from easy with all of the sleepless nights but you have helped me in so many ways and it has been a blessing to have you around when I have been working from home due to the pandemic. Thank you and may you never stop learning.

Many thanks to you all

David

To my Grandma, in loving memory

And to my son

"However difficult life may seem, there is always something you can do and succeed at."

-Professor Stephen Hawking

About the Author

Candidate degrees

2013 – 2014 MSc Medical Statistics, University of Leicester

2010 – 2013 BSc Mathematics, Keele University

Research Interests

David's research focuses on using real world health data to improve healthcare, with keen interests in both applied work and statistical methodology. His primary statistical research interest is predictive modelling. Specifically, developing methodology around validating prediction models and developing clinical prediction models which are able to update and evolve over time.

Publications

Published peer-reviewed papers arising directly from this PhD include:

Jenkins, D.A., Sperrin, M., Martin, G.P. and Peek, N., Dynamic models to predict health outcomes: current status and methodological challenges. *Diagnostic and prognostic research*. (2018) DOI: 10.1186/s41512-018-0045-2

Jenkins, D.A., Martin, G.P., Sperrin, M. et al. Continual updating and monitoring of clinical prediction models: time for dynamic prediction systems? *Diagnostic and prognostic research*. (2021) DOI: 10.1186/s41512-020-00090-3

Peer-reviewed papers published during my PhD but not directly arising from my PhD:

Jenkins, D.A., Hussein, H., Martina, R. et al. Methods for the inclusion of real-world evidence in network meta-analysis. *BMC Medical Research Methodology*. (2021). DOI: 10.1186/s12874-021-01399-3

Jenkins, D.A., Wade, K.H., Carslake, D., Bowden, J., Sattar, N., Loos, R.J., Timpson, N.J., Sperrin, M. and Rutter, M.K., 2021. Estimating the causal effect of BMI on mortality risk in people with heart disease, diabetes and cancer using Mendelian randomization. *International Journal of Cardiology*. DOI: 10.1016/j.ijcard.2021.02.027

Lin, L., Sperrin, M., **Jenkins, D.A.**, Martin, G.P. and Peek, N. A scoping review of causal methods enabling predictions under hypothetical interventions. *Diagnostic and prognostic research*. (2021). DOI: 10.1186/s41512-021-00092-9

Williams, R., **Jenkins, D. A.**, Ashcroft, D. M., Brown, B., Campbell, S., Carr, M. J., … Peek, N. Diagnosis of physical and mental health conditions in primary care during the COVID-19 pandemic : a retrospective cohort study. *The Lancet Public Health.* (2020). DOI: 10.1016/S2468-2667(20)30201-2

Martin, G. P., **Jenkins, D. A.**, Bull, L., Sisk, R., Lin, L., Hulme, W., … Peek, N. Toward a framework for the design, implementation, and reporting of methodology scoping reviews. Journal of Clinical Epidemiology. (2020). DOI: 10.1016/j.jclinepi.2020.07.014

Mistry, C., Palin, V., Li, Y., Martin, G.P., **Jenkins, D.**, Welfare, W., Ashcroft, D.M. and van Staa, T. Development and validation of a multivariable prediction model for infection-related complications in patients with common infections in UK primary care and the extent of risk-based prescribing of antibiotics. *BMC medicine.* (2020). DOI: 10.1186/s12916-020-01581-2

Jenkins, D.A., Mohamed, S., Taylor, J.K., Peek, N. and van der Veer, S.N., 2019. Potential prognostic factors for delayed healing of common, non-traumatic skin ulcers: A scoping review. *International wound journal.* (2019). DOI: 10.1111/iwj.13100

Chapter 1 General Introduction

The data revolution has led to the increase in collection and use of data throughout healthcare. Algorithms and statistical models, specifically clinical prediction models, are increasingly being used, but they degrade over time and methods to utilise the continual flow of data in healthcare are lacking.

This thesis will identify and compare statistical modelling methods, known as dynamic prediction models, for model updating and monitoring to overcome the problem of model degradation. This chapter provides the background to the research area by first describing health data that are often used for clinical prediction modelling in healthcare before outlining the concept of the learning health system. Following this, an introduction to clinical prediction models is provided which includes the current approach to develop and validate models, before detailing one of the major pitfalls with clinical prediction models, calibration drift. Current practice to overcome this issue is then discussed and the need for further improvement is highlighted. Next, dynamic prediction models are introduced as a potential solution before discussing the current challenges and further research needed in this area. Finally, the chapter ends with the research aims and objectives.

1.1 Electronic health records/ Health data

Electronic health records (EHRs) contain the medical and treatment histories of patients, including diagnosis, medications, treatment plans, laboratory tests and referrals. Such data are stored digitally and often in real-time. These data enable information to be available instantly and securely to those with access. The records consist of information from GP consultations and hospital inpatient and outpatient visits. They are used primarily to share information across the healthcare system and include information from numerous healthcare providers and organisations. As a secondary use, EHRs are increasingly used for research purposes given that the digitisation of routinely collected data provides data on a large number of individuals often across wide geographical areas. The widespread adoption of EHRs has provided a wealth of information for health research and has enabled opportunities to enhance patient care, embed performance measures into clinical practice, identify and recruit patients in research and improve productivity and efficiency of the healthcare system. Increases in digitisation, data collection and computing power have led to big data analytics and the data revolution[1–3].

1.2 Learning health system

Inspired by the data revolution, such as EHRs, health informaticians have proposed the concept of a learning health system[4] (LHS): a health system that improves itself by learning from data, continuously and in real time. This takes place through cyclical processes that mobilise health data, analyse it to create new knowledge, and apply that new knowledge to improve the health of individuals and populations (Figure 1.1).

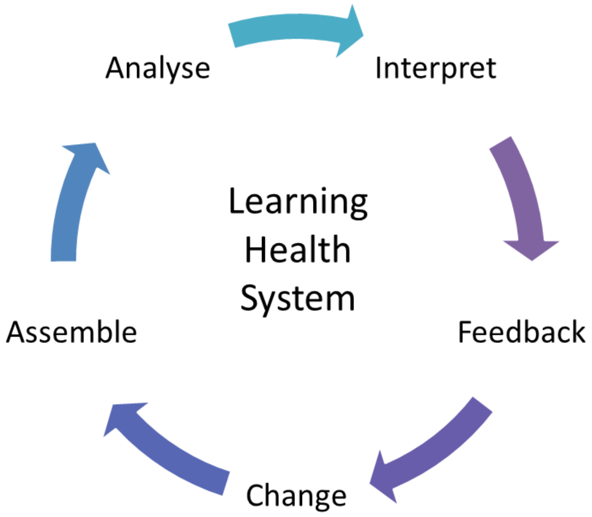


Figure 1.1 - Representation of the Learning health system concept adapted from Friedman et al[5]

LHSs are supported by infrastructures, such as cloud-based computing, that enable these processes to take place routinely and with efficiency of scale and scope. A key metric of the learning health systems is data-action latency[6]: the time lag between evidence being available and corresponding action being taken in clinical practice. Minimizing the data-action latency requires concerted data capture, data consolidation, and analysis followed swiftly by interpretation of results, assignment of responsibility for any actions, and recording of actions[6]. Ultimately, actions may be initiated in real-time, following every new data item recorded. The concept of a LHS is often applied to help improve clinical decision-making. Indeed, healthcare research uses the experiences of past patients to build understanding and learning for the future. A large area of work in this space is being able to build models that use the experiences of past patients (i.e. past data) to make predictions about the prognosis (or diagnosis) of similar patients in the future – a concept commonly referred to as predictive modelling.

1.3 Clinical prediction models

1.3.1 Background

Prognosis research is the investigation of the relationship between future outcomes among people with a given baseline health state in order to improve health[7]. The PROGnosis RESearch Strategy (PROGRESS) series[8–11] proposes a framework of research themes where the third[10] is the development, validation and impact of statistical models that predict individual risk of a future outcome, known as prognostic model research. These models are referred to as Clinical prediction models (CPMs).

CPMs have become fundamental to clinical decision support systems over recent years. They are tools/models/algorithms that compute the risk of an outcome (either in the future for prognosis, or in the past for diagnosis) given a set of patient characteristics[12,13]. Diagnostic CPMs calculate the probability that a patient currently has this outcome of interest, while prognostic CPMs calculate patient's risk of experiencing the outcome of interest at some timepoint in the future[10,14]. They are typically based on multivariable regression models, derived by analysing historical, routine healthcare data and have numerous uses across healthcare. Uses include disease prevention and management, treatment decision-making, manging supply chains and precision medicine. The Leicester diabetes risk score[15] is an example of a CPM used in clinical practice. The score is used to identify individuals at risk of developing type 2 diabetes and included in NICE guidelines for prevention of type 2 diabetes[16]. CPMs can also be used for auditing and quality assessment and the Percutaneous coronary intervention (PCI) mortality model[17] is currently used for this purpose. This model predicts 30-day mortality following PCI (a coronary revascularisation procedure) and is used for benchmarking and auditing of hospitals (i.e. predictions of the expected mortality within each hospital are calculated and then compared to the hospitals observed mortality).

Arising from the desire to move health systems away from cure to preventative medicine, CPMs have become popular and have now been routinely used over the past 2 decades. A recent systematic review showed that over 300 CPMs have been developed for cardiovascular disease alone[18]. This included two well established CPMs, the Framingham[19] and QRISK[20] models, used to compute an individual's risk of developing cardiovascular disease over the next 10 years. QRISK is included in clinical guidelines[21] and EHR systems in the UK now have QRISK2[22] embedded. Namely, if an individual has a risk above 10% they would be considered for statins and given lifestyle advice on how to reduce their risk. In response to this rise in CPM use across the health system, a guide on how to present clinical prediction models for use in clinical settings has been publication in the British Medical Journal[23].

1.3.2 Development and validation

Traditionally, cohort studies are used in the development of CPMs. Prospective cohort studies require patient recruitment and assessment of patients upon study entry. Individuals are then followed up and observed for outcomes of interest in the future. CPMs are usually developed using prospective studies but retrospective studies can also be used. Retrospective studies identify individuals who have experienced (or not experienced) an outcome of interest and then look back in their records or are interviewed about their medical history. EHRs are often a source for this data and are increasingly being used for CPM development as they become increasingly available for research and contain vast quantities of data. This is important as the more data used to develop the model, the less likely you are to observe overfitting, a problem that arises when the model captures idiosyncrasies in the development data. Riley et al[24] recently proposed how to calculate the minimum required sample size for developing a CPM and that large sample sizes are needed to ensure precise estimates.

Once the data has been collected, regression analysis is applied using a prespecified list of candidate predictors. These should be identified from previous research and clinical guidance. The final model will include a subset of predictors from the candidate predictors chosen based on statistical significance and clinical importance. Methods and best practice for the development of prediction models have been widely documented[25,26] and an MRC partnership, the MRC PROGRESS Partnership (www.progress-partnership.org), has developed guidelines and training for prognosis research[9–11,14].

The purpose of a prediction model is to provide outcome predictions for new patients. For CPMs to be used and accepted in practice we need to build trust in them and their predictive accuracy. Therefore, validation is an important aspect of prediction models that ensures the models are accurate, generalizable (to settings they would be applied to) and clinically credible. Validation can be performed internally, using data similar to the development data, and externally, using an independent (separate) dataset. Internal validation techniques include bootstrap, split-sample and cross validation; however, bootstrap validation is preferred as it leads to more accurate estimates of model performance[27]. Bootstrap validation involves sampling data, with replacement, from the raw (original) data to create a bootstrap sample. A CPM is then developed in the bootstrap data, following the same model development procedure as the CPM being validated, and then validated in both the bootstrap and original dataset. The difference in the performance between the original and bootstrap data gives what is known as the optimism. This is repeated many times, often 1000, and the average optimism across the bootstrap samples is calculated. Finally, the CPM developed on the original data is validated within the original data and the averaged optimism is subtracted from this to obtain the bootstrapped validation measure. The key metrics for model validation are calibration and discrimination. Calibration measures how well the model predictions match the observed data and discrimination refers to the models ability to distinguish between those with and without the outcome[28,29].

Calibration measures include calibration-in-the-large, calibration slope and ratio of expected and observed number of events[30,31]. Historically, the Hosmer-Lemeshow test statistic has also been used but it has limited statistical power, is sensitive to sample size and does not produce direction or magnitude of miscalibration[32]. It is therefore no longer advised for validating a CPM. Calibration-in-the-large, also known as calibration intercept, assesses the mean calibration of the model by comparing the average predicted risk with the average outcome. The target value of this measure is 0 and values below or above 0 indicate that predictions are systematically too high or too low, respectively. The calibration slope, however, has a target value of 1 and evaluates the spread of the estimated risks. If the calibration slope is less than 1, the model is overfitted, meaning the predictions are too extreme (too high for high risk and too low for low risk individuals). A slope greater than 1 means the model is underfitted and predictions are not varied enough (too high for low risk and too low for high risk individuals). The expected-observed ratio is calculated as the mean expected (predicted) outcome divided by the mean observed outcome. A value of 1 represents prefect calibration and this measure is related to the calibration-in-the-large, such that when expected-observed ratio is less than 1 the calibration-in-the-large will be greater than 0. In addition to calibration, discrimination is typically calculated using the (concordance) C-statistic, which is the proportion of concordant pairs of individuals, or D-statistic for time to event outcomes. The C-statistic bounded between 0 and 1 where a C-statistic of 1 indicates perfect discrimination but a C-statistic of 0.5 indicates the model has no discrimination and is no better than predicting the outcome randomly.

1.4 Calibration drift

1.4.1 Background

Currently CPM coefficients/parameters are estimated in such a way that they are time-invariant; hereto called static models. Hence, once a model has been developed the model and its coefficients remain fixed. Static models ignore the fact that the demographics of the population, disease prevalence and health policies may change over time. As a result, predictions based on static models quickly become 'outdated' and therefore do not provide accurate risk estimates[33]. Consequently, it is not uncommon for the agreement between the observed and predicted event rates (i.e. calibration) to worsen over time. This is known as calibration drift[34] and is one of the major pitfalls in using CPMs in practice. Calibration drift not only occurs over time but can also be present when a model is used in a population that it was not developed in and does not generalise well to the population it is being used to predict. This thesis does not consider these situations but instead focuses on calibration drift over time. Figure 1.2 illustrates calibration drift over time, from a fictional model on simulated data, where the models predicted mortality (Black line) diverges from the observed mortality (blue line) with increasing time.

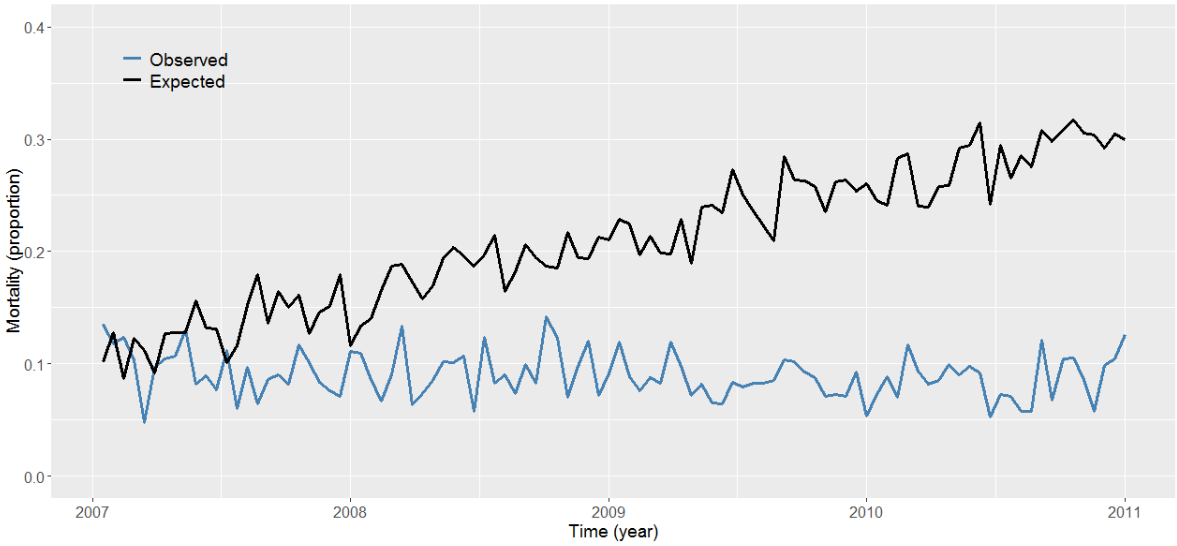


Figure 1.2 - Plot illustrating calibration drift using simulated data of a fictional model

Hickey et al[33] highlights this issue in the logistic EuroSCORE model, which is used to predict in-hospital mortality following cardiac surgery. They show that the EuroSCORE model over-predicts mortality, which thereby leads to a worsening of prediction accuracy over time. To address this issue, EuroSCORE II has been developed using more recent data. QRISK is another model that is updated yearly for this reason[22]. Calster et el[30] have also highlighted the need to improve efforts to avoid poor calibration when developing and validating prediction models. They describe the importance in calibration and the lack of attention it currently receives in the field of predictive analytics.

1.4.2 Model updating

Current practice to address calibration drift is to develop a new model, apply model updating to an existing CPM[35,36] or to aggregate existing models[37,38]. Traditionally, once calibration drift has been identified (or indeed any reduction in predictive performance of an existing CPM observed) new models would be developed de novo. However, over recent years, model updating and aggregation have become the preferred methods as they are do not discard data or existing knowledge[39,40].

Model updating uses data from more recent time points or data from a different setting to which the model was developed and regresses the original model's linear predictor against the outcome in the new data. This is known as model recalibration and provides an updated model intercept as well as a calibration slope which is a scalar value for the beta coefficients in the original model. Model revision is another updating technique which re-estimates the model coefficients to best fit the new data. Model revision can also be extended to allow additional predictors to be included in the model and is referred to as model extension[25,41]. Steyerberg et al[40] describes the approaches of model updating determines a hierarchy of complexity from model intercept adjustment, the simplest model recalibration method, to the more complex model revision methods.

Although these techniques do recalibrate models, calibration drift can still occur between updates, because model updating requires manual intervention by analysts. This means that the data-action latency period is often too long. For example, if a model is updated yearly on the first of January but clinical guidelines change on the first of February, then this could impact model calibration and cause calibration drift. Seasonal effects could also impact calibration. For example, a model might be well calibrated overall, but if it is being used to predict hospital admission for phneumonia then it is likely to under predict in the winter and over predict in the summer months. In addition to this, many models are never updated and those that do are often updated at arbitrary time points. For example, QRISK[20,22,42] is updated yearly and EuroSCORE[43,44] has only been updated once since it was originally developed in 1999. Arbitrary updating is suboptimal and is not sufficient to ensure our CPMs remain accurate over time.

In addition to this, we have recently experienced a global pandemic as a result of the SARS-CoV-2 virus, known as COVID-19. This had huge impact on healthcare services and patient populations, thus demonstrating the need to decrease the data-action latency and respond to global or regional changes that occur in the future. Hence, the healthcare system and disease populations are constantly evolving but our models remain static and are not evolving at the same rate. Therefore, there is a need to advance these approaches by ensuring a CPM is always as accurate as possible. For this to be achieved, we need to remove the latency period between observing calibration drift and updating a model.

1.4.3 Opportunity to improve model updating

The increase in data collection and sharing, along with acceptability of new statistical methods to analyse health data provides opportunity to use CPMs across healthcare and build new methodology and systems to improve patient care and outcomes. This opportunity had already resulted in the growing number of CPMs described above and has aided the growth of statistical modelling in healthcare. Therefore, this opportunity could be used to expand the CPM methodology and combine CPMs with the LHS to help address calibration drift and provide a data driven approach to model updating.

1.5 Dynamic models

1.5.1 Background

Dynamic prediction models have been developed as a potential solution to calibration drift and synergise with the learning health system framework[45]. They are clinical prediction models that estimate parameters that allow them to be time varying and not fixed values. Dynamic prediction models can therefore evolve over time with the collection of new data, continuously provide updated information and acknowledge the temporal nature of health data – thus reducing the data-action latency compared to static models or classic model updating. A dynamic model is often formulated and fitted within the Bayesian modelling framework and the model coefficients are updated as each new observation is recorded in the data. Dynamic models allow us to: 1) utilise historical data and models effectively, 2) tailor models to local populations, 3) reduce data-action latency, and 4) allow models to adapt over time. The term 'dynamic modelling' in this thesis is distinct to the use of the term in longitudinal data analysis[46], where updated prediction are made about an individual following new repeated measures about them.

1.5.2 Challenges

Hickey et al[47] has implemented a dynamic prediction model and compared the model to the standard model updating techniques discussed above. This study highlights the danger of calibration drift because it can "provide misleading indications of risk to support patient-level decision making" as well as provide “false reassurance to providers about quality of care”. The study also illustrates the potential for dynamic prediction models in healthcare but outlines some potential challenges with dynamic modelling approaches. The current challenges they highlight are: 1) difficulty in identifying model performance, 2) lack of model development and tools, 3) model complexity and 4) the arbitrariness in how to select/choose a number of hyperparameters in such models. Currently, dynamic models have limited exposure to healthcare and beyond the papers by Raftery et al[45] and Hickey et al[47] there has been little progression. McCormick et al[48] expands the dynamic modelling method from Raftery to allow for prediction of binary responses and produced the dma package[49] in R. Beyond this, there has been no methodological development and only a small number of studies have implemented dynamic modelling in healthcare.

Dynamic prediction modelling has huge potential within healthcare, and it is essential to increase our understanding of these models and how they can be developed to better facilitate healthcare by providing more accurate and precise predictions. This PhD thesis explores this field of work and some of these issues and problems that arise within the dynamic modelling framework. Section 1.6 outlines the aims and objectives for this thesis.

1.6 Aims and Thesis structure

This PhD project has three primary aims. First, to provide a comprehensive overview of the methodology available to develop dynamic prediction models, including the challenges associated with each; second to compare the model’s predictive performance; and third to develop a method to address the problem of arbitrary updating

These aims will be achieved by following the objectives:

1. Perform a literature review to identify existing methods that could be used for dynamic predictive modelling.
2. Determine the methodological challenges related to dynamic predictive modelling
3. Apply the identified dynamic modelling methods and compare their predictive performance to static CPMs using both real-world and synthetic health data.
4. Propose a method to address the problem of arbitrary updating and investigate the method under different magnitudes of miscalibration in synthetic health data.

This thesis is structured in “journal format”, as a series of papers which are previously accepted in, or about to be submitted to, peer-review journals. As recommended in the University of Manchester thesis guidelines, author contributions to each paper/chapter are described in section 1.6.1. The papers have been arranged thematically, following the chronological order of the objectives. Chapters 2 and 3 address objectives 1 and 2, while chapters 4 and 5 address objective 3 and chapter 6 addresses objective 4.

1.6.1 Author contributions

Chapter 2: Dynamic models to predict health outcomes: current status and methodological challenges. Diagnostic and Prognostic Research, 2018. DOI: 10.1186/s41512-018-0045-2

- David A. Jenkins, Matthew Sperrin, Glen P. Martin and Niels Peek designed the study. David A. Jenkins conducted the analysis and interpreted the findings in discussion with Matthew Sperrin, Glen P. Martin and Niels Peek. David A. Jenkins wrote the initial draft of the manuscript which was then critically reviewed for important intellectual content by all authors of the manuscript.

Chapter 3: Continual updating and monitoring of clinical prediction models: time for dynamic prediction systems? Diagnostic and Prognostic Research, 2021. DOI: 10.1186/s41512-020-00090-3

- David A. Jenkins conceived the commentary idea and then conceptualised it in discussion with Glen P. Martin, Matthew Sperrin and Niels Peek. David A. Jenkins wrote the initial draft of the manuscript which was then critically reviewed for important intellectual content by all authors.

Chapter 4: Development and validation of clinical prediction models in the presence of temporal trends: A simulation study comparing static and dynamic models. In preparation for submission

- David A. Jenkins, Matthew Sperrin, GPM, Thomas Debray, Mamas Mamas and Niels Peek designed the study. David A. Jenkins conducted the analysis and interpreted the findings in discussion with Matthew Sperrin, Glen P. Martin, Camilla Sammut-Powell, Mamas Mamas and Niels Peek. David A. Jenkins wrote the initial draft of the manuscript. All listed authors helped interpret the results and we plan to prepare this for submission to Statistics in Medicine.

Chapter 5: Comparing predictive performance of time invariant and time variant clinical prediction models in a UK cardiac surgery dataset. In preparation for submission

- David A. Jenkins, Matthew Sperrin, Glen P. Martin and Niels Peek designed the study. David A. Jenkins conducted the analysis and interpreted the findings in discussion with Matthew Sperrin, Glen P. Martin, Niels Peek, Benjamin Brown and Stuart Grant. David A. Jenkins wrote the initial draft of the manuscript. We plan to prepare this for submission to the Journal of the American Medical Informatics Association.

Chapter 6: Use of statistical process control to monitor the performance of a clinical prediction model.

- David A. Jenkins, Matthew Sperrin and Glen P. Martin designed the study. David A. Jenkins conducted the analysis and interpreted the findings in discussion with Matthew Sperrin and Glen P. Martin. David A. Jenkins wrote the initial draft of the manuscript. There are currently no plans for submission of this study.

1.7 References

1. Murdoch, T. B. & Detsky, A. S. The Inevitable Application of Big Data. *Jama* **309**, 1351–1352 (2014).

2. Wang, Y., Kung, L. A. & Byrd, T. A. Big data analytics: Understanding its capabilities and potential benefits for healthcare organizations. *Technol. Forecast. Soc. Change* **126**, 3–13 (2018).

3. Raghupathi, W. & Raghupathi, V. Big data analytics in healthcare: promise and potential. *Heal. Inf. Sci. Syst.* **2**, 3 (2014).

4. Friedman, C. P., Wong, A. K. & Blumenthal, D. Achieving a Nationwide Learning Health System. *Public Law* **2**, 8–10 (2010).

5. Friedman, C. P., Rubin, J. C. & Sullivan, K. J. Toward an Information Infrastructure for Global Health Improvement. *Yearb. Med. Inform.* **26**, 16–23 (2017).

6. J., A. & I., B. Combining health data uses to ignite health system learning. *Methods Inf. Med.* **54**, 479–487 (2015).

7. Kent, P., Cancelliere, C., Boyle, E., Cassidy, J. D. & Kongsted, A. A conceptual framework for prognostic research. *BMC Med. Res. Methodol.* **7**, 1–13 (2020).

8. Hemingway, H. *et al.* Prognosis research strategy (PROGRESS) 1: A framework for researching clinical outcomes. *BMJ* **346**, 1–11 (2013).

9. Riley, R. D. *et al.* Prognosis Research Strategy (PROGRESS) 2: Prognostic Factor Research. *PLoS Med.* **10**, e1001380 (2013).

10. Steyerberg, E. *et al.* Prognosis research strategy (PROGRESS) 3: prognostic model research. *PLoS Med* **10**, e1001381 (2013).

11. Hingorani, A. D. *et al.* Prognosis research strategy (PROGRESS) 4: Stratified medicine research. *BMJ* **346**, 1–9 (2013).

12. Riley, R. D., Windt, D. Van Der & Moons, K. G. M. Prognosis Research in Health Care. *Progn. Res. Heal. Care* 1–11 (2019). doi:10.1093/med/9780198796619.001.0001

13. Steyerberg, E. W. *Clinical Prediction Models. Statistics for Biology and Health. 2nd edition*. (2019).

14. Hemingway, H. *et al.* Prognosis research strategy (PROGRESS) 1: A framework for researching clinical outcomes. *BMJ* **346**, 1–11 (2013).

15. Gray, L. J. *et al.* The Leicester Risk Assessment score for detecting undiagnosed Type 2 diabetes and impaired glucose regulation for use in a multiethnic UK setting. *Diabet. Med.* **27**, 887–895 (2010).

16. (NICE), N. I. for H. and C. E. Type 2 diabetes: prevention in people at high risk. *Clinical guidelines [PH38]* (2012). Available at: https://www.nice.org.uk/guidance/ph38.

17. McAllister, K. S. L. *et al.* A contemporary risk model for predicting 30-day mortality following percutaneous coronary intervention in England and Wales. *Int. J. Cardiol.* **210**, 125–132 (2016).

18. Damen, J. A. A. G. *et al.* Prediction models for cardiovascular disease risk in the general population: Systematic review. *BMJ* **353**, (2016).

19. D'Agostino, R. B. *et al.* General Cardiovascular Risk Profile for Use in Primary Care: The Framingham Heart Study. *Circulation* **117**, 743–753 (2008).

20. Hippisley-Cox, J. *et al.* Derivation and validation of QRISK, a new cardiovascular disease risk score for the United Kingdom: Prospective open cohort study. *Br. Med. J.* **335**, 136–141 (2007).

21. (NICE), N. I. for H. and C. E. Cardiovascular disease: risk assessment and reduction, including lipid modification. *NICE Guidel. [CG181]* (2014).

22. Hippisley-Cox, J., Coupland, C., Robson, J. & Brindle, P. Derivation, validation, and evaluation of a new QRISK model to estimate lifetime risk of cardiovascular disease: Cohort study using QResearch database. *Bmj* **342**, 93 (2011).

23. Bonnett, L. J., Snell, K. I. E., Collins, G. S. & Riley, R. D. Guide to presenting clinical prediction models for use in clinical settings. *BMJ* **365**, 1–8 (2019).

24. Riley, R. D. *et al.* Minimum sample size for developing a multivariable prediction model: PART II - binary and time-to-event outcomes. *Stat. Med.* **38**, 1276–1296 (2019).

25. Steyerberg, E. W. *Clinical prediction models: a practical approach to development, validation, and updating*. (2008).

26. Harrell Jr, F. E. *Regression Modeling Strategies : With Applications to Linear Models, Logistic and Ordinal Regression, and Survival Analysis in Springer Series in Statistics*. *Springer* (Springer, 2015).

27. Steyerberg, E. W. *et al.* Internal validation of predictive models: Efficiency of some procedures for logistic regression analysis. *J. Clin. Epidemiol.* **54**, 774–781 (2001).

28. Steyerberg, E. W. *et al.* Assessing the performance of prediction models: a framework for some traditional and novel measures. *Epidemiology* **21**, 128–138 (2013).

29. Alba, A. C. *et al.* Discrimination and calibration of clinical prediction models: Users' guides to the medical literature. *JAMA - J. Am. Med. Assoc.* **318**, 1377–1384 (2017).

30. Van Calster, B. *et al.* Calibration: The Achilles heel of predictive analytics. *BMC Med.* **17**, 1–7 (2019).

31. Steyerberg, E. W. & Vergouwe, Y. Towards better clinical prediction models: Seven steps for development and an ABCD for validation. *Eur. Heart J.* **35**, 1925–1931 (2014).

32. Kramer, A. A. & Zimmerman, J. E. Assessing the calibration of mortality benchmarks in critical care: The Hosmer-Lemeshow test revisited. *Crit. Care Med.* **35**, 2052–2056 (2007).

33. Hickey, G. L. *et al.* Dynamic trends in cardiac surgery: Why the logistic euroscore is no longer suitable for contemporary cardiac surgery and implications for future risk models. *Eur. J. Cardio-thoracic Surg.* **43**, 1146–1152 (2013).

34. Davis, S. E., Lasko, T. A., Chen, G., Siew, E. D. & Matheny, M. E. Calibration drift in regression and machine learning models for acute kidney injury. *J. Am. Med. Informatics Assoc.* **24**, 1052–1061 (2017).

35. van Houwelingen, H. C. & Thorogood, J. Construction, validation and updating of a prognostic model for kidney graft survival. *Stat. Med.* **14**, 1999–2008 (1995).

36. Janssen, K. J. M., Moons, K. G. M., Kalkman, C. J., Grobbee, D. E. & Vergouwe, Y. Updating

methods improved the performance of a clinical prediction model in new patients. *J. Clin. Epidemiol.* **61**, 76–86 (2008).

37. Debray, T. P. A., Koffijberg, H., Vergouwe, Y., Moons, K. G. M. & Steyerberg, E. W. Aggregating published prediction models with individual participant data: A comparison of different approaches. *Stat. Med.* **31**, 2697–2712 (2012).

38. Martin, G. P., Mamas, M. A., Peek, N., Buchan, I. & Sperrin, M. Clinical prediction in defined populations: a simulation study investigating when and how to aggregate existing models. *BMC Med. Res. Methodol.* **17**, 1 (2017).

39. Martin, G. P., Mamas, M. A., Peek, N., Buchan, I. & Sperrin, M. A multiple-model generalisation of updating clinical prediction models. *Stat. Med.* **37**, 1343–1358 (2018).

40. Steyerberg, E. W., Borsboom, G. J. J. M., van Houwelingen, H. C., Eijkemans, M. J. C. & Habbema, J. D. F. Validation and updating of predictive logistic regression models: A study on sample size and shrinkage. *Stat. Med.* **23**, 2567–2586 (2004).

41. Toll, D. B., Janssen, K. J. M., Vergouwe, Y. & Moons, K. G. M. Validation, updating and impact of clinical prediction rules: A review. *J. Clin. Epidemiol.* **61**, 1085–1094 (2008).

42. Hippisley-Cox, J., Coupland, C. & Brindle, P. Development and validation of QRISK3 risk prediction algorithms to estimate future risk of cardiovascular disease : prospective cohort study. **2099**, 1–21 (2017).

43. Nashef, S. A. M. *et al.* European system for cardiac operative risk evaluation (EuroSCORE). *Eur. J. Cardio-thoracic Surg.* **16**, 9–13 (1999).

44. Nashef, S. A. M. *et al.* Euroscore II. *Eur. J. Cardio-thoracic Surg.* **41**, 734–745 (2012).

45. Raftery, A. E. & Ettler, P. Online Prediction Under Model Uncertainty via Dynamic Model Averaging : Application to a Cold Rolling Mill. **52**, 52–66 (2010).

46. Bull, L. M., Lunt, M., Martin, G. P., Hyrich, K. & Sergeant, J. C. Harnessing repeated measurements of predictor variables for clinical risk prediction: a review of existing methods. *Diagnostic Progn. Res.* **4**, (2020).

47. Hickey, G. L. *et al.* Dynamic prediction modeling approaches for cardiac surgery. *Circ. Cardiovasc. Qual. Outcomes* **6**, 649–658 (2013).

48. Mccormick, T. H., Raftery, A. E., Madigan, D. & Burd, R. S. Dynamic Logistic Regression and Dynamic Model Averaging for Binary Classification. *Biometrics* **68**, 23–30 (2012).

49. McCormick, T. H., Raftery, A. & Madigan, D. dma: Dynamic Model Averaging. (2018).

50. Five Year Forward View. (2014).

51. Salive, M. E. Multimorbidity in older adults. *Epidemiol. Rev.* **35**, 75–83 (2013).

52. Divo, M. J., Martinez, C. H. & Mannino, D. M. Ageing and the epidemiology of multimorbidity. *Eur. Respir. J.* **44**, 1055–1068 (2014).

53. Watkins, J. *et al.* Effects of health and social care spending constraints on mortality in England: a time trend analysis. *BMJ Open* **7**, e017722 (2017).

54. Abu-Hanna, A. & Lucas, P. J. F. Prognostic Models in Medicine. AI and Statistical Approaches.

Method Inf. Med **40**, 1–5 (2001).

55. Damen, J. A. A. G. *et al.* Prediction models for cardiovascular disease risk in the general population: systematic review. *Bmj* i2416 (2016). doi:10.1136/bmj.i2416

56. Siregar, S. *et al.* Improved Prediction by Dynamic Modeling. *Circ. Cardiovasc. Qual. Outcomes* **9**, 171–181 (2016).

57. Su, T.-L., Jaki, T., Hickey, G. L., Buchan, I. & Sperrin, M. A review of statistical updating methods for clinical prediction models. *Stat. Methods Med. Res.* 1–16 (2016). doi:10.1177/0962280215626466

58. van Houwelingen, H. C. & Thorogood, J. Construction, validation and updating of a prognostic model for kidney graft survival. *Stat. Med.* **14**, 1999–2008 (1995).

59. Draper, N. R., Nostrand, R. C. Van & Draper, N. R. Ridge Regression and James-Stein Estimation : Review and Comments Linked references are available on JSTOR for this article : Ridge Regression and James-Stein Estimation : Review and Comments. **21**, 451–466 (2016).

60. Copas, J. . Regression, Prediction and Shrinkage. *R. Stat. Soc.* **45**, 311–354 (1983).

61. Finkelman, B. S., French, B. & Kimmel, S. E. The prediction accuracy of dynamic mixed-effects models in clustered data. *BioData Min.* **9**, 5 (2016).

62. Fan, J. & Zhang, W. Statistical Methods with Varying Coefficient Models. *Stat Interface* **1**, 179–195 (2008).

63. Hoover, D. R., Rice, J. A., Wu, C. O. & Yang, L.-P. Nonparametric smoothing estimates of time-varying coefficient models with longitudinal data. *Biometrika* **85**, 809–822 (1998).

64. Madigan, D. & Raftery, A. E. Model Selection and Accounting for Model Uncertainty in Graphical Models Using Occam's Window. (1991).

65. Onorante, L. & Raftery, A. E. Dynamic model averaging in large model spaces using dynamic Occam's window. *Eur. Econ. Rev.* **81**, 2–14 (2016).

66. Ohata, T., Kaneko, M., Kuratani, T., Ueda, H. & Shimamura, K. Using the EuroSCORE to assess changes in the risk profiles of the patients undergoing coronary artery bypass grafting before and after the introduction of less invasive coronary surgery. *Ann. Thorac. Surg.* **80**, 131–135 (2005).

67. Ramsay, J. O., Wickham, H., Graves, S. & Hooker, G. fda: Functional Data Analysis. (2017).

68. Altman, D. G. & Royston, P. What do we mean by validating a prognistic model? *Stat. Med.* **19**, 453–473 (2000).

69. Justice, A. C., Covinsky, K. E. & Berlin, J. A. Assessing the generalizability of prognostic information. *Ann. Intern. Med.* **130**, 515–524 (1999).

70. Zimmerman, J. E., Kramer, A. A., McNair, D. S. & Malila, F. M. Acute Physiology and Chronic Health Evaluation (APACHE) IV: Hospital mortality assessment for today's critically ill patients. *Crit. Care Med.* **34**, 1297–1310 (2006).

71. Vergouwe, Y. *et al.* A closed testing procedure to select an appropriate method for updating prediction models. *Stat. Med.* **36**, 4529–4539 (2017).

72. Hafkamp-De Groen, E. *et al.* Predicting asthma in preschool children with asthma-like symptoms: Validating and updating the PIAMA risk score. *J. Allergy Clin. Immunol.* **132**, (2013).

73. Genders, T. S. S. *et al.* A clinical prediction rule for the diagnosis of coronary artery disease: Validation, updating, and extension. *Eur. Heart J.* **32**, 1316–1330 (2011).

74. Moons, K. G. M. *et al.* Risk prediction models: II. External validation, model updating, and impact assessment. *Heart* **98**, 691–698 (2012).

75. Hippisley-Cox, J., Coupland, C. & Brindle, P. Development and validation of QRISK3 risk prediction algorithms to estimate future risk of cardiovascular disease: Prospective cohort study. *BMJ* **357**, 1–21 (2017).

76. Martin, G. P., Sperrin, M. & Sotgiu, G. Performance of Prediction Models for Covid-19: The Caudine Forks of the External Validation. *Eur. Respir. J.* 2003728 (2020). doi:10.1183/13993003.03728-2020

77. Wynants, L. *et al.* Prediction models for diagnosis and prognosis of covid-19: Systematic review and critical appraisal. *BMJ* **369**, (2020).

78. Vergouwe, Y., Moons, K. G. M. & Steyerberg, E. W. External validity of risk models: Use of benchmark values to disentangle a case-mix effect from incorrect coefficients. *Am. J. Epidemiol.* **172**, 971–980 (2010).

79. Debray, T. P. A. *et al.* A guide to systematic review and meta-analysis of prediction model performance. *BMJ* **356**, (2017).

80. Luijken, K. *et al.* Changing predictor measurement procedures affected the performance of prediction models in clinical examples. *J. Clin. Epidemiol.* **119**, 7–18 (2020).

81. Riley, R. D. *et al.* External validation of clinical prediction models using big datasets from e-health records or IPD meta-analysis: Opportunities and challenges. *BMJ* **353**, 27–30 (2016).

82. Debray, T. P. A., Riley, R. D., Rovers, M. M., Reitsma, J. B. & Moons, K. G. M. Individual Participant Data (IPD) Meta-analyses of Diagnostic and Prognostic Modeling Studies: Guidance on Their Use. *PLoS Med.* **12**, 1–12 (2015).

83. Hippisley-Cox, J. *et al.* Predicting cardiovascular risk in England and Wales: Prospective derivation and validation of QRISK2. *Bmj* **336**, 1475–1482 (2008).

84. Jenkins, D. A., Sperrin, M., Martin, G. P. & Peek, N. Dynamic models to predict health outcomes: current status and methodological challenges. *Diagnostic Progn. Res.* **2**, 23 (2018).

85. Halabi, S. *et al.* Updated prognostic model for predicting overall survival in first-line chemotherapy for patients with metastatic castration-resistant prostate cancer. *J. Clin. Oncol.* **32**, 671–677 (2014).

86. Dawid, A. Present Position and Potential Developments : Some Personal Views : Statistical Theory : The Prequential Approach Author (s): A . P . Dawid Source : Journal of the Royal Statistical Society . Series A (General), Vol . 147 , No . 2 , The 150th Annivers. *J. R. Stat. Soc. Ser. A* **147**, 278–292 (1984).

87. Lenert, M. C., Matheny, M. E. & Walsh, C. G. Prognostic models will be victims of their own success, unless…. *J. Am. Med. Inform. Assoc.* **26**, 1645–1650 (2019).

88. Adibi, A., Sadatsafavi, M. & Ioannidis, J. P. A. Validation and Utility Testing of Clinical Prediction

Models. *JAMA* **2004**, (2020).

89. Booth, S., Riley, R. D., Ensor, J., Lambert, P. C. & Rutherford, M. J. Temporal recalibration for improving prognostic model development and risk predictions in settings where survival is improving over time. *Int. J. Epidemiol.* 1–10 (2020). doi:10.1093/ije/dyaa030

90. Jenkins, D. A. *et al.* Continual updating and monitoring of clinical prediction models: time for dynamic prediction systems? *Diagnostic Progn. Res.* **5**, 1–7 (2021).

91. Mccormick, T. H. *et al.* Dynamic Logistic Regression and Dynamic Model Averaging for Binary Classification. 23–30 (2012). doi:10.1111/j.1541-0420.2011.01645.x

92. R Core Team. R: A language and environment for statistical computing. R Foundation for Statistical. (2014).

93. Martin, G. P., Riley, R. D., Collins, G. S. & Sperrin, M. Developing clinical prediction models when adhering to minimum sample size recommendations: The importance of quantifying bootstrap variability in tuning parameters and predictive performance. *Stat. Methods Med. Res.* **30**, 2545–2561 (2021).

94. Sperrin, M., Jenkins, D., Martin, G. P. & Peek, N. Explicit causal reasoning is needed to prevent prognostic models being victims of their own success. *J. Am. Med. Informatics Assoc.* **26**, 1675–1676 (2019).

95. Hickey, G. L. *et al.* Clinical registries: Governance, management, analysis and applications. *Eur. J. Cardio-thoracic Surg.* **44**, 605–614 (2013).

96. Moons, K. G. M. *et al.* Risk prediction models: II. External validation, model updating, and impact assessment. *Heart* **98**, 691–698 (2012).

97. Davis, S. E. *et al.* A nonparametric updating method to correct clinical prediction model drift. *J. Am. Med. Informatics Assoc.* **26**, 1448–1457 (2019).

98. Minne, L. *et al.* Statistical process control for monitoring standardized mortality ratios of a classification tree model. *Methods Inf. Med.* **51**, 353–358 (2012).

99. Statistics, M. SUMS OF NONINDEPENDENT BERNOULLI RANDOM VARIABLES Author (s): Jose M . Gonzalez-Barrios Source : Brazilian Journal of Probability and Statistics , JUNE 1998 , Vol . 12 , No . 1 (JUNE Published by : Institute of Mathematical Statistics Stable URL : http. **12**, 55–64 (1998).

100. Koetsier, A., De Keizer, N. F., De Jonge, E., Cook, D. A. & Peek, N. Performance of risk-adjusted control charts to monitor in-hospital mortality of intensive care unit patients: A simulation study. *Crit. Care Med.* **40**, 1799–1807 (2012).

101. Team, R. core. R: A language and environment for statistical computing. *R Foundation for Statistical Computing, Vienna, Austria.* (2021). Available at: https://www.r-project.org/.

102. Albert, A. A. *et al.* On-line variable live-adjusted displays with internal and external risk-adjusted mortalities. A valuable method for benchmarking and early detection of unfavourable trends in cardiac surgery. *Eur. J. Cardio-thoracic Surg.* **25**, 312–319 (2004).

103. Lovegrove, J., Valencia, O., Treasure, T., Sherlaw-Johnson, C. & Gallivan, S. Monitoring the results of cardiac surgery by variable life-adjusted display. *Lancet* **350**, 1128–1130 (1997).

104. Poloniecki, J., Valencia, O. & Littlejohns, P. Correction: Cumulative risk adjusted mortality chart

for detecting changes in death rate: Observational study of heart surgery (British Medical Journal (1998) (1697-700)). *Br. Med. J.* **316**, 1947 (1998).

105. Minne, L. *et al.* Effect of changes over time in the performance of a customized SAPS-II model on the quality of care assessment. *Intensive Care Med.* **38**, 40–46 (2012).

106. Shi, X., Gallagher, C., Lund, R. & Killick, R. A Comparison of Single and Multiple Changepoint Techniques for Time Series Data. 1–29 (2021). doi:10.1016/j.csda.2022.107433

107. Wittenberg, P., Gan, F. F. & Knoth, S. A simple signaling rule for variable life-adjusted display derived from an equivalent risk-adjusted CUSUM chart. *Stat. Med.* **37**, 2455–2473 (2018).

108. Davis, S. E., Greevy, R. A., Lasko, T. A., Walsh, C. G. & Matheny, M. E. Detection of calibration drift in clinical prediction models to inform model updating. *J. Biomed. Inform.* **112**, 103611 (2020).

109. Riley, R. D. *et al.* Minimum sample size for external validation of a clinical prediction model with a binary outcome. *Stat. Med.* **40**, 4230–4251 (2021).

110. Steyerberg, E. W., Borsboom, G. J. J. M., van Houwelingen, H. C., Eijkemans, M. J. C. & Habbema, J. D. F. Validation and updating of predictive logistic regression models: A study on sample size and shrinkage. *Stat. Med.* **23**, 2567–2586 (2004).

111. Grigg, O. A., Farewell, V. T. & Spiegelhalter, D. J. Use of risk-adjusted CUSUM and RSPRT charts for monitoring in medical contexts. *Stat. Methods Med. Res.* **12**, 147–170 (2003).

112. Pagel, C. *et al.* Real time monitoring of risk-adjusted paediatric cardiac surgery outcomes using variable life-adjusted display: Implementation in three UK centres. *Heart* **99**, 1445–1450 (2013).

113. Barrett, J. & Su, L. Dynamic predictions using flexible joint models of longitudinal and time-to-event data. *Stat. Med.* **36**, 1447–1460 (2017).

114. van Houwelingen, H., & Putter, H. *Dynamic Prediction in Clinical Survival Analysis*. (2012). doi:https://doi.org/10.1201/b11311

Chapter 2 Dynamic models to predict health outcomes: current status and methodological challenges

David A. Jenkins; Matthew Sperrin; Glen P. Martin; Niels Peek

Diagnostic and Prognostic Research 2, 23 (2018). DOI: 10.1186/s41512-018-0045-2

2.1 Abstract

2.1.1 Background

Disease populations, clinical practice, and healthcare systems are constantly evolving. This can result in clinical prediction models quickly becoming outdated and less accurate over time. A potential solution is to develop 'dynamic' prediction models capable of retaining accuracy by evolving over time in response to observed changes. Our aim was to review the literature in this area to understand the current state-of-the-art in dynamic prediction modelling and identify unresolved methodological challenges.

2.1.2 Methods

MEDLINE, Embase and Web of Science were searched for papers which used or developed dynamic clinical prediction models. Information was extracted on: methods for model updating, choice of update windows and decay factors, and validation of models. We also extracted reported limitations of methods and recommendations for future research.

2.1.3 Results

We identified eleven papers that discussed seven dynamic clinical prediction modelling methods which split into three categories. The first category uses frequentist methods to update models in discrete steps, the second uses Bayesian methods for continuous updating and the third, based on varying coefficients, explicitly describes the relationship between predictors and outcome variable as a function of calendar time. These methods have been applied to a limited number of healthcare problems and few empirical comparisons between them have been made.

2.1.4 Conclusion

Dynamic prediction models are not well established but they overcome one of the major issues with static clinical prediction models, calibration drift. However, there are challenges in choosing decay factors and in dealing with sudden changes. The validation of dynamic prediction models is still largely unexplored terrain.

2.3 Introduction

Healthcare systems have limited resources and their budgets are being reduced[50], while there are increasing numbers of people living with one or more long term conditions[51,52]. This can have a negative effect on health outcomes[53], and systems therefore need to be more efficient. One way to improve efficiency is by implementing preventative measures which delay or prevent onset of disease and increase the overall health of the population. Increased data collection in healthcare systems, and availability of large scale data sources provide an opportunity to effectively target care and resources in a data-driven way. This could also be used to guide health policies, assist in healthcare auditing and select appropriate therapies in individual patient management, along with other uses[54] to improve the healthcare system as a whole.

Clinical prediction models (CPMs) are used for diagnosis or prediction of future outcomes for individuals[10,55], and thus have the potential to be used for decision making and effective targeting of resources. CPMs use information about an individual at a given time, to compute the risk/probability of a future outcome; they have been increasingly used over the past 2 decades. CPMs are currently used to support various decisions. For example, QRISK[20] computes an individual's risk of developing cardiovascular disease over the next 10 years and if the individual's risk is above 10% then they would be considered for statins.

Over time, population demographics, prevalence of disease, clinical practice, and the healthcare system as a whole may change, meaning that predictions based on static data can become outdated and hence no longer accurate. This is known as calibration drift[34] and is one of the major pitfalls in using CPMs in practice[33]. It can lead to over or under prescribing of treatment and, if the model is used for audit purposes, it can provide misleading results because it does not correctly adjust for case mix. QRISK[22] is updated yearly for this reason. However, this provides periodic updates, and although this is a step in the right direction, it is problematic because patients' calculated risk changes abruptly when updates are applied, while patients' actual outcomes do not.

It would be advantageous if models could be produced that would continuously update over time as more information is collected and made available, thus providing accurate risk predictions that respond rapidly to new information. This could reduce the use of outdated models and avoid multiple models being produced and used, reducing both time and effort. This approach is known as dynamic prediction modelling. We define dynamic models (DMs) as those which acknowledge the real time of each point, are designed to evolve over time and address the problem of calibration drift. The model could, in principle, change after a single new observation, which could be a structural change or a coefficient change. Models can evolve over time and an individual's risk can also change over time. Here we focus on models evolving over time as opposed to the alternative where we observe repeated measures for an individual and observe time varying coefficients.

Our aim was to review methods for developing and validating dynamic prediction models, in order to understand the current state-of-the-art in this field and identify unresolved methodological challenges.

2.4 Methods

2.4.1 Search strategy

The literature search was conducted in three electronic databases, Medline, Embase and Web of Science. OVID was used to search the former two databases, and searches were restricted to English language because of limited translation resources but were not restricted by publication year. The Medline search terms comprise terms the authors considered to best describe dynamic prediction modelling (Table 2.1). The search was tailored to each database and supplemented with relevant papers that were identified from the reference list of the included papers. Further snowballing, using Google Scholar, was also performed by conducting a citation search which identified papers referencing our initial relevant paper list.

Table 2.1 - Ovid search terms

1	dynamic model*.mp. [mp=ti, ab, hw, tn, ot, dm, mf, dv, kw, fx, nm, kf, px, rx, ui, sy]
2	dynamic prediction*.mp. [mp=ti, ab, hw, tn, ot, dm, mf, dv, kw, fx, nm, kf, px, rx, ui, sy]
3	clinical prediction model*.mp. [mp=ti, ab, hw, tn, ot, dm, mf, dv, kw, fx, nm, kf, px, rx, ui, sy]
4	dynamic model* prediction.mp. [mp=ti, ab, hw, tn, ot, dm, mf, dv, kw, fx, nm, kf, px, rx, ui, sy]
5	dynamic regression.mp. [mp=ti, ab, hw, tn, ot, dm, mf, dv, kw, fx, nm, kf, px, rx, ui, sy]
6	dynamic logistic regression.mp. [mp=ti, ab, hw, tn, ot, dm, mf, dv, kw, fx, nm, kf, px, rx, ui, sy]
7	model updating.mp. [mp=ti, ab, hw, tn, ot, dm, mf, dv, kw, fx, nm, kf, px, rx, ui, sy]
8	clinical prediction.mp. [mp=ti, ab, hw, tn, ot, dm, mf, dv, kw, fx, nm, kf, px, rx, ui, sy]
9	(dynamic model* and updat*).af.
10	dynamic prediction model*.af.
11	model revision.mp. [mp=ti, ab, hw, tn, ot, dm, mf, dv, kw, fx, nm, kf, px, rx, ui, sy]
12	model recalibration.mp. [mp=ti, ab, hw, tn, ot, dm, mf, dv, kw, fx, nm, kf, px, rx, ui, sy]
13	1 or 2 or 4 or 5 or 6 or 9 or 10
14	3 or 8
15	13 and 14
16	7 or 11 or 12
17	15 or 16
18	dynamic.mp. [mp=ti, ab, hw, tn, ot, dm, mf, dv, kw, fx, nm, kf, px, rx, an, ui, sy]
19	14 and 18
20	17 or 19

2.4.2 Selection of studies

A two-stage screening process was conducted by one author (DJ) to assess the relevance of studies and was applied after the initial search and again after the two snowballing approaches. The first stage consisted of screening the titles and abstracts of citations to exclude articles that did not meet the inclusion criteria. The eligible criteria for inclusion were original methodological peer reviewed journal articles which considered: 1) dynamic prediction models (DPMs); 2) model updating methods that could be performed in real time; or 3) model coefficients as functions of time. Exclusion criteria were determined in advance and included: conference proceedings, papers with methods that could not change over time or update in real time, static prediction models and models that only consider a single time point (eg models for cross sectional data). Dynamic survival models were also excluded because they do not fall under our definition of dynamic prediction. Applied research, without any methodological work, was excluded because our interest was around the current state-of-the-art and methodology in the area.

2.4.3 Extraction

We evaluated papers on two general domains: modelling methods, and validation and evaluation. We extracted each method we considered to meet, or have the potential to meet, our definition of dynamic modelling. For validation we extracted how the models implemented were evaluated. For all the methods found during the search, we also extracted any modelling challenges and further work discussed by the authors and provide our suggestions for the future work needed in the area.

No specific study measures or synthesis were calculated across studies.

2.5 Results

Our initial search resulted in 1034 papers, with 61 considered potentially relevant after abstract and title screening. After full article screening 8 were identified for which information was extracted and snowballing was taken place. An additional 14 papers were then considered relevant but after title and abstract screening only 3 were included for which information was extracted. Hence, in total, 11 papers were deemed relevant for final inclusion (see figure 2.1).

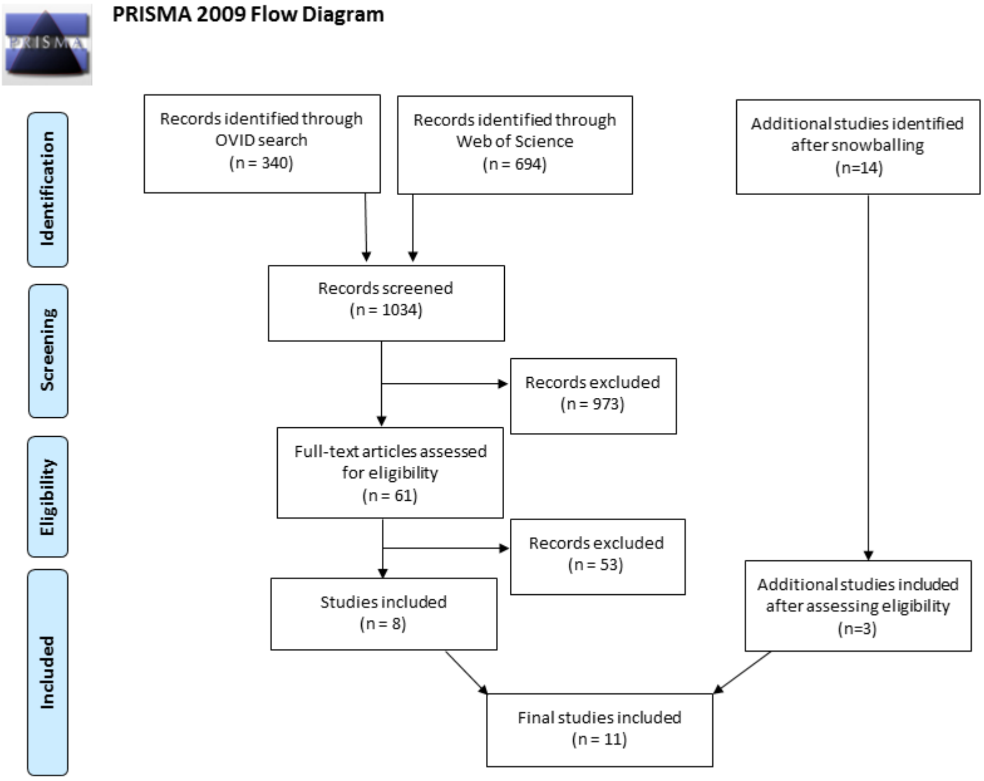

PRISMA 2009 Flow Diagram
Identification
Screening
Eligibility
Included
Records identified through OVID search (n = 340)
Records identified through Web of Science (n = 694)
Additional studies identified after snowballing (n=14)
Records screened (n = 1034)
Records excluded (n = 973)
Full-text articles assessed for eligibility (n = 61)
Records excluded (n = 53)
Studies included (n = 8)
Additional studies included after assessing eligibility (n=3)
Final studies included (n = 11)

Figure 2.1 - PRISMA diagram of included studies

Seven methods were reported across 11 papers which could be used to deal with calibration drift in prediction models (see table 2.2). These can be split into three categories: discrete model updating, Bayesian model updating and varying coefficient modelling.

Table 2.2 - Tick table of methods included in each paper

	Modelling methods						
	Discrete model updating				Bayesian model updating		Varying coefficient modelling
Author	Intercept update	Overall slope update	Individual slopes update	Model revision	Bayesian dynamic modelling	Bayesian model averaging	
Fan							✓
Finkelman					✓		
Hickey				✓	✓		
Hoover							✓
Janssen	✓	✓	✓	✓			
McCormick					✓	✓	
Raftery					✓	✓	
Siregar	✓	✓		✓	✓		
Steyerberg	✓	✓	✓	✓			
Su	✓	✓	✓	✓	✓	✓	
Van Houwelingen			✓				

To illustrate the approaches, we will focus on prognosis and consider a regression model with either a continuous or binary end point at a fixed point in time. A response y_t is observed for an individual at a time $t = (t_1, \dots, t_n)$, and a vector of predictors $x_t = (x_{tk}: k = 1, \dots, K)$ such that:

$$g\big(E(y_t)\big) = \beta_0(t) + \beta_K(t)x_t, \qquad (1)$$

for link g, where $\beta_0(t)$ is the intercept and $\beta_K(t)$ is a vector of the regression coefficients for the *K* predictors at time t.

Equation 1 is a general form of a dynamic prediction model, but the methods described in the literature vary on how to estimate the coefficient functions $\beta_0(t)$ and $\beta_K(t)$ and update the model. Below we outline each of the methods found in the literature, followed by a discussion of the various challenges highlighted within the papers.

2.5.1 Modelling methods

2.5.1.1 Discrete model updating – Model recalibration and revision

Discrete model updating methods use new data over time to update the model. Using a single observation or small group of observations can result in an unstable and less accurate model. Hence, these methods are updated in batches at set times, for example, each month or year, to ensure a sufficient amount of data is collected and used for the update. We denote these batch times as $T_j = (T_1, \dots, T_B)$ where $B \ll n$.

Four discrete updating methods are explained/used in the final included papers. All four methods consider a frequentist approach. 'Intercept update', 'overall slope update', 'individual slopes update', and 'model revision'.

The 'intercept update' method[36,40,56,57] fits a regression model to the new data, at updating batch time T_j, using the linear predictor of Equation 1 as an offset. This recalculates a new intercept, $\beta_0\big(T_j\big)$ with $\beta_K\big(T_j\big) = \beta_K$ remaining constant over time.

'Overall slope update'[36,40,56,57] re-estimates both the intercept and an overall slope $\alpha\big(T_j\big)$ for each update time T_j. This factor is used to proportionally adjust the original coefficients and thus creates a new predictor-outcome association $\beta_K(T_j) = \alpha(T_j)\beta_K\big(T_{j-1}\big)$ and a new intercept, $\beta_0(T_j)$.

'Individual slopes update'[36,40,57,58] is a two-step method where the overall slope updating is first used and then a subset of the coefficients, which are statistically different in the new data compared with the historic data, are re-estimated. Thus, $\beta_K\big(T_j\big) = \alpha\big(T_j\big)\beta_K\big(T_{j-1}\big) + \gamma_K(T_j)$ where $\gamma_K(T_j)$ is a vector of length *k* which has zeros located in the elements corresponding to the parameter estimates that are not re-estimated. The choice of which variable coefficients to re-estimate can be decided by a likelihood ratio test, stepwise variable selection or obtaining expert opinion. A special case[36,40,47,56–58] would be to update all model coefficients and not only those that are statistically different. Hence, all prognostic effects are updated and the original CPM is only used to select the covariates included in the updated model. After revision, shrinkage can be conducted, where the coefficient estimates are shrunk towards the recalibration estimates[36,40,57]. This can be done, for example, using ridge regression[59,60].

'Model revision'[40,57] considers adding predictors into the model. This method re-estimates existing coefficients as in the above approaches, but also tests if any additional predictors now have a statistically significant effect in model fit by performing likelihood ratio tests in a forward stepwise variable selection manner. This builds a new model

$$g\big(E(y_t)\big) = \beta_0(T) + \beta_L(T)\, x_t \qquad (2)$$

Where L is the total number of predictors in the updated model at time T_j, such that, $L \geq K$ and $L - K$ is the number of additional predictors added to the model at time T_j. This model is applied for individuals t such that $T_j \leq t < T_{j+1}$.

2.5.1.2 Continuous model updating - Bayesian updating

Two continuous updating methods are discussed in five of the final included papers[45,47,48,56,57,61]. The first method is known as Bayesian dynamic modelling and the second, known as dynamic model averaging, is a generalisation of the first. In both methods, the information obtained from past data is used as prior information and combined with the new data to obtain updated estimates. Thus, the updating (posterior) equation is proportional to the product of the likelihood (at time t) and the prior (prediction equation at time t-1),

$$\mathrm{p}(\,\beta_K(t) \mid Y_t) \propto \mathrm{p}(\,\beta_K(t) \mid\; Y_{t-1})\mathrm{p}(\,y_t \mid \beta_K(t)\,) \propto \text{Prior x Likelihood} \qquad (3)$$

Where the prediction equation (Prior) is obtained through Kalman filtering by supposing $\mathrm{p}(\beta(\mathrm{t}-1) \mid\; \mathrm{Y}_{\mathrm{t}-1}) \sim \mathrm{N}(\hat{\beta}(\mathrm{t}-1), \varphi_{t-1})$, where $\mathrm{Y}_{\mathrm{t}-1} = \{\mathrm{y}_1, \dots, \mathrm{y}_{\mathrm{t}-1}\}$. This results in the prediction equation

$$\mathrm{p}(\beta_K(t) \mid Y_{t-1}) \sim \mathrm{N}\big(\hat{\beta}_K(\mathrm{t}-1), \mathrm{R}_\mathrm{t}\big);\; \mathrm{R}_\mathrm{t} = \varphi_{t-1} + W_t, \qquad (4)$$

where W_t represents the covariance matrix.

We can also introduce a forgetting factor, λ_t, such that $\mathrm{R}_\mathrm{t} = {}^{\varphi_{t-1}}\!/_{\lambda_\mathrm{t}}$. This down weights (or decays) historical observations so they have less influence/weight than new data by essentially inflating the variance of the prior. Typically, λ_t is constant over time, and close to 1. In principle the forgetting factor could change over time, but this has yet to be done in practice. A forgetting factor of 0.99 was used in one study[48], while another[57] performed sensitivity analysis using different values for the forgetting factor. Su et al[57] suggests that λ_t can be selected using an auto tuning procedure at each time point which could result in a time varying forgetting factor. However, this would result in a much higher computation load.

Two advantages are discussed in using the forgetting factor. The first is that the model becomes less computationally demanding than when forgetting is not applied, which can make the dynamic model more feasible to use in practice. The second is that the model relies less on the historical data. If the model coefficients are changing over time then giving a lot of weight to past data may decrease prediction accuracy. Also, the historical data used for the prior could anchor the results and provide inaccurate predictions.

The first method described is for a single model case, but if there exist multiple models $\mathrm{M}_1, \dots, \mathrm{M}_\mathrm{m}$ implemented at the same times then the above approach can be applied simultaneously to each model. We can then combine each of the m models together to create one final model, thus resulting in dynamic model averaging (DMA). In DMA a weighted average of models is used at each time point, where the ‘better’ models contribute more weight in the final model and the weights can vary over time. One major advantage of this approach is that it allows parameters to be down weighted/excluded and emerging factors to become present over time. Hence, there is extra flexibility in this approach that the others do not have and forgetting can also be applied within DMA.

All of the above methods, both continuous and discrete, are two-step approaches in which the initial CPM is computed using the first batch of data and that model is subsequently updated in light of new data. The initial model will generally fix which parameters are included within the model, although, as described above, there is discussion in the literature[41] about adding or deleting predictors during the updating. The majority of studies set a specific time interval where all the data within that window would be used for the next update. One study[48] had data to perform monthly updates but another[56] only considered updating yearly, and one[33] considered updating models on either a monthly, yearly or 2 yearly basis. Finckleman et al[61] was the only paper to consider the batch as observation numbers and not length of time. They considered 250, 500, 1000 and 5000 for the updating batch numbers and concluded the results were “fairly insensitive to changes” in the size of the update. Some papers[33,61] have suggested, for the discrete methods, that a sufficient number of new data are needed in each batch to ensure enough is obtained for accurate and stable predictions. Step one of these models will not always consider the same time period as step two (model updating). For example, one of the models Hickey et al[47] conducts, uses a first step of 12 months to create the initial CPM but then uses monthly updates for step two.

Some of the studies also compared which of the methods performed best. However, not all methods were included in each paper. Raftery[45] used mean square error (MSE) and maximum absolute error (MAE) to compare the Bayesian models. After 200 sample updates, the models become stable and differences between models become smaller than in the initial sample updates where the DMA performs better because “it’s more adaptive”. Finkelman also used MAE to compare models, but to improve interpretability computed the ‘relative improvement’, which is the improvement of the current model compared with the intercept-only model. McCormick on the other hand, used the Brier score to compare model performance.

2.5.1.3 Varying coefficient model

Varying coefficient models[62] (also known as functional response models) were developed to explore dynamic patterns in data. These are particularly useful when we encounter multiple data from the same individuals over time, known as longitudinal data, and/or have data changing over time, known as functional data. Varying coefficient models are often used to model longitudinal data, for example, risk of HIV after birth[63] and $\beta_K(t)$ is modelled as a function of time from birth. We can also use it as an approach to dynamic modelling in which the relationship between predictors and outcome variable is described as a function of calendar time. This approach has been used in other areas, such as economics, but not yet in healthcare.

Following the form of Equation 1, in this case we have $\beta_0(t)$ and $\beta_K(t) = (\beta_1(t), \dots, \beta_K(\mathrm{t}))$ as functions of time which are assumed to be smooth. Hoover et al[63] presents three ways the coefficients can be estimated: kernel, polynomial and smoothing splines.

A special case of this method is where only the intercept is dependent on time. Equation 1 would then become

$$g(E(y_t)) = \beta_0 + \beta_K x_t + \beta_{time} t,$$

where the betas are no longer functions of time and β_{time} adjust the intercept for observed calibration drift in the development dataset, i.e $\beta_0(t) = \beta_0 + \beta_{time} t$. This is arguably the simplest approach to overcome the problem of calibration drift but to the best of our knowledge it has not been applied in healthcare settings to improve calibration. Compared with the previous methods, the varying coefficient model does not regularly update at each time but rather attempts to estimate the complete function of the coefficients over time given data up to a certain time point. Hence, this method does not view data as a stream but rather assumes all data are available over time and then estimates $\beta_K(t)$. No study considering varying coefficient models also considered discrete or continuous updating approaches. A comparison between the different methods has yet to be explored.

All included papers discuss dynamic models as a way of using all the data available to create models that are evolving over time and have the flexibility of adapting to a changing landscape over time. The discrete and continuous updating models use current/new data to update past knowledge, rather than using a static time frame and assuming the prediction model remains the same over time. However, the weight applied to the historic data varies. On the one hand, all data, historic and new, is used equally. On the other hand, the historic data may be given no weight in the update, so only the new data is used to update the model. These are just two extremes and dynamic model updating can be anywhere within this space. The functional varying coefficient models differ because they are not updating over time. These models use the complete data available to estimate the coefficient function over time in order to provide future predictions. However, they have the potential to be updated using discrete updating but this has yet to be explored.

2.5.2 Model Validation

Once a model has been computed and selection of appropriate predictors has taken place, it is not sufficient to assume the model is accurate and predicts well. We therefore need to formally validate our models. For static CPMs, cross-validation and bootstrap validation are the recommended methods over split sample or external sample validation[56], but validation is more complex when it comes to DPMs. The literature around validating a dynamic CPM is much less established, meaning that it was not possible to identify different validation techniques for each of the dynamic modelling methods separately.

Siregar et al[56] and Su et al[57] assess calibration and discrimination in all of their models. Both validate their models in subsequent years (after model updating has stopped) but in reality the model would continue updating and so a way is needed to validate in this framework to provide real time validation without using the same data that is then used for the model. Su et al[57] also note that because validation is conducted at a separate time to the model, then their validation constitutes transportability rather than validation. Split sampling could be performed, where part of the sample at each update is used to validate the updated model, but this was not explored in any of the papers and would need doing in a dynamic way which could add to the computational aspect of the models. Van Houweingen[58] conducts a split sample validation on the original CPM and uses this to determine if an update is needed as the new data is collected, however, validation of the updating model was not undertaken. There would also be a lag time from determining if a model is valid, such that, the model would possibly have been updated many more times. McCormick[64] on the other hand, designed their model for a setting when data is not stored and so validating can be an issue here. They suggest maximizing the average one-step-ahead prediction by updating a tuning parameter, but not through numerical optimisation because of computational infeasibility. Using an Occams window[65] approach was also discussed as a possible solution to computational problems. This would consider a smaller model space at each time (after time 1) by using a subset of all possible models based on a cut off value for each models weight contribution at the previous time. However, none of these methods have been formally implemented in any of the included papers or across healthcare, although this has been applied within economics[43].

Hickey et al[33] produced time series plots of the beta values to obtain inferences of the association between the outcome and risk factor. This allowed for comparison of methods, as well as the ability to visually detect any abrupt changes. Although this provides a better understanding of how the models are working, it is not a formal way in which to validate, test or compare models. Hickey et al[33] acknowledge that not performing validation was a limitation of the study and suggests that to do so one would need to compute and monitor the models discrimination over time, in a continuous way. Conducting time series on the coefficients could potentially be used to detect patterns in the coefficient estimate over time or even be used as a way to predict future beta estimates which could then be compared to the DM predictions, but either has yet to be explored.

Therefore, validation is a clear issue in this area and was only used in a small number of studies which mainly considered the discrete approaches.

2.5.3 Other challenges

All of the methods described above assume a steady change in the model coefficients over time. However, sudden large changes are possible and could result in poor model performance. These changes can occur for many reasons, such as, a change in policy, introduction of new interventions, a change in data collection, or the introduction of clinical decision support that is based on the CPM. An example of a step change in clinical practice is the introduction of less invasive coronary surgery[33]. This change in surgery, along with a change in the case-mix of the population undergoing cardiac surgery resulted in the EUROSCORE CPM[66] largely over predicting patient risk[55]. One way to model these changes in a CPM would be to include a time factor but it has yet to be discussed in the literature how well dynamic models react to these changes and which models provide the most accurate predictions and should be used in these circumstances. However, this assumes that a step change is anticipated for a known reason. However, in practice it is not always anticipated or known. Therefore, it would also be advantageous to account for, and model, unexpected step changes. McCormick et al[48] suggests that when these occur, a smaller forgetting factor should be chosen to allow for these changes. However, the windowed approach in Hickey et al[47] is used to dampen any abrupt changes. Step changes have the potential to impact model accuracy and being able to identify them, as well as, knowing how to deal with them could have great benefit. There is currently little work discussing what to do when they occur and how to detect or define a true step change. Analysing the impact of these changes (with various magnitudes and frequency) on model performance and understanding how best to weight past data (if at all) when they occur would be largely beneficial for future work. Also, ability to detect step changes would be valuable and could be used to either identify when models need to be updated or inform the user a change has occurred and investigation into the data in needed.

Finally, computational complexity was discussed as a limitation of DMs but only two papers[45,61] formally considered computation time. Finkelman et al[61] discuses that computation time linearly increased with the number of updates, but around the same number of subjects were included in each update and computation time could vary if the numbers varied across iterations. Raftery et al[45] discusses that although DMs and DMA does increase computation time, they are still well within a range for practical application. In a large system when updating is to be applied when each new data point is collected, then this could be problematic if the computation time associated with updates exceeds the time between subsequent data points. Continuous model updating is then not feasible.

Software is available to perform dynamic modelling, the dma[49] and fda[67] packages in R can be used for the continuous updating and the varying coefficient methods, respectively. To our knowledge, no package is available for discrete updating, but it can easily be programmed manually in many software packages. Extension of these, along with user friendly tutorials, would aid widespread implementation into the clinical setting.

2.6 Discussion

In this study, we conducted a literature review which has identified three main types of dynamic modelling, with the main differences between the methods emerging in relation to how the coefficients are estimated. Our review has enabled us to draw together all the methods within one paper and highlight gaps in the literature for future research. Discrete and continuous updating have been used a small number of times within the healthcare setting to address the issue of calibration drift. These methods update the model over time, which provides the dynamic aspect of these models. We have also identified an additional method, varying coefficient modelling, that could be used in healthcare but has yet to be implemented for dynamic prediction in this setting. This method differs in comparison to the others as it does not update but uses the data up to time t to estimate the function for each coefficient in the model over time. The continuous updating and varying coefficient methods both assume a smooth function over time and discrete updating differs by assuming discrete changes. These dynamic prediction models have the potential to be extremely useful but currently have limited exposure to healthcare problems and validation of these models in practice is challenging. Further work is needed to develop ways to validate these models and assess how these models perform under different healthcare settings and scenarios.

To our knowledge, only two other studies have performed a review of dynamic modelling methods. Su et al[57] describes both the discrete and continuous updating methods and then applies them to a clinical data set, updating on a monthly basis. Comparisons of model performance and accuracy of future predictions were then made. Siregar et al[56] also describes the discrete and continuous updating methods, excluding dynamic model averaging. The methods were then applied to a cardiac data set by updating the EUROSCORE model and comparing model discrimination across all methods. Overall, our work is consistent with these two papers but extends the findings by conducting an up to date literature search and includes the use of varying coefficient modelling as a possible method to maintain model performance over time. Comparisons of the intercept updating method with different updating times and population size were compared with the standard continuous updating method by Hickey et al[47]. This work compares the methods in a real-world situation and discusses limitations of the methods, but it is not a complete review of dynamic modelling. Our review draws together all methods in the literature and identifies gaps in the literature but does not provide practical examples and direct comparisons of all the methods found.

The most pressing problem to address, which we have highlighted in this study, is that of validation. The purpose of any model validation is collected incremental evidence that the model works satisfactorily in populations where it is applied – thus provided trust among its potential users and enabling adoption[68,69]. Many well-established (static) prognostic models, such as the Apache IV model[70] for predicting mortality in critically ill patients, were validated in numerous studies before they were broadly adopted in clinical practice. Because dynamic prediction models are moving targets, it is fundamentally impossible to follow the same approach. We can validate each of the individual iterations, but by the time that users have taken notice of the validation results, the model will have already progressed to a next iteration and those results might be outdated. So, to enable a similar mechanism that instills trust and fosters adoption, validation methods are needed that can provide evidence of good performance of the entire dynamic 'system'. These methods should convince us that both the initial model and all its future iterations have good performance, regardless of the new data points that are used for updating.

Future work would also benefit from assessing the impact of step changes, as well as the impact size and frequency of updates could have on predictions. A close test procedure has previously been used[71] to select which discrete updating method should be used when updating your model. However, this has only been used for transportability to a new population, opposed to updating regularly over time. Exploring this method to address calibration drift, as well as, extending the method to include Bayesian updating and decide when/if updating should occur would be extremely useful and increase the utility of the approach. Testing and comparing these dynamic models in more complex data structures, such as clustered data, would also be beneficial. This could be done with the use of random effects or generalised estimating equations, as previously suggested[61,63]. Also, only a small number of studies have applied and considered these dynamic modelling methods for use within healthcare, with the majority of applications only considering the discrete updating methods[72] and focusing on transportability for models to different populations[73,74] rather than using the methods discussed to address temporal changes over time. Therefore, more practical examples and comparisons of the methods found are warranted for further work. This would help aid the broader adoption of these methods into clinical practice, which is a current issue with CPMs as a whole. While this is not confined to dynamic prediction models, this is a common problem with prediction models and refinements, such as, improved reporting and better use of existing CPMs (e.g. a focus on external validation rather than de novo development) could improve the adoption of CPMs in clinical practice. Also, incorporating models into hand-held technology (e.g. mobile apps to allow calculation of complex models a patient's bedside) and extending the methods into software with user friendly tutorials would be of value.

Because dynamic prediction models are an emerging field and not a well-established concept, different authors may have used different terminologies to describe dynamic prediction models; further, there are currently no MeSH terms for these methods and this could have resulted in some studies not being captured within our search. Our search focussed on the methodological papers and it was not possible to go through all of the applied work. This may have resulted in some methods, or adaptations of existing methods, not being captured within our search. Nevertheless, we believe that we have identified the main methodological approaches to dynamic model development, updating, and validation.

Although the focus of this review was in methods accounting for temporal differences over time, some of the methods and issues raised would apply to geographic or contextual updating, for example, where a model is to be used in a different population to which it was originally developed. Also, although we restrict our attention to prognostic models, the findings are generalizable to diagnostic modelling.

2.7 Conclusion

Several statistical methods for creating dynamic prediction models have been described in the literature. These methods are well developed but their application to real-world clinical prediction problems is sparse and no dynamic prediction models have been deployed in clinical practice. Validation of dynamic prediction models is an unresolved issue that needs to be addressed urgently.

2.8 References

1. Murdoch, T. B. & Detsky, A. S. The Inevitable Application of Big Data. *Jama* **309**, 1351–1352 (2014).

2. Wang, Y., Kung, L. A. & Byrd, T. A. Big data analytics: Understanding its capabilities and potential benefits for healthcare organizations. *Technol. Forecast. Soc. Change* **126**, 3–13 (2018).

3. Raghupathi, W. & Raghupathi, V. Big data analytics in healthcare: promise and potential. *Heal. Inf. Sci. Syst.* **2**, 3 (2014).

4. Friedman, C. P., Wong, A. K. & Blumenthal, D. Achieving a Nationwide Learning Health System. *Public Law* **2**, 8–10 (2010).

5. Friedman, C. P., Rubin, J. C. & Sullivan, K. J. Toward an Information Infrastructure for Global Health Improvement. *Yearb. Med. Inform.* **26**, 16–23 (2017).

6. J., A. & I., B. Combining health data uses to ignite health system learning. *Methods Inf. Med.* **54**, 479–487 (2015).

7. Kent, P., Cancelliere, C., Boyle, E., Cassidy, J. D. & Kongsted, A. A conceptual framework for prognostic research. *BMC Med. Res. Methodol.* **7**, 1–13 (2020).

8. Hemingway, H. *et al.* Prognosis research strategy (PROGRESS) 1: A framework for researching clinical outcomes. *BMJ* **346**, 1–11 (2013).

9. Riley, R. D. *et al.* Prognosis Research Strategy (PROGRESS) 2: Prognostic Factor Research. *PLoS Med.* **10**, e1001380 (2013).

10. Steyerberg, E. *et al.* Prognosis research strategy (PROGRESS) 3: prognostic model research. *PLoS Med* **10**, e1001381 (2013).

11. Hingorani, A. D. *et al.* Prognosis research strategy (PROGRESS) 4: Stratified medicine research. *BMJ* **346**, 1–9 (2013).

12. Riley, R. D., Windt, D. Van Der & Moons, K. G. M. Prognosis Research in Health Care. *Progn. Res. Heal. Care* 1–11 (2019). doi:10.1093/med/9780198796619.001.0001

13. Steyerberg, E. W. *Clinical Prediction Models. Statistics for Biology and Health. 2nd edition*. (2019).

14. Hemingway, H. *et al.* Prognosis research strategy (PROGRESS) 1: A framework for researching clinical outcomes. *BMJ* **346**, 1–11 (2013).

15. Gray, L. J. *et al.* The Leicester Risk Assessment score for detecting undiagnosed Type 2 diabetes and impaired glucose regulation for use in a multiethnic UK setting. *Diabet. Med.* **27**, 887–895 (2010).

16. (NICE), N. I. for H. and C. E. Type 2 diabetes: prevention in people at high risk. *Clinical guidelines [PH38]* (2012). Available at: https://www.nice.org.uk/guidance/ph38.

17. McAllister, K. S. L. *et al.* A contemporary risk model for predicting 30-day mortality following percutaneous coronary intervention in England and Wales. *Int. J. Cardiol.* **210**, 125–132 (2016).

18. Damen, J. A. A. G. *et al.* Prediction models for cardiovascular disease risk in the general population: Systematic review. *BMJ* **353**, (2016).

19. D'Agostino, R. B. *et al.* General Cardiovascular Risk Profile for Use in Primary Care: The Framingham Heart Study. *Circulation* **117**, 743–753 (2008).

20. Hippisley-Cox, J. *et al.* Derivation and validation of QRISK, a new cardiovascular disease risk score for the United Kingdom: Prospective open cohort study. *Br. Med. J.* **335**, 136–141 (2007).

21. (NICE), N. I. for H. and C. E. Cardiovascular disease: risk assessment and reduction, including lipid modification. *NICE Guidel. [CG181]* (2014).

22. Hippisley-Cox, J., Coupland, C., Robson, J. & Brindle, P. Derivation, validation, and evaluation of a new QRISK model to estimate lifetime risk of cardiovascular disease: Cohort study using QResearch database. *Bmj* **342**, 93 (2011).

23. Bonnett, L. J., Snell, K. I. E., Collins, G. S. & Riley, R. D. Guide to presenting clinical prediction models for use in clinical settings. *BMJ* **365**, 1–8 (2019).

24. Riley, R. D. *et al.* Minimum sample size for developing a multivariable prediction model: PART II - binary and time-to-event outcomes. *Stat. Med.* **38**, 1276–1296 (2019).

25. Steyerberg, E. W. *Clinical prediction models: a practical approach to development, validation, and updating*. (2008).

26. Harrell Jr, F. E. *Regression Modeling Strategies : With Applications to Linear Models, Logistic and Ordinal Regression, and Survival Analysis in Springer Series in Statistics*. *Springer* (Springer, 2015).

27. Steyerberg, E. W. *et al.* Internal validation of predictive models: Efficiency of some procedures for logistic regression analysis. *J. Clin. Epidemiol.* **54**, 774–781 (2001).

28. Steyerberg, E. W. *et al.* Assessing the performance of prediction models: a framework for some traditional and novel measures. *Epidemiology* **21**, 128–138 (2013).

29. Alba, A. C. *et al.* Discrimination and calibration of clinical prediction models: Users' guides to the medical literature. *JAMA - J. Am. Med. Assoc.* **318**, 1377–1384 (2017).

30. Van Calster, B. *et al.* Calibration: The Achilles heel of predictive analytics. *BMC Med.* **17**, 1–7 (2019).

31. Steyerberg, E. W. & Vergouwe, Y. Towards better clinical prediction models: Seven steps for development and an ABCD for validation. *Eur. Heart J.* **35**, 1925–1931 (2014).

32. Kramer, A. A. & Zimmerman, J. E. Assessing the calibration of mortality benchmarks in critical care: The Hosmer-Lemeshow test revisited. *Crit. Care Med.* **35**, 2052–2056 (2007).

33. Hickey, G. L. *et al.* Dynamic trends in cardiac surgery: Why the logistic euroscore is no longer suitable for contemporary cardiac surgery and implications for future risk models. *Eur. J. Cardio-thoracic Surg.* **43**, 1146–1152 (2013).

34. Davis, S. E., Lasko, T. A., Chen, G., Siew, E. D. & Matheny, M. E. Calibration drift in regression and machine learning models for acute kidney injury. *J. Am. Med. Informatics Assoc.* **24**, 1052–1061 (2017).

35. van Houwelingen, H. C. & Thorogood, J. Construction, validation and updating of a prognostic model for kidney graft survival. *Stat. Med.* **14**, 1999–2008 (1995).

36. Janssen, K. J. M., Moons, K. G. M., Kalkman, C. J., Grobbee, D. E. & Vergouwe, Y. Updating

methods improved the performance of a clinical prediction model in new patients. *J. Clin. Epidemiol.* **61**, 76–86 (2008).

37. Debray, T. P. A., Koffijberg, H., Vergouwe, Y., Moons, K. G. M. & Steyerberg, E. W. Aggregating published prediction models with individual participant data: A comparison of different approaches. *Stat. Med.* **31**, 2697–2712 (2012).

38. Martin, G. P., Mamas, M. A., Peek, N., Buchan, I. & Sperrin, M. Clinical prediction in defined populations: a simulation study investigating when and how to aggregate existing models. *BMC Med. Res. Methodol.* **17**, 1 (2017).

39. Martin, G. P., Mamas, M. A., Peek, N., Buchan, I. & Sperrin, M. A multiple-model generalisation of updating clinical prediction models. *Stat. Med.* **37**, 1343–1358 (2018).

40. Steyerberg, E. W., Borsboom, G. J. J. M., van Houwelingen, H. C., Eijkemans, M. J. C. & Habbema, J. D. F. Validation and updating of predictive logistic regression models: A study on sample size and shrinkage. *Stat. Med.* **23**, 2567–2586 (2004).

41. Toll, D. B., Janssen, K. J. M., Vergouwe, Y. & Moons, K. G. M. Validation, updating and impact of clinical prediction rules: A review. *J. Clin. Epidemiol.* **61**, 1085–1094 (2008).

42. Hippisley-Cox, J., Coupland, C. & Brindle, P. Development and validation of QRISK3 risk prediction algorithms to estimate future risk of cardiovascular disease : prospective cohort study. **2099**, 1–21 (2017).

43. Nashef, S. A. M. *et al.* European system for cardiac operative risk evaluation (EuroSCORE). *Eur. J. Cardio-thoracic Surg.* **16**, 9–13 (1999).

44. Nashef, S. A. M. *et al.* Euroscore II. *Eur. J. Cardio-thoracic Surg.* **41**, 734–745 (2012).

45. Raftery, A. E. & Ettler, P. Online Prediction Under Model Uncertainty via Dynamic Model Averaging : Application to a Cold Rolling Mill. **52**, 52–66 (2010).

46. Bull, L. M., Lunt, M., Martin, G. P., Hyrich, K. & Sergeant, J. C. Harnessing repeated measurements of predictor variables for clinical risk prediction: a review of existing methods. *Diagnostic Progn. Res.* **4**, (2020).

47. Hickey, G. L. *et al.* Dynamic prediction modeling approaches for cardiac surgery. *Circ. Cardiovasc. Qual. Outcomes* **6**, 649–658 (2013).

48. Mccormick, T. H., Raftery, A. E., Madigan, D. & Burd, R. S. Dynamic Logistic Regression and Dynamic Model Averaging for Binary Classification. *Biometrics* **68**, 23–30 (2012).

49. McCormick, T. H., Raftery, A. & Madigan, D. dma: Dynamic Model Averaging. (2018).

50. Five Year Forward View. (2014).

51. Salive, M. E. Multimorbidity in older adults. *Epidemiol. Rev.* **35**, 75–83 (2013).

52. Divo, M. J., Martinez, C. H. & Mannino, D. M. Ageing and the epidemiology of multimorbidity. *Eur. Respir. J.* **44**, 1055–1068 (2014).

53. Watkins, J. *et al.* Effects of health and social care spending constraints on mortality in England: a time trend analysis. *BMJ Open* **7**, e017722 (2017).

54. Abu-Hanna, A. & Lucas, P. J. F. Prognostic Models in Medicine. AI and Statistical Approaches.

Method Inf. Med **40**, 1–5 (2001).

55. Damen, J. A. A. G. *et al.* Prediction models for cardiovascular disease risk in the general population: systematic review. *Bmj* i2416 (2016). doi:10.1136/bmj.i2416

56. Siregar, S. *et al.* Improved Prediction by Dynamic Modeling. *Circ. Cardiovasc. Qual. Outcomes* **9**, 171–181 (2016).

57. Su, T.-L., Jaki, T., Hickey, G. L., Buchan, I. & Sperrin, M. A review of statistical updating methods for clinical prediction models. *Stat. Methods Med. Res.* 1–16 (2016). doi:10.1177/0962280215626466

58. van Houwelingen, H. C. & Thorogood, J. Construction, validation and updating of a prognostic model for kidney graft survival. *Stat. Med.* **14**, 1999–2008 (1995).

59. Draper, N. R., Nostrand, R. C. Van & Draper, N. R. Ridge Regression and James-Stein Estimation : Review and Comments Linked references are available on JSTOR for this article : Ridge Regression and James-Stein Estimation : Review and Comments. **21**, 451–466 (2016).

60. Copas, J. . Regression, Prediction and Shrinkage. *R. Stat. Soc.* **45**, 311–354 (1983).

61. Finkelman, B. S., French, B. & Kimmel, S. E. The prediction accuracy of dynamic mixed-effects models in clustered data. *BioData Min.* **9**, 5 (2016).

62. Fan, J. & Zhang, W. Statistical Methods with Varying Coefficient Models. *Stat Interface* **1**, 179–195 (2008).

63. Hoover, D. R., Rice, J. A., Wu, C. O. & Yang, L.-P. Nonparametric smoothing estimates of time-varying coefficient models with longitudinal data. *Biometrika* **85**, 809–822 (1998).

64. Madigan, D. & Raftery, A. E. Model Selection and Accounting for Model Uncertainty in Graphical Models Using Occam's Window. (1991).

65. Onorante, L. & Raftery, A. E. Dynamic model averaging in large model spaces using dynamic Occam's window. *Eur. Econ. Rev.* **81**, 2–14 (2016).

66. Ohata, T., Kaneko, M., Kuratani, T., Ueda, H. & Shimamura, K. Using the EuroSCORE to assess changes in the risk profiles of the patients undergoing coronary artery bypass grafting before and after the introduction of less invasive coronary surgery. *Ann. Thorac. Surg.* **80**, 131–135 (2005).

67. Ramsay, J. O., Wickham, H., Graves, S. & Hooker, G. fda: Functional Data Analysis. (2017).

68. Altman, D. G. & Royston, P. What do we mean by validating a prognistic model? *Stat. Med.* **19**, 453–473 (2000).

69. Justice, A. C., Covinsky, K. E. & Berlin, J. A. Assessing the generalizability of prognostic information. *Ann. Intern. Med.* **130**, 515–524 (1999).

70. Zimmerman, J. E., Kramer, A. A., McNair, D. S. & Malila, F. M. Acute Physiology and Chronic Health Evaluation (APACHE) IV: Hospital mortality assessment for today's critically ill patients. *Crit. Care Med.* **34**, 1297–1310 (2006).

71. Vergouwe, Y. *et al.* A closed testing procedure to select an appropriate method for updating prediction models. *Stat. Med.* **36**, 4529–4539 (2017).

72. Hafkamp-De Groen, E. *et al.* Predicting asthma in preschool children with asthma-like symptoms: Validating and updating the PIAMA risk score. *J. Allergy Clin. Immunol.* **132**, (2013).

73. Genders, T. S. S. *et al.* A clinical prediction rule for the diagnosis of coronary artery disease: Validation, updating, and extension. *Eur. Heart J.* **32**, 1316–1330 (2011).

74. Moons, K. G. M. *et al.* Risk prediction models: II. External validation, model updating, and impact assessment. *Heart* **98**, 691–698 (2012).

75. Hippisley-Cox, J., Coupland, C. & Brindle, P. Development and validation of QRISK3 risk prediction algorithms to estimate future risk of cardiovascular disease: Prospective cohort study. *BMJ* **357**, 1–21 (2017).

76. Martin, G. P., Sperrin, M. & Sotgiu, G. Performance of Prediction Models for Covid-19: The Caudine Forks of the External Validation. *Eur. Respir. J.* 2003728 (2020). doi:10.1183/13993003.03728-2020

77. Wynants, L. *et al.* Prediction models for diagnosis and prognosis of covid-19: Systematic review and critical appraisal. *BMJ* **369**, (2020).

78. Vergouwe, Y., Moons, K. G. M. & Steyerberg, E. W. External validity of risk models: Use of benchmark values to disentangle a case-mix effect from incorrect coefficients. *Am. J. Epidemiol.* **172**, 971–980 (2010).

79. Debray, T. P. A. *et al.* A guide to systematic review and meta-analysis of prediction model performance. *BMJ* **356**, (2017).

80. Luijken, K. *et al.* Changing predictor measurement procedures affected the performance of prediction models in clinical examples. *J. Clin. Epidemiol.* **119**, 7–18 (2020).

81. Riley, R. D. *et al.* External validation of clinical prediction models using big datasets from e-health records or IPD meta-analysis: Opportunities and challenges. *BMJ* **353**, 27–30 (2016).

82. Debray, T. P. A., Riley, R. D., Rovers, M. M., Reitsma, J. B. & Moons, K. G. M. Individual Participant Data (IPD) Meta-analyses of Diagnostic and Prognostic Modeling Studies: Guidance on Their Use. *PLoS Med.* **12**, 1–12 (2015).

83. Hippisley-Cox, J. *et al.* Predicting cardiovascular risk in England and Wales: Prospective derivation and validation of QRISK2. *Bmj* **336**, 1475–1482 (2008).

84. Jenkins, D. A., Sperrin, M., Martin, G. P. & Peek, N. Dynamic models to predict health outcomes: current status and methodological challenges. *Diagnostic Progn. Res.* **2**, 23 (2018).

85. Halabi, S. *et al.* Updated prognostic model for predicting overall survival in first-line chemotherapy for patients with metastatic castration-resistant prostate cancer. *J. Clin. Oncol.* **32**, 671–677 (2014).

86. Dawid, A. Present Position and Potential Developments : Some Personal Views : Statistical Theory : The Prequential Approach Author (s): A . P . Dawid Source : Journal of the Royal Statistical Society . Series A (General), Vol . 147 , No . 2 , The 150th Annivers. *J. R. Stat. Soc. Ser. A* **147**, 278–292 (1984).

87. Lenert, M. C., Matheny, M. E. & Walsh, C. G. Prognostic models will be victims of their own success, unless…. *J. Am. Med. Inform. Assoc.* **26**, 1645–1650 (2019).

88. Adibi, A., Sadatsafavi, M. & Ioannidis, J. P. A. Validation and Utility Testing of Clinical Prediction

Models. *JAMA* **2004**, (2020).

89. Booth, S., Riley, R. D., Ensor, J., Lambert, P. C. & Rutherford, M. J. Temporal recalibration for improving prognostic model development and risk predictions in settings where survival is improving over time. *Int. J. Epidemiol.* 1–10 (2020). doi:10.1093/ije/dyaa030

90. Jenkins, D. A. *et al.* Continual updating and monitoring of clinical prediction models: time for dynamic prediction systems? *Diagnostic Progn. Res.* **5**, 1–7 (2021).

91. Mccormick, T. H. *et al.* Dynamic Logistic Regression and Dynamic Model Averaging for Binary Classification. 23–30 (2012). doi:10.1111/j.1541-0420.2011.01645.x

92. R Core Team. R: A language and environment for statistical computing. R Foundation for Statistical. (2014).

93. Martin, G. P., Riley, R. D., Collins, G. S. & Sperrin, M. Developing clinical prediction models when adhering to minimum sample size recommendations: The importance of quantifying bootstrap variability in tuning parameters and predictive performance. *Stat. Methods Med. Res.* **30**, 2545–2561 (2021).

94. Sperrin, M., Jenkins, D., Martin, G. P. & Peek, N. Explicit causal reasoning is needed to prevent prognostic models being victims of their own success. *J. Am. Med. Informatics Assoc.* **26**, 1675–1676 (2019).

95. Hickey, G. L. *et al.* Clinical registries: Governance, management, analysis and applications. *Eur. J. Cardio-thoracic Surg.* **44**, 605–614 (2013).

96. Moons, K. G. M. *et al.* Risk prediction models: II. External validation, model updating, and impact assessment. *Heart* **98**, 691–698 (2012).

97. Davis, S. E. *et al.* A nonparametric updating method to correct clinical prediction model drift. *J. Am. Med. Informatics Assoc.* **26**, 1448–1457 (2019).

98. Minne, L. *et al.* Statistical process control for monitoring standardized mortality ratios of a classification tree model. *Methods Inf. Med.* **51**, 353–358 (2012).

99. Statistics, M. SUMS OF NONINDEPENDENT BERNOULLI RANDOM VARIABLES Author (s): Jose M . Gonzalez-Barrios Source : Brazilian Journal of Probability and Statistics , JUNE 1998 , Vol . 12 , No . 1 (JUNE Published by : Institute of Mathematical Statistics Stable URL : http. **12**, 55–64 (1998).

100. Koetsier, A., De Keizer, N. F., De Jonge, E., Cook, D. A. & Peek, N. Performance of risk-adjusted control charts to monitor in-hospital mortality of intensive care unit patients: A simulation study. *Crit. Care Med.* **40**, 1799–1807 (2012).

101. Team, R. core. R: A language and environment for statistical computing. *R Foundation for Statistical Computing, Vienna, Austria.* (2021). Available at: https://www.r-project.org/.

102. Albert, A. A. *et al.* On-line variable live-adjusted displays with internal and external risk-adjusted mortalities. A valuable method for benchmarking and early detection of unfavourable trends in cardiac surgery. *Eur. J. Cardio-thoracic Surg.* **25**, 312–319 (2004).

103. Lovegrove, J., Valencia, O., Treasure, T., Sherlaw-Johnson, C. & Gallivan, S. Monitoring the results of cardiac surgery by variable life-adjusted display. *Lancet* **350**, 1128–1130 (1997).

104. Poloniecki, J., Valencia, O. & Littlejohns, P. Correction: Cumulative risk adjusted mortality chart

for detecting changes in death rate: Observational study of heart surgery (British Medical Journal (1998) (1697-700)). *Br. Med. J.* **316**, 1947 (1998).

105. Minne, L. *et al.* Effect of changes over time in the performance of a customized SAPS-II model on the quality of care assessment. *Intensive Care Med.* **38**, 40–46 (2012).

106. Shi, X., Gallagher, C., Lund, R. & Killick, R. A Comparison of Single and Multiple Changepoint Techniques for Time Series Data. 1–29 (2021). doi:10.1016/j.csda.2022.107433

107. Wittenberg, P., Gan, F. F. & Knoth, S. A simple signaling rule for variable life-adjusted display derived from an equivalent risk-adjusted CUSUM chart. *Stat. Med.* **37**, 2455–2473 (2018).

108. Davis, S. E., Greevy, R. A., Lasko, T. A., Walsh, C. G. & Matheny, M. E. Detection of calibration drift in clinical prediction models to inform model updating. *J. Biomed. Inform.* **112**, 103611 (2020).

109. Riley, R. D. *et al.* Minimum sample size for external validation of a clinical prediction model with a binary outcome. *Stat. Med.* **40**, 4230–4251 (2021).

110. Steyerberg, E. W., Borsboom, G. J. J. M., van Houwelingen, H. C., Eijkemans, M. J. C. & Habbema, J. D. F. Validation and updating of predictive logistic regression models: A study on sample size and shrinkage. *Stat. Med.* **23**, 2567–2586 (2004).

111. Grigg, O. A., Farewell, V. T. & Spiegelhalter, D. J. Use of risk-adjusted CUSUM and RSPRT charts for monitoring in medical contexts. *Stat. Methods Med. Res.* **12**, 147–170 (2003).

112. Pagel, C. *et al.* Real time monitoring of risk-adjusted paediatric cardiac surgery outcomes using variable life-adjusted display: Implementation in three UK centres. *Heart* **99**, 1445–1450 (2013).

113. Barrett, J. & Su, L. Dynamic predictions using flexible joint models of longitudinal and time-to-event data. *Stat. Med.* **36**, 1447–1460 (2017).

114. van Houwelingen, H., & Putter, H. *Dynamic Prediction in Clinical Survival Analysis*. (2012). doi:https://doi.org/10.1201/b11311

Chapter 3 Continual updating and monitoring of clinical prediction models: time for dynamic prediction systems?

David A. Jenkins; Glen P. Martin; Matthew Sperrin; Richard D. Riley; Thomas P.A. Debray; Gary S. Collins and Niels Peek

Diagnostic and Prognostic Research (2021). DOI: 10.1186/s41512-020-00090-3

3.1 Abstract

Clinical prediction models (CPMs) have become fundamental for risk stratification across healthcare. The CPM pipeline (development, validation, deployment and impact assessment) is commonly viewed as a one-time activity, with model updating rarely considered and done in a somewhat ad-hoc manner. This fails to address the fact that the performance of a CPMs worsen over time as natural changes in populations and care pathways occur. CPMs need constant surveillance to maintain adequate predictive performance. Rather than reactively updating a developed CPM once evidence of deteriorated performance accumulates, it is possible to proactively adapt CPMs whenever new data becomes available. Approaches for validation then need to be changed accordingly, making validation a continuous rather than a discrete effort. As such, "living" (dynamic) CPMs represent a paradigm shift, where the analytical methods dynamically generate updated versions of a model through time; one then needs to validate the system rather than each subsequent model revision.

3.2 Background

Clinical prediction models (CPMs) are tools that compute the risk of an outcome given a set of patient characteristics (‘predictors’), and can be used for informing diagnosis or prognosis in individuals[10,14]. They are typically based on multivariable regression models, for example as derived by analysing historical cohort data or routinely collected healthcare data. Arising from the desire to move health systems away from managing or curing disease towards preventative medicine, CPMs have become popular and several are now embedded in clinical practice (e.g. QRISK3[75] and the Leicester diabetes risk score[15]).

Commonly, the process of developing a CPM equation is a one-time activity, with estimates of model parameters obtained from a single Dataset ignoring time. Once a model has been developed, usually the model equation remains fixed until a revision is conducted. However, revisions are rare and usually undertaken at an arbitrary time, or following an external validation that suggests the model is miscalibrated. Model validation is an important aspect of the CPM pipeline, and aims to evaluate whether model predictions are accurate (in settings they would be applied to in practice). Similarly to model development, validation is often a one-time activity. Commonly, the literature refers to CPMs as being “validated”, but this may create a false impression that no more model testing needs to be performed. In this paper, we propose moving away from one-time model development and validation, and rather embed CPM development, validation and updating into a dynamic system that reflects an evolving healthcare service. For example, the current covid-19 pandemic represents a situation where this would be particularly useful, given how quickly healthcare processes have changed, meaning that any prediction models for covid-19 need to be updated rapidly[76,77] . For example, in the future, vaccinations, immunity build up and virus mutation may affect the strength of predictor effects over time.

3.3 Calibration drift prediction problem

CPM production pipelines are built on the assumption that once produced and verified, evidence can be translated into practice ad infinitum. But the distribution of patient characteristics, disease prevalence and health policies change over time. When these changes occur, the estimated CPM parameters and corresponding predictions may no longer be valid[7833]. Consequently, the agreement between the observed and predicted event rates worsens over time[79]: so called calibration drift[34]. Hickey et al[33] highlights this issue in the logistic EuroSCORE model[43], which quickly became outdated as improvements in patient outcomes were rapid. Therefore, there is evidence that model coefficients need to change through time, as illustrated with EuroSCORE. In addition, Luijken et al[80] observed that changing predictor measurement procedures induced miscalibration in nine real-world examples.

Traditional practice to address this is to develop another CPM de novo. However, alternative approaches such as updating[35,36] aggregating existing CPMs[37,39], or meta-analysis of individual participant data[81,82], are preferable because they do not discard historical data and previous research efforts[38]. For example, models such as QRISK are now updated yearly[20,75] using contemporary data and also revised to include additional predictors (such as the revision of QRISK2[83] into QRISK3[75]). Nonetheless, this updating (recalibration) is still relatively uncommon, often occurs a substantial time after model development, is often undertaken at arbitrary time points and is typically dependent on funding. For example, EuroSCORE II[44] was developed in 2012, some 13 years after the original model, and it is unclear when this will be updated again. The problem with this approach to model validation and revision is that predictive performance of a CPM may only be investigated many years after the model has been developed. Although this can subsequently result in the CPM being updated, incorrect decisions may have already been made as a result of the mis-calibrated model and harm already caused.

Typically, a model is developed or updated under the assumption that the data are well described by a fixed underlying model where the coefficients are constant across the observation period used to develop the model. If the prevalence of an outcome is increasing at a steady rate during a 5 year window of data collection and then used to develop the model, the CPM will be calibrated to the middle of the window and not the most recent data. The overarching issue here, for both development and validation, is that the data generating process could change through time. While frequent model updating will mitigate these issues, it does not eliminate the problem since commonly used methods do not acknowledge temporal changes. Rather, we propose embedding prediction models in practice to ensure development, validation and updating is a continual process. We now discuss how this might be implemented and the challenges involved.

3.4 Possible solution and challenges

The healthcare system and disease populations are constantly changing but the CPMs we deploy are not updating at the same rate. Therefore, we need to ensure a CPM is maintained on a continual (rather than an ad hoc) basis. For this to be achieved we need to reduce the latency period between observing calibration drift and updating a model. Thus, moving towards a service that constantly monitors a model, and has an embedded feedback loop where the monitoring information is then relayed back to the model, and used to modify and maintain it.

3.4.1 Dynamic models

Dynamic prediction models have been proposed as a potential solution to calibration drift and to allow prediction models to evolve simultaneously with the healthcare system[45,84]. They are a collection of analytical methods that allow CPMs to continuously adapt as data on new patients arises– thus reducing the data-action latency compared with traditional methods of developing CPMs at a single point in time. By dynamic model we mean models that update over calendar time as data on new individuals arises, not models that update predictions for individuals as new data on them arises. A dynamic model is formulated to account for the calendar time that a prediction is made, that is the calendar time predictors are recorded for each individual (e.g. date of GP appointment), and is designed to evolve over time, such that the parameter estimates are not constrained to remain fixed as (calendar) time evolves. Thus, given a fixed set of patient characteristics, a dynamic model could produce different predicted risks at different times of prediction, for example, if two individuals with the same predictor values are observed at different times then the model could produce different predicted risks.

The simplest approach to develop a dynamic CPM is to include (calendar) time as a predictor[62,63]. Alternatively, the Bayesian dynamic model could be implemented, where information obtained from past data is used as prior information and combined with new data to obtain updated estimates, thereby updating with new observations in real-time[45,48,84]. More weight can also be given to the most recent data by 'forgetting' past data at a given rate. For more detail on these methods see the reviews by Jenkins et al[84] and Su et al[57].In summary, dynamic models allow us: 1) to utilise historical data and models more effectively, 2) to reduce data-action latency (time between changes in the data and reacting to them), and 3) to "automatically" adapt model parameters over time. Hickey et al[47] illustrates the use of dynamic modelling in EuroSCORE and shows how the coefficients change over time.

Although there is much potential in dynamic models, they are rarely used in healthcare. There are both methodological and practical reasons why this is so. Methodological reasons include: 1) a lack of methods on how to validate dynamic prediction models[84]; 2) uncertainty on when to include new or exclude existing predictors; 3) deciding how much to discount historical data; 4) uncertainty around when to update the model; 5) the potential lack of model transparency; and 6) inconsistent outputs over time (e.g. a patient with the stable risk factors could have changing predicted risks because the model has changed). Practical considerations include: 1) lack of robust and suitable new data to be able to update the models continuously; 2) complexity of the dynamic modelling approach; 3) lack of software implementations; 4) lack of requisite expertise by those developing the model; and 5) lack of infrastructure and funding. However, many of these problems are not specific to dynamic CPMs, for example, the problem of how to handle historical data in traditional CPMs is often ignored but a problem is still present. When updating CPMs we often append the new data to past data or use only the recent data to perform the update. This is an arbitrary choice by the researcher performing the update and neither is likely to be optimal. Raftery et al[45] attempted to address this in dynamic modelling by using an approach to choose how to discount past data at each update by optimising the predictive performance over past samples, but this is computationally expensive. More of these challenges have also attempted to be addressed in statistical literature, for example, use of the time dependent AUC[85], but have yet to be applied to continual prognostic modelling. Other theoretical methods to address these challenges also exist, but their application in prognostic model research is generally lacking and it remains unclear how this would and should affect prediction model research.

3.4.2 Model surveillance

If a dynamic model evolves with every new data point, then there is only ever the next data point in which to validate each evolution of the model. Furthermore, validation at a given time point is only a single snapshot in time. It does not follow that if a CPM, dynamic or otherwise, has high performance at a given point in time that it will always perform well. However, as we continue to make predictions for new patients, we can record and monitor the accuracy, essentially continuously monitoring and testing for calibration drift (prequential testing[86]). This leads to the idea of model surveillance, where the CPM monitoring could be performed after every new data point or at given intervals. Prequential testing approaches have a long history in the statistical literature and have been used in areas such as economic forecasting. However, they have yet to be transported and used in prediction model research. Lenert et al[87] discuss the notion of having surveillance of models used in practice as the models themselves can directly impact the data and subsequently their own performance. They explain that without surveillance, models will have limited effectiveness and can become hazardous. We propose prequential testing as a potential solution to these issues but further research is required.

3.4.3 Feedback loop

Model surveillance, and the use of prequential testing, could also allow us to address some of the issues discussed above. However, continuous monitoring of performance will not address all of these problems. The results of continuous monitoring need to be transported back into the model providing a feedback loop, which allows the model to learn and ensures the model continually provides accurate predictions (Figure 3.1). Ideally this would be conducted in a timely manner to reduce the data-action latency, which is a key metric of the learning health system (LHS)[5], a system that improves itself by learning from new data through cyclic processes that mobilise data to create new knowledge and then use that knowledge to improve. We therefore need a system approach, where one encompasses clinical prediction modelling into a learning health system, thus resulting in a learning prediction system. This system could improve itself by learning from data, continually and in real time and would take place through cyclical processes (Figure 3.1).

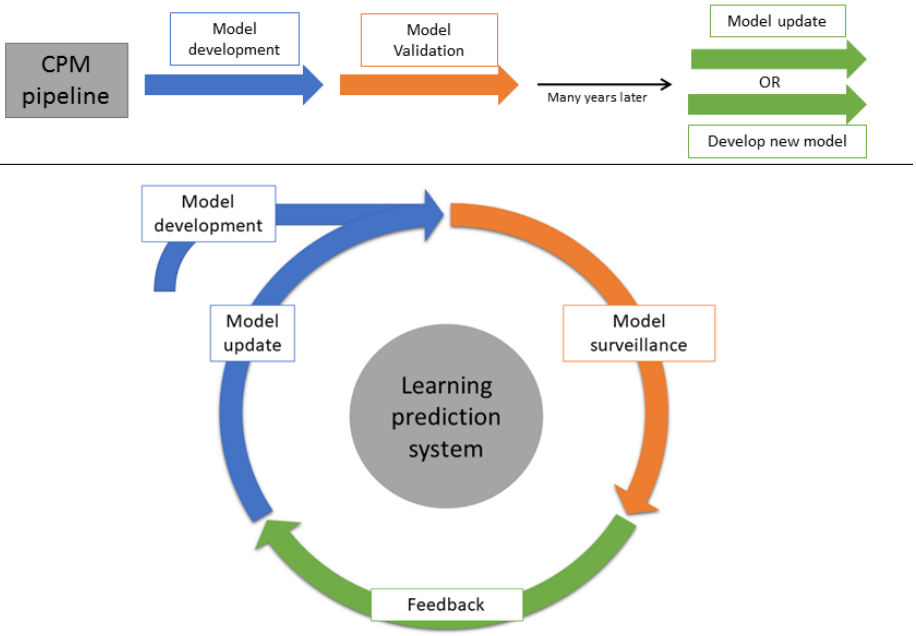


Figure 3.1 - Illustration of the current CPM pipeline (top) and the proposed learning prediction system (bottom)

Minimizing the data-action latency, and doing so efficiently, requires concerted data capture, aggregation, and analysis followed swiftly by interpretation of results, assignment of responsibility for any actions, and recording of actions. Not only can a learning prediction system allow a model to evolve over time, but it could also decide when and how to evolve each iteration of the cycle. This is achievable in LHSs that are supported by infrastructures that enable these processes to take place routinely and with efficiency of scale and scope. Dynamic methods (updating and/or monitoring) offer a flexible solution, requiring less manual labour, but need the infrastructure and sustained resources in place to implement them. Adibi et al[88] discusses an integrated infrastructure for CPMs and highlights that much of the technology is available, but not yet fully utilised in healthcare. For dynamic updating to work, a system is needed where patient data is automatically collected and stored in a database and subsequently used to update parameter estimates.

3.4.4 Further considerations

We acknowledge that continual updating a CPM might not always be needed. For example, comparative audit requires a standardised method to adjust for case-mix differences, so dynamic methods might not be appropriate. Also, updating all of the coefficients in a model may not always be a good idea. Booth et al[89] recently proposed temporal recalibration in settings where survival is improving over time. This approach develops a model using all the available data but then recalibrates the baseline survival function using a subset of the data from a recent time window. Vergouwe et al[71] described a closed test procedure to select methods for updating prediction models, something which could be embedded into the learning prediction system. This study also found that model revision, updating all model coefficients, can be chosen over intercept-only-updating, even in small sample sizes. Further supporting the need for a continual system. Although we could redevelop or update traditional models on a daily basis, the use of dynamic methods may offer a more flexible solution. Both traditional and dynamic approaches to CPM development/updating have their advantages and disadvantages (see Table 3.1), but ultimately all CPMs need their performance to be monitored regularly and thus require a continual flow of data.

Table 3.1 - Summary of the characteristics and pros and cons for different modelling approaches

Models	**Characteristics**	**Advantages**	**Disadvantages**
Existing approaches			
Fixed model never updated	• Model and coefficients fixed • Never updated	• Cheap (funding available) • Low complexity and easy to communicate	• Can become miscalibrated quickly • Dethroned by new model likely developed in future • Ends up as research waste • Loss of information
Model with ad hoc updating (e.g. EuroSCORE)	• Updated when opportunity allows • Fixed coefficients between updates	• Easy to maintain • Cheap (funding available) • Low complexity • Little manual labour • Advantageous over developing a completely new model	• Non responsive to calibration drift • Long data-action latency
Models that get periodically updated (e.g. QRISK)	• Fixed regular updates • Set time period between updates	• Lower chance of miscalibration than above • Allows predictors to be included/excluded from the model • Relatively low complexity	• Funding required • Can still observe calibration drift between updates • Increased maintenance • Requires more than manual labour to maintain • Uncertainty on length of time needed between updates
Proposed approaches			
Models with discrete updating and continual validation/monitoring (learning prediction system with discrete updating and continual monitoring)	• Updated when opportunity allows • Continuously monitors new data • Updated as a result of the monitoring • Feeds back information to the model on how and when to update	• Monitoring informs updates • Only update when required • Reactive to changes • Transports well across settings and populations	• Funding and infrastructure required • Update does not immediately follow after suggestion from monitoring • Requires some manual labour to maintain

Complete dynamic system (continual model update with continual validation/monitoring) (learning prediction system with continual updating and monitoring)	• Dynamic model • Continuously monitors new data • Feeds back information to the model	• Efficient • Potential to be more accurate • Provides less miscalibrated results • ‘Reacts’ quicker to change (responsive) • Possible to automate • Less manual labour to maintain • Transports well across settings and populations • Do not need to store the data	• Requires access to an appropriate “living” data source that is linked to the relevant outcomes. • Uncertainty on how one should validate dynamic prediction models • Uncertainty on when to include/exclude predictors • Deciding how much to discount historical data • Uncertainty around when to update the model • Lack of software packages • Complexity of approach • Lack of requisite expertise by those developing the model • Lack of transparency • Inconsistent outputs from day to day • Funding

Dynamic CPMs requires a continual flow of data. These are typically provided by routine data sources such as audit data, registries and electronic health records. Dynamic CPMs also offer opportunity in remote monitoring data, such as wearable device or app data, which provides large quantities of data in real time that is otherwise challenging to analyse. However, continuous data flows are usually not supported by epidemiological studies and clinical trials. This could raise concerns about the quality of dynamic CPMs because routine data sources tend to have poorer data quality and higher levels of missingness than study datasets. A possible solution is to develop CPMs using high quality study data (e.g. from a prospective observational study) and dynamically revise and monitor them using the routine data. However, quality checks and comparisons between the datasets would still be required.

Throughout this article we have focused on the temporal aspect of miscalibration, however, miscalibration can also occur when CPMs are transferred to different settings and/or populations[68,79]. It may be possible to generalise the concept of dynamic CPMs to address this type of calibration variation in space. For example, dynamic approaches could be used to tailor a model to a local population or transfer a model to a different setting. This is an area that requires further research.

3.5 Conclusion

Static CPMs are at risk of being always one step behind on reality. Through an alliance between information technology and statistics, clinical prediction can be progressed to a continual service that minimizes the data-action latency in preventative medicine.

3.7 References

1. Murdoch, T. B. & Detsky, A. S. The Inevitable Application of Big Data. *Jama* **309**, 1351–1352 (2014).

2. Wang, Y., Kung, L. A. & Byrd, T. A. Big data analytics: Understanding its capabilities and potential benefits for healthcare organizations. *Technol. Forecast. Soc. Change* **126**, 3–13 (2018).

3. Raghupathi, W. & Raghupathi, V. Big data analytics in healthcare: promise and potential. *Heal. Inf. Sci. Syst.* **2**, 3 (2014).

4. Friedman, C. P., Wong, A. K. & Blumenthal, D. Achieving a Nationwide Learning Health System. *Public Law* **2**, 8–10 (2010).

5. Friedman, C. P., Rubin, J. C. & Sullivan, K. J. Toward an Information Infrastructure for Global Health Improvement. *Yearb. Med. Inform.* **26**, 16–23 (2017).

6. J., A. & I., B. Combining health data uses to ignite health system learning. *Methods Inf. Med.* **54**, 479–487 (2015).

7. Kent, P., Cancelliere, C., Boyle, E., Cassidy, J. D. & Kongsted, A. A conceptual framework for prognostic research. *BMC Med. Res. Methodol.* **7**, 1–13 (2020).

8. Hemingway, H. *et al.* Prognosis research strategy (PROGRESS) 1: A framework for researching clinical outcomes. *BMJ* **346**, 1–11 (2013).

9. Riley, R. D. *et al.* Prognosis Research Strategy (PROGRESS) 2: Prognostic Factor Research. *PLoS Med.* **10**, e1001380 (2013).

10. Steyerberg, E. *et al.* Prognosis research strategy (PROGRESS) 3: prognostic model research. *PLoS Med* **10**, e1001381 (2013).

11. Hingorani, A. D. *et al.* Prognosis research strategy (PROGRESS) 4: Stratified medicine research. *BMJ* **346**, 1–9 (2013).

12. Riley, R. D., Windt, D. Van Der & Moons, K. G. M. Prognosis Research in Health Care. *Progn. Res. Heal. Care* 1–11 (2019). doi:10.1093/med/9780198796619.001.0001

13. Steyerberg, E. W. *Clinical Prediction Models. Statistics for Biology and Health. 2nd edition*. (2019).

14. Hemingway, H. *et al.* Prognosis research strategy (PROGRESS) 1: A framework for researching clinical outcomes. *BMJ* **346**, 1–11 (2013).

15. Gray, L. J. *et al.* The Leicester Risk Assessment score for detecting undiagnosed Type 2 diabetes and impaired glucose regulation for use in a multiethnic UK setting. *Diabet. Med.* **27**, 887–895 (2010).

16. (NICE), N. I. for H. and C. E. Type 2 diabetes: prevention in people at high risk. *Clinical guidelines [PH38]* (2012). Available at: https://www.nice.org.uk/guidance/ph38.

17. McAllister, K. S. L. *et al.* A contemporary risk model for predicting 30-day mortality following percutaneous coronary intervention in England and Wales. *Int. J. Cardiol.* **210**, 125–132 (2016).

18. Damen, J. A. A. G. *et al.* Prediction models for cardiovascular disease risk in the general population: Systematic review. *BMJ* **353**, (2016).

19. D’Agostino, R. B. *et al.* General Cardiovascular Risk Profile for Use in Primary Care: The Framingham Heart Study. *Circulation* **117**, 743–753 (2008).

20. Hippisley-Cox, J. *et al.* Derivation and validation of QRISK, a new cardiovascular disease risk score for the United Kingdom: Prospective open cohort study. *Br. Med. J.* **335**, 136–141 (2007).

21. (NICE), N. I. for H. and C. E. Cardiovascular disease: risk assessment and reduction, including lipid modification. *NICE Guidel. [CG181]* (2014).

22. Hippisley-Cox, J., Coupland, C., Robson, J. & Brindle, P. Derivation, validation, and evaluation of a new QRISK model to estimate lifetime risk of cardiovascular disease: Cohort study using QResearch database. *Bmj* **342**, 93 (2011).

23. Bonnett, L. J., Snell, K. I. E., Collins, G. S. & Riley, R. D. Guide to presenting clinical prediction models for use in clinical settings. *BMJ* **365**, 1–8 (2019).

24. Riley, R. D. *et al.* Minimum sample size for developing a multivariable prediction model: PART II - binary and time-to-event outcomes. *Stat. Med.* **38**, 1276–1296 (2019).

25. Steyerberg, E. W. *Clinical prediction models: a practical approach to development, validation, and updating*. (2008).

26. Harrell Jr, F. E. *Regression Modeling Strategies : With Applications to Linear Models, Logistic and Ordinal Regression, and Survival Analysis in Springer Series in Statistics*. *Springer* (Springer, 2015).

27. Steyerberg, E. W. *et al.* Internal validation of predictive models: Efficiency of some procedures for logistic regression analysis. *J. Clin. Epidemiol.* **54**, 774–781 (2001).

28. Steyerberg, E. W. *et al.* Assessing the performance of prediction models: a framework for some traditional and novel measures. *Epidemiology* **21**, 128–138 (2013).

29. Alba, A. C. *et al.* Discrimination and calibration of clinical prediction models: Users’ guides to the medical literature. *JAMA - J. Am. Med. Assoc.* **318**, 1377–1384 (2017).

30. Van Calster, B. *et al.* Calibration: The Achilles heel of predictive analytics. *BMC Med.* **17**, 1–7 (2019).

31. Steyerberg, E. W. & Vergouwe, Y. Towards better clinical prediction models: Seven steps for development and an ABCD for validation. *Eur. Heart J.* **35**, 1925–1931 (2014).

32. Kramer, A. A. & Zimmerman, J. E. Assessing the calibration of mortality benchmarks in critical care: The Hosmer-Lemeshow test revisited. *Crit. Care Med.* **35**, 2052–2056 (2007).

33. Hickey, G. L. *et al.* Dynamic trends in cardiac surgery: Why the logistic euroscore is no longer suitable for contemporary cardiac surgery and implications for future risk models. *Eur. J. Cardio-thoracic Surg.* **43**, 1146–1152 (2013).

34. Davis, S. E., Lasko, T. A., Chen, G., Siew, E. D. & Matheny, M. E. Calibration drift in regression and machine learning models for acute kidney injury. *J. Am. Med. Informatics Assoc.* **24**, 1052–1061 (2017).

35. van Houwelingen, H. C. & Thorogood, J. Construction, validation and updating of a prognostic model for kidney graft survival. *Stat. Med.* **14**, 1999–2008 (1995).

36. Janssen, K. J. M., Moons, K. G. M., Kalkman, C. J., Grobbee, D. E. & Vergouwe, Y. Updating

methods improved the performance of a clinical prediction model in new patients. *J. Clin. Epidemiol.* **61**, 76–86 (2008).

37. Debray, T. P. A., Koffijberg, H., Vergouwe, Y., Moons, K. G. M. & Steyerberg, E. W. Aggregating published prediction models with individual participant data: A comparison of different approaches. *Stat. Med.* **31**, 2697–2712 (2012).

38. Martin, G. P., Mamas, M. A., Peek, N., Buchan, I. & Sperrin, M. Clinical prediction in defined populations: a simulation study investigating when and how to aggregate existing models. *BMC Med. Res. Methodol.* **17**, 1 (2017).

39. Martin, G. P., Mamas, M. A., Peek, N., Buchan, I. & Sperrin, M. A multiple-model generalisation of updating clinical prediction models. *Stat. Med.* **37**, 1343–1358 (2018).

40. Steyerberg, E. W., Borsboom, G. J. J. M., van Houwelingen, H. C., Eijkemans, M. J. C. & Habbema, J. D. F. Validation and updating of predictive logistic regression models: A study on sample size and shrinkage. *Stat. Med.* **23**, 2567–2586 (2004).

41. Toll, D. B., Janssen, K. J. M., Vergouwe, Y. & Moons, K. G. M. Validation, updating and impact of clinical prediction rules: A review. *J. Clin. Epidemiol.* **61**, 1085–1094 (2008).

42. Hippisley-Cox, J., Coupland, C. & Brindle, P. Development and validation of QRISK3 risk prediction algorithms to estimate future risk of cardiovascular disease : prospective cohort study. **2099**, 1–21 (2017).

43. Nashef, S. A. M. *et al.* European system for cardiac operative risk evaluation (EuroSCORE). *Eur. J. Cardio-thoracic Surg.* **16**, 9–13 (1999).

44. Nashef, S. A. M. *et al.* Euroscore II. *Eur. J. Cardio-thoracic Surg.* **41**, 734–745 (2012).

45. Raftery, A. E. & Ettler, P. Online Prediction Under Model Uncertainty via Dynamic Model Averaging : Application to a Cold Rolling Mill. **52**, 52–66 (2010).

46. Bull, L. M., Lunt, M., Martin, G. P., Hyrich, K. & Sergeant, J. C. Harnessing repeated measurements of predictor variables for clinical risk prediction: a review of existing methods. *Diagnostic Progn. Res.* **4**, (2020).

47. Hickey, G. L. *et al.* Dynamic prediction modeling approaches for cardiac surgery. *Circ. Cardiovasc. Qual. Outcomes* **6**, 649–658 (2013).

48. Mccormick, T. H., Raftery, A. E., Madigan, D. & Burd, R. S. Dynamic Logistic Regression and Dynamic Model Averaging for Binary Classification. *Biometrics* **68**, 23–30 (2012).

49. McCormick, T. H., Raftery, A. & Madigan, D. dma: Dynamic Model Averaging. (2018).

50. Five Year Forward View. (2014).

51. Salive, M. E. Multimorbidity in older adults. *Epidemiol. Rev.* **35**, 75–83 (2013).

52. Divo, M. J., Martinez, C. H. & Mannino, D. M. Ageing and the epidemiology of multimorbidity. *Eur. Respir. J.* **44**, 1055–1068 (2014).

53. Watkins, J. *et al.* Effects of health and social care spending constraints on mortality in England: a time trend analysis. *BMJ Open* **7**, e017722 (2017).

54. Abu-Hanna, A. & Lucas, P. J. F. Prognostic Models in Medicine. AI and Statistical Approaches.

Method Inf. Med **40**, 1–5 (2001).

55. Damen, J. A. A. G. *et al.* Prediction models for cardiovascular disease risk in the general population: systematic review. *Bmj* i2416 (2016). doi:10.1136/bmj.i2416

56. Siregar, S. *et al.* Improved Prediction by Dynamic Modeling. *Circ. Cardiovasc. Qual. Outcomes* **9**, 171–181 (2016).

57. Su, T.-L., Jaki, T., Hickey, G. L., Buchan, I. & Sperrin, M. A review of statistical updating methods for clinical prediction models. *Stat. Methods Med. Res.* 1–16 (2016). doi:10.1177/0962280215626466

58. van Houwelingen, H. C. & Thorogood, J. Construction, validation and updating of a prognostic model for kidney graft survival. *Stat. Med.* **14**, 1999–2008 (1995).

59. Draper, N. R., Nostrand, R. C. Van & Draper, N. R. Ridge Regression and James-Stein Estimation : Review and Comments Linked references are available on JSTOR for this article : Ridge Regression and James-Stein Estimation : Review and Comments. **21**, 451–466 (2016).

60. Copas, J. . Regression, Prediction and Shrinkage. *R. Stat. Soc.* **45**, 311–354 (1983).

61. Finkelman, B. S., French, B. & Kimmel, S. E. The prediction accuracy of dynamic mixed-effects models in clustered data. *BioData Min.* **9**, 5 (2016).

62. Fan, J. & Zhang, W. Statistical Methods with Varying Coefficient Models. *Stat Interface* **1**, 179–195 (2008).

63. Hoover, D. R., Rice, J. A., Wu, C. O. & Yang, L.-P. Nonparametric smoothing estimates of time-varying coefficient models with longitudinal data. *Biometrika* **85**, 809–822 (1998).

64. Madigan, D. & Raftery, A. E. Model Selection and Accounting for Model Uncertainty in Graphical Models Using Occam's Window. (1991).

65. Onorante, L. & Raftery, A. E. Dynamic model averaging in large model spaces using dynamic Occam's window. *Eur. Econ. Rev.* **81**, 2–14 (2016).

66. Ohata, T., Kaneko, M., Kuratani, T., Ueda, H. & Shimamura, K. Using the EuroSCORE to assess changes in the risk profiles of the patients undergoing coronary artery bypass grafting before and after the introduction of less invasive coronary surgery. *Ann. Thorac. Surg.* **80**, 131–135 (2005).

67. Ramsay, J. O., Wickham, H., Graves, S. & Hooker, G. fda: Functional Data Analysis. (2017).

68. Altman, D. G. & Royston, P. What do we mean by validating a prognistic model? *Stat. Med.* **19**, 453–473 (2000).

69. Justice, A. C., Covinsky, K. E. & Berlin, J. A. Assessing the generalizability of prognostic information. *Ann. Intern. Med.* **130**, 515–524 (1999).

70. Zimmerman, J. E., Kramer, A. A., McNair, D. S. & Malila, F. M. Acute Physiology and Chronic Health Evaluation (APACHE) IV: Hospital mortality assessment for today's critically ill patients. *Crit. Care Med.* **34**, 1297–1310 (2006).

71. Vergouwe, Y. *et al.* A closed testing procedure to select an appropriate method for updating prediction models. *Stat. Med.* **36**, 4529–4539 (2017).

72. Hafkamp-De Groen, E. *et al.* Predicting asthma in preschool children with asthma-like symptoms: Validating and updating the PIAMA risk score. *J. Allergy Clin. Immunol.* **132**, (2013).

73. Genders, T. S. S. *et al.* A clinical prediction rule for the diagnosis of coronary artery disease: Validation, updating, and extension. *Eur. Heart J.* **32**, 1316–1330 (2011).

74. Moons, K. G. M. *et al.* Risk prediction models: II. External validation, model updating, and impact assessment. *Heart* **98**, 691–698 (2012).

75. Hippisley-Cox, J., Coupland, C. & Brindle, P. Development and validation of QRISK3 risk prediction algorithms to estimate future risk of cardiovascular disease: Prospective cohort study. *BMJ* **357**, 1–21 (2017).

76. Martin, G. P., Sperrin, M. & Sotgiu, G. Performance of Prediction Models for Covid-19: The Caudine Forks of the External Validation. *Eur. Respir. J.* 2003728 (2020). doi:10.1183/13993003.03728-2020

77. Wynants, L. *et al.* Prediction models for diagnosis and prognosis of covid-19: Systematic review and critical appraisal. *BMJ* **369**, (2020).

78. Vergouwe, Y., Moons, K. G. M. & Steyerberg, E. W. External validity of risk models: Use of benchmark values to disentangle a case-mix effect from incorrect coefficients. *Am. J. Epidemiol.* **172**, 971–980 (2010).

79. Debray, T. P. A. *et al.* A guide to systematic review and meta-analysis of prediction model performance. *BMJ* **356**, (2017).

80. Luijken, K. *et al.* Changing predictor measurement procedures affected the performance of prediction models in clinical examples. *J. Clin. Epidemiol.* **119**, 7–18 (2020).

81. Riley, R. D. *et al.* External validation of clinical prediction models using big datasets from e-health records or IPD meta-analysis: Opportunities and challenges. *BMJ* **353**, 27–30 (2016).

82. Debray, T. P. A., Riley, R. D., Rovers, M. M., Reitsma, J. B. & Moons, K. G. M. Individual Participant Data (IPD) Meta-analyses of Diagnostic and Prognostic Modeling Studies: Guidance on Their Use. *PLoS Med.* **12**, 1–12 (2015).

83. Hippisley-Cox, J. *et al.* Predicting cardiovascular risk in England and Wales: Prospective derivation and validation of QRISK2. *Bmj* **336**, 1475–1482 (2008).

84. Jenkins, D. A., Sperrin, M., Martin, G. P. & Peek, N. Dynamic models to predict health outcomes: current status and methodological challenges. *Diagnostic Progn. Res.* **2**, 23 (2018).

85. Halabi, S. *et al.* Updated prognostic model for predicting overall survival in first-line chemotherapy for patients with metastatic castration-resistant prostate cancer. *J. Clin. Oncol.* **32**, 671–677 (2014).

86. Dawid, A. Present Position and Potential Developments : Some Personal Views : Statistical Theory : The Prequential Approach Author (s): A . P . Dawid Source : Journal of the Royal Statistical Society . Series A (General), Vol . 147 , No . 2 , The 150th Annivers. *J. R. Stat. Soc. Ser. A* **147**, 278–292 (1984).

87. Lenert, M. C., Matheny, M. E. & Walsh, C. G. Prognostic models will be victims of their own success, unless…. *J. Am. Med. Inform. Assoc.* **26**, 1645–1650 (2019).

88. Adibi, A., Sadatsafavi, M. & Ioannidis, J. P. A. Validation and Utility Testing of Clinical Prediction

Models. *JAMA* **2004**, (2020).

89. Booth, S., Riley, R. D., Ensor, J., Lambert, P. C. & Rutherford, M. J. Temporal recalibration for improving prognostic model development and risk predictions in settings where survival is improving over time. *Int. J. Epidemiol.* 1–10 (2020). doi:10.1093/ije/dyaa030

90. Jenkins, D. A. *et al.* Continual updating and monitoring of clinical prediction models: time for dynamic prediction systems? *Diagnostic Progn. Res.* **5**, 1–7 (2021).

91. Mccormick, T. H. *et al.* Dynamic Logistic Regression and Dynamic Model Averaging for Binary Classification. 23–30 (2012). doi:10.1111/j.1541-0420.2011.01645.x

92. R Core Team. R: A language and environment for statistical computing. R Foundation for Statistical. (2014).

93. Martin, G. P., Riley, R. D., Collins, G. S. & Sperrin, M. Developing clinical prediction models when adhering to minimum sample size recommendations: The importance of quantifying bootstrap variability in tuning parameters and predictive performance. *Stat. Methods Med. Res.* **30**, 2545–2561 (2021).

94. Sperrin, M., Jenkins, D., Martin, G. P. & Peek, N. Explicit causal reasoning is needed to prevent prognostic models being victims of their own success. *J. Am. Med. Informatics Assoc.* **26**, 1675–1676 (2019).

95. Hickey, G. L. *et al.* Clinical registries: Governance, management, analysis and applications. *Eur. J. Cardio-thoracic Surg.* **44**, 605–614 (2013).

96. Moons, K. G. M. *et al.* Risk prediction models: II. External validation, model updating, and impact assessment. *Heart* **98**, 691–698 (2012).

97. Davis, S. E. *et al.* A nonparametric updating method to correct clinical prediction model drift. *J. Am. Med. Informatics Assoc.* **26**, 1448–1457 (2019).

98. Minne, L. *et al.* Statistical process control for monitoring standardized mortality ratios of a classification tree model. *Methods Inf. Med.* **51**, 353–358 (2012).

99. Statistics, M. SUMS OF NONINDEPENDENT BERNOULLI RANDOM VARIABLES Author (s): Jose M . Gonzalez-Barrios Source : Brazilian Journal of Probability and Statistics , JUNE 1998 , Vol . 12 , No . 1 (JUNE Published by : Institute of Mathematical Statistics Stable URL : http. **12**, 55–64 (1998).

100. Koetsier, A., De Keizer, N. F., De Jonge, E., Cook, D. A. & Peek, N. Performance of risk-adjusted control charts to monitor in-hospital mortality of intensive care unit patients: A simulation study. *Crit. Care Med.* **40**, 1799–1807 (2012).

101. Team, R. core. R: A language and environment for statistical computing. *R Foundation for Statistical Computing, Vienna, Austria.* (2021). Available at: https://www.r-project.org/.

102. Albert, A. A. *et al.* On-line variable live-adjusted displays with internal and external risk-adjusted mortalities. A valuable method for benchmarking and early detection of unfavourable trends in cardiac surgery. *Eur. J. Cardio-thoracic Surg.* **25**, 312–319 (2004).

103. Lovegrove, J., Valencia, O., Treasure, T., Sherlaw-Johnson, C. & Gallivan, S. Monitoring the results of cardiac surgery by variable life-adjusted display. *Lancet* **350**, 1128–1130 (1997).

104. Poloniecki, J., Valencia, O. & Littlejohns, P. Correction: Cumulative risk adjusted mortality chart

for detecting changes in death rate: Observational study of heart surgery (British Medical Journal (1998) (1697-700)). *Br. Med. J.* **316**, 1947 (1998).

105. Minne, L. *et al.* Effect of changes over time in the performance of a customized SAPS-II model on the quality of care assessment. *Intensive Care Med.* **38**, 40–46 (2012).

106. Shi, X., Gallagher, C., Lund, R. & Killick, R. A Comparison of Single and Multiple Changepoint Techniques for Time Series Data. 1–29 (2021). doi:10.1016/j.csda.2022.107433

107. Wittenberg, P., Gan, F. F. & Knoth, S. A simple signaling rule for variable life-adjusted display derived from an equivalent risk-adjusted CUSUM chart. *Stat. Med.* **37**, 2455–2473 (2018).

108. Davis, S. E., Greevy, R. A., Lasko, T. A., Walsh, C. G. & Matheny, M. E. Detection of calibration drift in clinical prediction models to inform model updating. *J. Biomed. Inform.* **112**, 103611 (2020).

109. Riley, R. D. *et al.* Minimum sample size for external validation of a clinical prediction model with a binary outcome. *Stat. Med.* **40**, 4230–4251 (2021).

110. Steyerberg, E. W., Borsboom, G. J. J. M., van Houwelingen, H. C., Eijkemans, M. J. C. & Habbema, J. D. F. Validation and updating of predictive logistic regression models: A study on sample size and shrinkage. *Stat. Med.* **23**, 2567–2586 (2004).

111. Grigg, O. A., Farewell, V. T. & Spiegelhalter, D. J. Use of risk-adjusted CUSUM and RSPRT charts for monitoring in medical contexts. *Stat. Methods Med. Res.* **12**, 147–170 (2003).

112. Pagel, C. *et al.* Real time monitoring of risk-adjusted paediatric cardiac surgery outcomes using variable life-adjusted display: Implementation in three UK centres. *Heart* **99**, 1445–1450 (2013).

113. Barrett, J. & Su, L. Dynamic predictions using flexible joint models of longitudinal and time-to-event data. *Stat. Med.* **36**, 1447–1460 (2017).

114. van Houwelingen, H., & Putter, H. *Dynamic Prediction in Clinical Survival Analysis*. (2012). doi:https://doi.org/10.1201/b11311

Chapter 4 Development and validation of clinical prediction models in the presence of temporal trends: A simulation study comparing static and dynamic models

David A Jenkins, Matthew Sperrin, Niels Peek, Camilla Sammut-Powell, Thomas P.A. Debray, Mamas Mamas, and Glen P Martin

In preparation for submission

4.1 Abstract

4.1.1 Background

Clinical prediction models (CPMs) are used across the healthcare system, specifically in preventative medicine and to support clinical decision making. However, healthcare and populations change through time. Most CPMs are developed using methods where the model coefficients remain static with time, thereby frequently leading to worsening predictive performance over time, due to the temporal changes in healthcare and the population. Dynamic prediction models, CPMs developed using methods that allow coefficients to be a function of time, have been proposed as a solution. However, the improvement in predictive performance of these methods, if any, over time-invariant CPMs has received little attention and varying coefficient models have yet to be evaluated in this context. Therefore, we aimed to assess performance of traditional and dynamic CPMs under a variety of temporal trend scenarios.

4.1.2 Methods

We simulated continuous and binary outcome data under a variety of scenarios where the data-generating predictor-outcome associations and intercept were changing over time. Traditional regression models and two dynamic modelling approaches: varying coefficient and Bayesian continuous updating models, were fitted to the data and predictive performance was assessed. We also applied the methods, and compared the performance of the resulting models, in a cardiac dataset.

4.1.3 Results

The Bayesian continuous updating model either outperforming or performed as well as the time-invariant model. In the cardiac data, the Bayesian model had a calibration-in-the-large of -0.018 (95% confidence interval: -0.039 - 0.002) and was the only calibration-in-the-large confidence interval to include 1. However, the varying coefficient model outperformed the other models in the simulation study but performed similar to the time-invariant model in the real-world example. Both the simulation study and real-world example show that dynamic prediction models retain accuracy over time.

4.1.4 Conclusion

How we acknowledge time in predictive modelling impacts model performance. Dynamic models offer a solution to model coefficient drift and also perform as well as time-invariant CPMs when there are little to no temporal changes.

4.2 Introduction

The desire to move health systems away from treatment/management to preventative medicine has resulted in the pervasiveness of clinical prediction models (CPMs) throughout healthcare. CPMs are statistical models/algorithms that compute the risk of an outcome given a set of patient characteristics and are typically developed using multivariable regression[25]. Classically, the methods used to develop CPMs mean that, once developed, the model coefficients remain fixed for the duration of its use (i.e. the coefficients are not a function of time). Thus, CPMs are often implemented under the implicit assumption that the processes that generated the data are constant, and evidence can be implemented into practice indefinitely. However, in practice healthcare is constantly evolving and the demographics of the population, disease prevalence and health policies may also change over time. As a result, the predictive performance of models that remain fixed with time can decrease and therefore not provide accurate risk estimates[33]. Consequently, it is not uncommon for the agreement between the observed and predicted event rates (i.e. calibration) to worsen over time. This is known as calibration drift[34] and is one of the major pitfalls in using CPMs in practice. Also, when the effect of multiple covariates change over time, discrimination can diminish.

Discrete model updating[36], using new batches of data at arbitrary time points to recalibrate the model, is currently used to tackle this and models such as QRISK and EuroSCORE have been updated and successfully deployed in practice[42,44]. This is where the existing model is revised using more recent data and following the update, model coefficients remain fixed. However, this is often done at arbitrary times, often creating a large data-action latency between calibration drift occurring and action being taken to correct it. This has been observed with EuroSCORE[33]. Jenkins et al also discuss that it does not stop drift occurring again in the future and it is only a temporary fix[90].

Dynamic models are a collection of analytical methods that allow model parameters to evolve over time and continually provide updated information as new data arrive[84]. Specifically, these methods acknowledge the time of each data point and are formulated such that the parameter estimates are not constrained to remain fixed over time. Therefore, these models have the potential to address diminishing performance of CPMs over time. Jenkins et al[84] recently conducted a review of these methods and found little use of dynamic models in healthcare. They also identified a method, varying coefficient modelling, yet to be used in this setting. Before application of these methods in practice, research is needed to assess the potential utility of varying coefficient models in this area and to explore the properties of the dynamic methods, under a range of temporal changes, compared to time-invariant CPMs.

This study aims to compare the predictive performance of time-invariant CPMs and dynamic prediction models under a range of temporal changes through the use of simulation and a real-world example in cardiovascular disease. Therefore, this study seeks to investigate the added value (if any) in predictive performance of dynamic modelling methods compared to time-invariant CPMs, and when dynamic models would be preferred to time-invariant CPMs.

The paper is structured as follows. In Section 4.3 we introduce the simulation study, including the aims, data-generating mechanisms, and methods. In Section 4.4, we present the results of our simulations. In Sections 4.5, we compare the different models using real data from a study on percutaneous coronary intervention. Finally, we conclude this paper in Section 4.6 with a discussion.

4.3 Simulation study

4.3.1 Aims

The primary aim of this simulation study is to assess and compare the predictive performance of dynamic modelling approaches and time-invariant CPMs under a variety of temporal change scenarios in which baseline-risk and predictor-outcome associations may change over time.

4.3.2 Data-generating mechanisms

We generate two continuous predictor variables observed at discrete times $\mathrm{t} = (1,2,\dots,\mathrm{T})$ derived from a standard normal distribution, thus

$$X_{j,t} \sim Normal(0,1) \text{ for } j = 1,2$$

At each time t, we allow for there to be multiple individuals (batches of data) and we denote the number of individuals at time t to be n_t; we also allow $n_t = 1$ such that data are arriving in real-time. We chose to simulate data such that n_t =1, 10, 25 or 100 data points at each time t. In this study we only consider situations where n_t is fixed for all t, though in practice this could vary at each time point. If there are multiple individuals, $X_{j,t}$ is a vector of size n_t. In either case, we assume that each individual had either a continuous or binary outcome, y_t, (depending on simulation scenario), which is observed for an individual a short time after t, which we index by t for convenience. We generated these outcomes such that:

$$g\big(E(Y_t)\big) = \beta^0{}_0(t) + \beta^0{}_1(t)x_{1,t} + \beta^0{}_2 x_{2,t}, \tag{1}$$

for a suitable link-function g, and where $\beta^0{}_0(t)$ is the time-dependent intercept and $\beta^0{}_1(t)$ is the time-dependent regression coefficient for the predictor, x_1, at time t. $\beta^0{}_2$ was chosen to remain fixed across all simulations, $\beta^0{}_2 = -1$, as this enables the risk order to change over time and hence the discrimination. To clarify, we consider situations where each individual is observed only once in the dataset; that is, we do not consider longitudinal data for a given individual.

We assume a temporal development and internal validation process, where we define the development data (available to derive the models) to include all observations made between time $t = 1 \; and \; t = d$ and the validation data (to test predictive performance) to include all observations made between time $t = d + 1 \; and \; t = T \; where \; d < T$. For all scenarios we chose $d = 365$ and $T = 730$. Hence, the sample sizes for model development were 365, 3650, 9125 and 36500 respectively. Note, observing 10 data points at each timepoint is also equivalent to observing one observation at each time t and updating every 10th time point over a time period ten times longer. Therefore, we did not choose to vary time across simulations.

Across all simulations $\beta^0{}_0(t) \; and \; \beta^0{}_1(t)$ were varied in three different scenarios such that the underlying functions were:

1. Fixed with no change over time (i.e. $\beta^0{}_0(t) = \beta^0{}_0$ or $\beta^0{}_1(t) = \beta^0{}_1$)
2. Linearly increasing over time such that,

$$\beta^0{}_p(t) = \beta^0{}_p(1) + \alpha_t, for \; p = 0,1 \; where \; \alpha_t = \alpha \; {}^{t}/_{T}$$

3. Static until a given time s, where $s < T$ when a step/sudden change occurred. Specifically,

$$\mathrm{β}^0{}_p(t) = \begin{cases} \mathrm{β}^0{}_p(1) & if\ t < s \\ \mathrm{β}^0{}_p(1) + \alpha & otherwise \end{cases}$$

For $p = 0{,}1$.

In the above scenarios, 2 and 3, α controls the magnitude of the change in the coefficients over time. In this study, we considered five values of α which were, 0, 0.2, 0.5, 1 and 2. For the linear model, this corresponds to a standard deviation change of 0, 1, 2.5, 5 and 10 for y_t between $t = 1$ and $t = T$. Additionally, for the choice of s, the time of a step change, we chose 3 times; 1) midway in the development data ($t = 183$), 2) three quarters of the way through the development data ($t = 274$) and 3) early in the validation data ($t = 395$).

The outcome was simulated for each set of possible combinations of the fixed, linear and step change scenarios (and for each $p = 0{,}1$), under the data generating model shown in equation 1. In simulation scenarios where we simulate a continuous outcome, Y_t was generated as

$$Y_t \sim N(\mathrm{β}^0{}_0(t) + \mathrm{β}^0{}_1(t)x_{1,t} + \mathrm{β}^0{}_2 x_{2,t}, 0.2^2)$$

We choose 0.2 as the error standard deviation to ensure random variation between iterations. Without loss of generality, we fixed the standard error of the outcome generation to be 0.2; this was chosen to fix the signal-noise ratio across scenarios. A higher standard deviation would simply change the absolute value of resulting performance, but this would be the same across all methods such that conclusions would not change. For simulation scenarios where we simulate a binary outcome, y_t was simulated as follows,

$$Y_t \sim Binomial(\frac{\exp{(LP)}}{1 + \exp{(LP)}})$$

$$where\ LP = \mathrm{β}^0{}_0(t) + \mathrm{β}^0{}_1(t)x_{1,t} + \mathrm{β}^0{}_2 x_{2,t}$$

Sample size needs to be considered during model development and Riley et al[24] recently proposed an approach for sample size in both linear and logistic CPMs. Rearranging the calculation, for the logistic model, we calculated that for a sample size of 365 when two parameters are to be considered in the model, the prevalence of the outcome needs to be less than 0.39 or above 0.61. Taking the prevalence of 0.39, the linear predictor therefore needs to be below -0.447 because $E(y) = \frac{\exp{(LP)}}{1+\exp{(LP)}}$, so solving $0.39 = \frac{\exp{(LP)}}{1+\exp{(LP)}}$ gives $LP = -0.447$. As a result, we chose $\beta_0(1) = -0.5$. We chose arbitrary values for the other betas such that $\beta_1(1) = 1\ and\ \beta_2 = 1$. The minimum sample size required for the linear model with the same parameter choices was 237, less than the 365 in the smallest simulation, so the same parameter choices were applied in the linear model.

Table 4.1 displays all parameter choices and for each unique combination of parameters we simulated 1000 datasets. The results were then averaged over the 1000 simulations for each scenario. A total of 1600 unique parameter choices (simulation scenarios) were considered but we focus on a subset of them in the results. Specifically, we present the results from the linear and logistic models where we observe a single observation at each time point from scenarios when alpha = 0, 0.5 and 1 and both $\mathrm{β}^0{}_0$ and $\mathrm{β}^0{}_1$ have the same rate of change over time. We also present the results for one of the step changes, where t=274, but the focus in the results is on linear changes as they are more common in clinical practice. For example, prevalence of a disease is often changing gradually rather than suddenly.

Table 4.1 - Parameter choices for each simulation

Parameter	**Model**	
Outcome	Linear	Logistic
Beta change (α)	0, 0.2, 0.5, 1, 2	0, 0.2, 0.5, 1, 2
How beta changes across all combinations of p=0,1	Linear Step change at t=s	Linear Step change at t=s
Error SD	0.2	NA
Observations at each timepoint	1, 10, 25, 100	1, 10, 25, 100
Time of step change in beta (s)	183, 274, 395	183, 274, 395
$\boldsymbol{\beta_0(t=1)}$	-0.5	-0.5
$\boldsymbol{\beta_1(t=1)}$	1	1
$\boldsymbol{\beta_2}$	1	1

4.3.3 Modelling approaches

This study compared three modelling approaches, a time-fixed CPM approach, to represent how models are currently derived and used in practice, as well as two dynamic modelling approaches: Bayesian updating and varying coefficient modelling.

4.3.3.1 Linear and logistic regression

Within each simulation, we fit either a linear or logistic model (depending on the outcome being simulated for that simulation scenario), where the coefficients are derived using the complete development data set, with no acknowledgment of time included in the model. The coefficients, estimated through maximum likelihood estimation, therefore remain fixed throughout the validation data. Specifically, we fit the following model to the development data,

$$g(E(y)) = \beta_0 + \beta_1 x_1 + \beta_2 x_2$$

4.3.3.2 Varying coefficient model

The varying coefficient model estimates the betas as smooth functions of time[62,63]. This approach uses all the development data set to estimate beta as a function of time. In the simulation we consider two varying coefficient models. The first, and most simplistic, is where only the intercept is dependent on time and we assume the functional form is linear. Specifically,

$$g(E(y_t)) = \beta_0(\mathrm{t}) + \beta_1 x_1 + \beta_2 x_2 = \beta_0 + \beta_3 t + \beta_1 x_1 + \beta_2 x_2,$$

Where $\beta_0(t) = \beta_0 + \beta_3 t$. The second varying coefficient model considers all coefficients as linear functions of time such that,

$$g(E(y_t)) = \beta_0(\mathrm{t}) + \beta_1(t) x_1 + \beta_2(t) x_2$$

$$= \beta_0 + \beta_3 t + \beta_1 x_1 + \beta_4 t x_1 + \beta_2 x_2 + \beta_5 t x_2,$$

Where,

$$\beta_0(t) = \beta_0 + \beta_3 t$$

$$\beta_1(t) = \beta_1 + \beta_4 t$$

$$\beta_2(t) = \beta_2 + \beta_5 t$$

When predicting using these models, one may extrapolate the functions to the time of the prediction or fix time to be the last time of the development data, t=d. The latter is similar to the approach described by Booth et al[89] for survival models. We consider both in the simulation and therefore have 4 varying-coefficient models.

4.3.3.3 Bayesian updating model

The Bayesian updating model continually updates at each time point where the information obtained from past data is used as prior information and combined with the new data to obtain updated estimates. Thus, the updating (posterior) equation is proportional to the product of the likelihood (at time t) and the prior (prediction equation at time t-1)[45,91].

$$\mathrm{p}(\,\beta_K(t)\mid Y_t) \propto \mathrm{p}(\,\beta_K(t)\mid\ Y_{t-1})\mathrm{p}(\,y_t|\,\beta_K(t)\,) \propto \text{Prior x Likelihood}$$

Where the prediction equation (Prior) is obtained through Kalman filtering by supposing $\mathrm{p}(\beta(\mathrm{t}-1)|\ \mathrm{Y}_{\mathrm{t}-1}) \sim \mathrm{N}(\hat{\beta}(\mathrm{t}-1), \varphi_{t-1})$, where $\mathrm{Y}_{\mathrm{t}-1} = \{\mathrm{y}_1, \dots, \mathrm{y}_{\mathrm{t}-1}\}$. This results in the prediction equation

$$\mathrm{p}(\beta_K(t)\ |Y_{t-1}) \sim \mathrm{N}\big(\hat{\beta}_K(\mathrm{t}-1), \mathrm{R}_\mathrm{t}\big);\ \mathrm{R}_\mathrm{t} = {}^{\varphi_{t-1}}/_{\lambda}$$

The forgetting factor, λ, down-weights (or decays) historical data so they have less influence/weight than new data and was used to down weighted at a constant rate over time. Previous research suggests λ should be between 0.9 and 1[47,48]. However, this does not guarantee the sample size at each time, *t*, will be sufficient. Minimum sample size criteria have recently been developed for clinical prediction models. Therefore, λ should be chosen to ensure adequate sample size. Forgetting is comparable to windowing where the effective window size is $h = \frac{1}{1-\lambda}$ and the data from the last *h* time points are used for estimation and equally weighted. Therefore, when updating at each observation, *h* should be at least the minimum required sample size, *minsamp*, based on Riley et al[24], such that $h\ \geq\ minsamp$. Hence, we suggest λ should be

$$^{minsamp - 1}/_{minsamp} \leq \lambda \leq 1$$

For each scenario in our study, we use the minimum lambda to allow the model more flexibility to change over time. Hence, if we have a linear model with 2 candidate variables to consider in the model then the minimum require sample size is 237. If we then observe one observation at each timepoint, λ would be $^{237-1}/_{237} = 0.9958$. If multiple observations are observed at each timepoint then λ needs to be adapted such that

$$\lambda = {}^{\frac{minsamp}{obs}-1}\Big/_{\frac{minsamp}{obs}}$$

Where *obs* is the number of observations at each time, t. Hence with 5 observations at each update and a minimum sample size of 237, $\lambda = 0.979$.

We include two Bayesian updating models in our study, the first updates at each time *t* in the development data set, and the coefficients remain fixed during the validation data. The second model continues to update during the validation data.

4.3.4 Performance measure

The performance measure of interest was predictive accuracy. To compare models in each scenario, we calculated calibration slope, calibration-in-the-large and mean square error (MSE). For the logistic models, we also computed discrimination.

Each performance measure was computed over the complete validation data to provide an average measure of performance for each model over the validation data. This was chosen as it represents current practice for validation where performance is evaluated over the complete validation. Currently it is challenging to validate the continuously updating Bayesian model in clinical practice but prequential testing[86,90] is a solution whereby we evaluate the model at time *t* using the data up to time *t-1*. Here we illustrate its potential use in practice.

4.3.5 Coding and execution

We used R[92] version 3.6.2 to generate the data, fit all the models and run the simulations. The simulation code can be found at https://github.com/David-A-Jenkins/Thesis. The Bayesian updating model was implemented using a modified version of the DMA package[49]. The current package only enabled updating at each single observation. We extended the package to allow updates to occur with more than one observation at a given time point and to only update at given times. The modified code can be found on github (https://github.com/David-A-Jenkins/Thesis).

4.4 Results

Among the 1600 simulated scenarios, we select a subset of them to focus on here. Specifically, we select a subset of choices for the drift in $\beta_0(t)$ and $\beta_1(t)$ and present the results (in section 4.4.1) where we observe one observation at each time point. Section 4.4.2 provides an overview of the results for the other scenarios not presented in section 4.4.1.

4.4.1 Predictive performance of selected scenarios

The predictive performance of the selected scenarios for the continuous and logistic outcomes are presented in figures 4.1 and 4.2. For the continuous outcomes, the varying coefficient models had the best predictive performance for all measures. For the linear change over time the varying coefficient model extrapolated using prediction time in the validation data had a calibration-in-the-large closest to 0, a calibration slope close to 1 and lowest mean-squared error. However, note that in this scenario the model is correctly specified with respect to the data generating model. When there is a step change in the betas of the data-generating model, these models no longer perform ‘perfect’. The continuously updating Bayesian model calibration-in-the-large was closer to 0 than the linear model and the Bayesian model which stopped updating at the end of the development data. The continuously updated model, on average, also had a calibration slope closer to 1 and mean-squared error closer to 0 compared to the linear model and Bayesian model updated only in development data. However, the varying coefficient models outperformed all of the models with respect to calibration-in-the-large and mean-squared error. For the calibration slope the varying coefficient models where all betas were function of time were closest to one out of all the models and the varying coefficient models with only the intercept as a function of time had a calibration slope similar to the Bayesian models.

For the logistic outcome, little differences were seen between the predictive performance of the models (figure 4. 2) except for the confidence intervals. The performance measure confidence intervals of the varying coefficient models were much wider than the other models and the continuously updated Bayesian model had the narrowest confidence interval for all performance measures in each scenario (figure 4.2). Little differences were seen between models in discrimination and mean-squared error. The largest differences between models were observed in the calibration-in-the-large and the varying coefficient models calibration-in-the-large were closest to zero. Specifically, the varying coefficient model with all betas as functions of time and extrapolated in the validation data had the best calibration-in-the-large, but it also had the widest confidence interval of any model. The time-invariant model was the worst performing model, except when there was no change over time and all models performed very well.

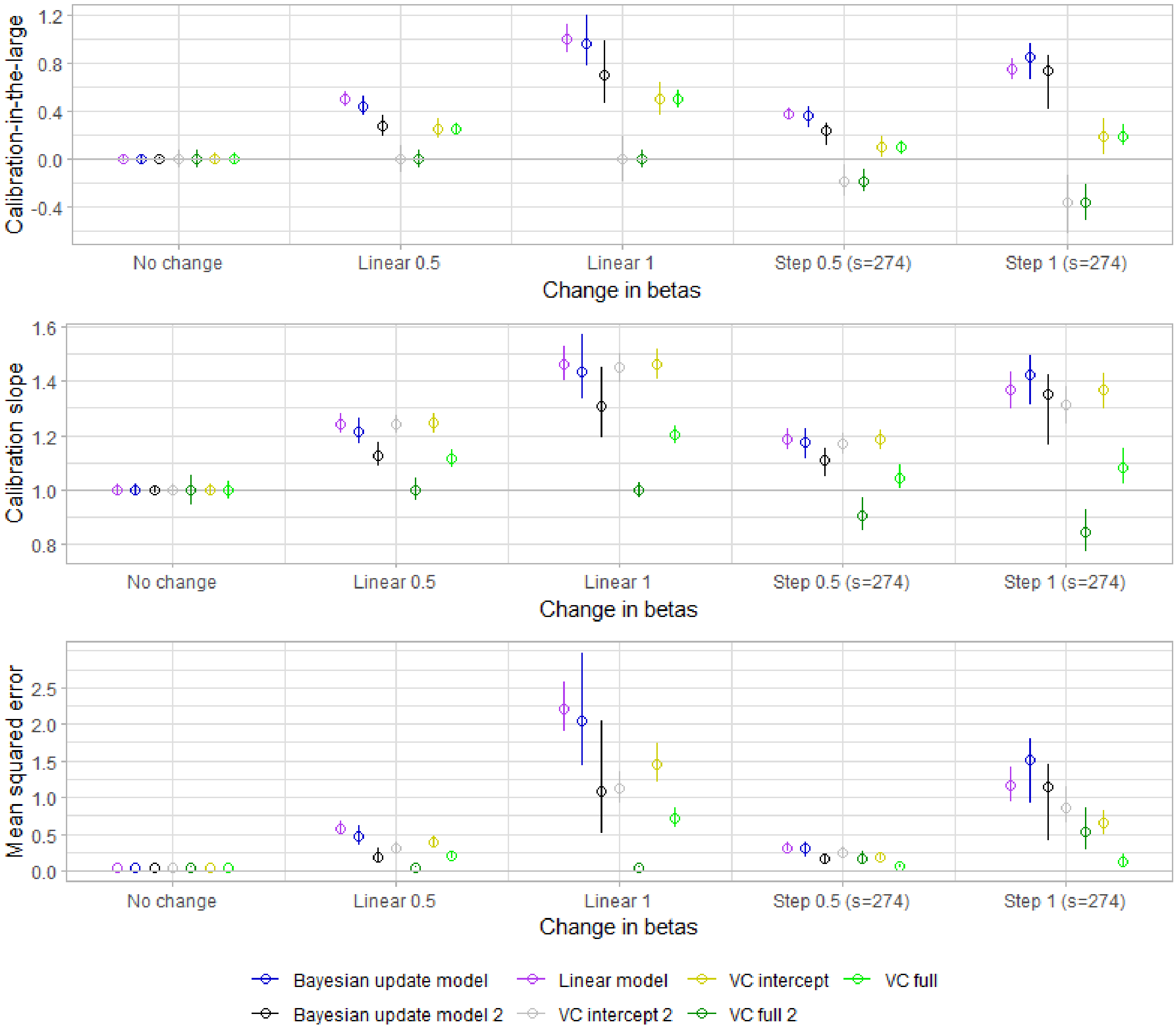


Figure 4.1 - Predictive performance averaged over the full validation data for each model separately for each of the selected scenarios for continuous outcomes. Model 2 refers to the model updating in the validation data (for the Bayesian model) and the model extrapolated to the validation time (varying coefficient model). The circles represent median values across the 1000 iterations and vertical lines represent the 95% quantiles.

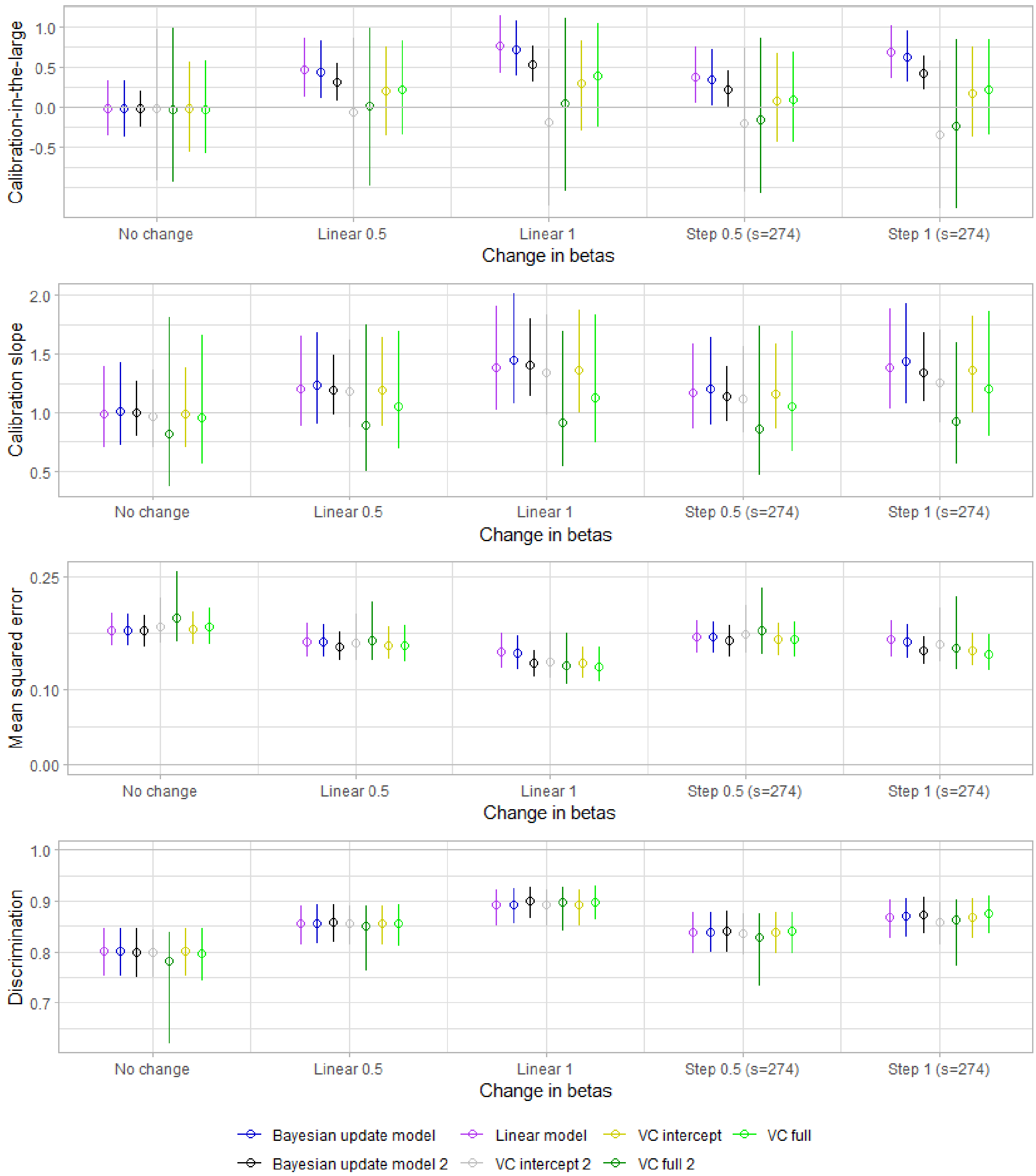


Figure 4.2 - Predictive performance averaged over the full validation data for each model separately for each of the selected scenarios for binary outcomes. Model 2 refers to the model updating in the validation data (for the Bayesian model) and the model extrapolated to the validation time (for the varying coefficient model). The circles represent median values across the 1000 iterations and vertical lines represent the 95% quantiles.

4.4.2 Overview of predictive performance for the other scenarios

For each combination of parameter choices, the results were similar when multiple observations were observed at each time point. The number of observations impacted the confidence intervals but did not impact the conclusions or the ordering of which models performed best for each performance measure. For the other combinations of drift that are not shown above, when both $\beta_0(t)$ and $\beta_1(t)$ are changing over time in the same manner the ordering of the models does not change. Only the magnitude of the performance value is affected by the values of drift chosen. When $\beta_1(t)$ changed over time but $\beta_0(t)$ was static, the calibration-in-the-large for all models was approximately zero (supplementary figure S4.1) but the calibration slope values were not effected, compared to the results above. The MSE values were also lower than the above results, but the ordering of the models remained the same. Conversely, when $\beta_1(t)$ was static and $\beta_0(t)$ was not static, the calibration slope was always approximately 1 except for the varying coefficient models extrapolated using prediction time in the validation data. These models were misscalibrated and had calibration slope below 1. The larger the change in $\beta_0(t)$ the more miscalibrated these models were (supplementary figure S4.2). Also, when $\beta_1(t)$ was static the calibration-in-the-large was similar to the above results, except for narrower confidence intervals, and the ordering of the best performing models did not change. The MSEs were also, on average, lower but the ordering of the models remained the same.

Hence, changes in $\beta_0(t)$ over time impacted the magnitude of the calibration-in-the-large and MSE for all models but only impacted the calibration slope results for the varying coefficient model extracted using validation data. Changes in $\beta_1(t)$ over time impacted the magnitude of the calibration slope and MSE values for all the models but it did not impact the ordering of which models had the best calibration. Ordering of the models for each performance value was the same for all parameter choices except for when $\beta_0(t)$ or $\beta_1(t)$ was static

4.5 Empirical Study of Percutaneous Coronary Intervention

Percutaneous coronary intervention (PCI) is a procedure that uses a catheter to insert a stent to open up blood vessels. It is used to improve blood flow and used to reduce symptoms of coronary heart disease or reduce damage after a heart attack.

Since 2005 The British Cardiovascular Intervention Society (BCIS), has incorporated patient data into annual audits for PCI in the UK and developed a registry collecting data on all PCI procedures. The audits are used for benchmarking and use 30-day mortality following PCI as one of the outcomes. In 2016, McAllister et al[17] developed a CPM to predict the risk of 30-day mortality after PCI, which is currently used in clinical practice for said benchmarking. The model was developed using a logistic regression model within the BCIS registry data from 2007 to 2011. Model validation was performed using the 2012 BCIS data and showed the model was well calibrated. This registry now has available data on all PCI procedures until March 2018 but now only records in-hospital mortality. Here we apply the models described in the simulation study to this cardiovascular data set and follow the data cleaning and model building outlined by McAllister to compare the performance of each model in the new data.

4.5.1 Study Population and BCIS Registry

The BCIS registry collects data on every PCI procedure performed in the UK through a Web-based interface provided by the National Institute of Cardiovascular Outcomes Research. This study included data on all PCI procedures between January 2007 and March 2018. The registry includes 113 variables which contain information on patient baseline demographics, risk factors for intervention, procedural details and patient outcomes, including discharge status. Time of each PCI procedure was recorded as the date and time the procedure took place.

4.5.2 Data pre-processing

Data pre-processing steps and exclusion criteria were applied to match those described by McAllister et al[17] as closely as possible. Individuals were excluded if they: 1) were not between the ages of 18 and 100 years, 2) were ventilated pre-operatively, 3) had fractional flow ultrasound without PCI, or 4) had missing outcome data. An indication-urgency variable was created as a five-group classification by combining clinical indication and urgency of the procedure to avoid collinearity. Indication group 1 was used as the reference group and were individuals with stable condition attending for a scheduled procedure. Groups increased in severity with group 5 including individuals with acute coronary syndrome who had PCI as an emergency procedure. See McAllister et al. paper for specific details[17]. The renal function, creatinine level and dialysis variables were combined to create one renal function variable as per McAllister et al[17]. Individuals who were on dialysis were assigned to the 'Renal (dialysis)' group. If they were not on dialysis but creatinine levels were above 200 µmol/l individuals were assigned to the 'Renal (creatinine)' group. All other patients were assigned to the no renal impairment group, including those with functioning transplants.

4.5.3 Statistical Analysis

Four models were developed and validated in the data. All models included mostly the same predictors that were in the model developed by McAlister et al[17]. However, renal function and the indication-shock interaction were excluded from the models due to their rarity (<1%) and thus issues with estimation in the data. The first model was a Logistic regression fitted to the BCIS data collected between 2007 and 2011. The second was a logistic regression model updated yearly. Again, this used the 2007 to 2011 data to develop the model but was then subsequently recalibrated at the start of each year between 2012 and 2018 by fitting a logistic regression model and using the linear predictor of the model as the only covariate. Hence, the coefficients remained fixed throughout each year, but are revised between years. This was chosen as it represents how some models in clinical practice, for example QRISK, are currently updated. The third model was a continuously updated Bayesian dynamic model, updated at each new observation. We applied the sample size calculation by Riley et al[24]. and derived the minimum forgetting factor that would ensure adequate sample size (as described in section 4.3.3.3 earlier). Finally, a varying coefficient model developed using the data between 2007 and 2011. Only the intercept term was dependent on time and the functional form was assumed to be linear.

The models were then validated in the data collected between January 2012 and March 2018. For each model, we calculated the calibration-in-the-large, calibration slope, discrimination and the mean-squared error. We computed the validation measures over the full validation data and for each month between January 2012 and March 2018. The mean and standard deviation of the monthly performance measures were also calculated for each model. Prequential testing was used to validate the continuously updated dynamic model.

All analyses are performed using R (version 3.6.2) and the dynamic models were fitted using functions adapted from the dma package[49].

4.5.4 PCI Results

After data cleaning, the final data comprised of 1,038,978 procedures and a total of 17,057 (1.64%) patients died in hospital following PCI procedure. 434,561 procedures were undertaken before 2012 and 5,368 (1.24%) of those died in hospital following PCI procedure. The minimum sample size required for model development was approximately 3500, resulting in a minimum forgetting factor of 0.9998.

4.5.4.1 Model performance

Table 4.2 displays the validation results for each model in the data between 2012 and 2018. The Bayesian updating model had a calibration-in-the-large closest to 0, -0.018 (95% CI: -0.039 – 0.002), but was the only model where the confidence interval for calibration slope did not include 1. The yearly updated logistic model had a calibration slope closest to 1, 0.995 (95% CI: 0.984 – 1.006). Over the entire validation data, no differences in discrimination or mean-squared error were observed between the models (Table 4.2).

Table 4.2 - Performance measures for each of the models validated on the complete validation data between January 2012 and March 2018

Models	Calibration-in-the-large	Calibration slope	Discrimination	Mean-squared error
Logistic model	0.158 (0.137 - 0.178)	1.016 (1.005 - 1.028)	0.904 (0.901 - 0.907)	0.015 (0.015 - 0.016)
Yearly updated logistic model	0.027 (0.006 - 0.047)	0.995 (0.984 - 1.006)	0.904 (0.901 - 0.907)	0.015 (0.015 - 0.016)
Bayesian updating model	-0.018 (-0.039 - 0.002)	0.971 (0.960 - 0.982)	0.903 (0.900 - 0.906)	0.015 (0.015 - 0.016)
Varying-coefficient model	0.158 (0.137 - 0.178)	1.016 (1.005 - 1.028)	0.904 (0.901 - 0.907)	0.015 (0.015 - 0.016)

Figure 4.3 displays each of the models' performances in the data computed monthly between 2012 and 2018. No difference in model discrimination was observed between any of the models. This was also observed for the calibration slope, but the dynamic model estimates fluctuated less and estimates more consistently close to 1 than the other models (figure 4.3). The standard deviation of the calibration slope in the monthly estimates was 0.034 compared to 0.051 in the logistic and varying coefficient models and 0.053 in the yearly updated logistic model (supplementary table S4.3). The largest difference between the model performances for the monthly validation was observed in the calibration-in-the-large. The dynamic model calibration-in-the-large remained stable around 0 and 1 respectively. The logistic regression and varying-coefficient model calibration-in-the-large was significantly above 0 for most of 2012 and all 3 months of the 2018 data. The yearly updated logistic model calibration-in-the-large was significantly higher than 0 for most 2012 and then was consistently below 0 in 2013. However, the calibration-in-the-large then remained stable and did not significantly differ, except for one month, between 2014 and 2018. The standard deviation of calibration-in-the-large for the dynamic model was 0.036 compared to 0.098, 0.123 and 0.101 observed in the logistic

regression, yearly updated logistic and varying-coefficient models respectively (supplementary table S4.3)

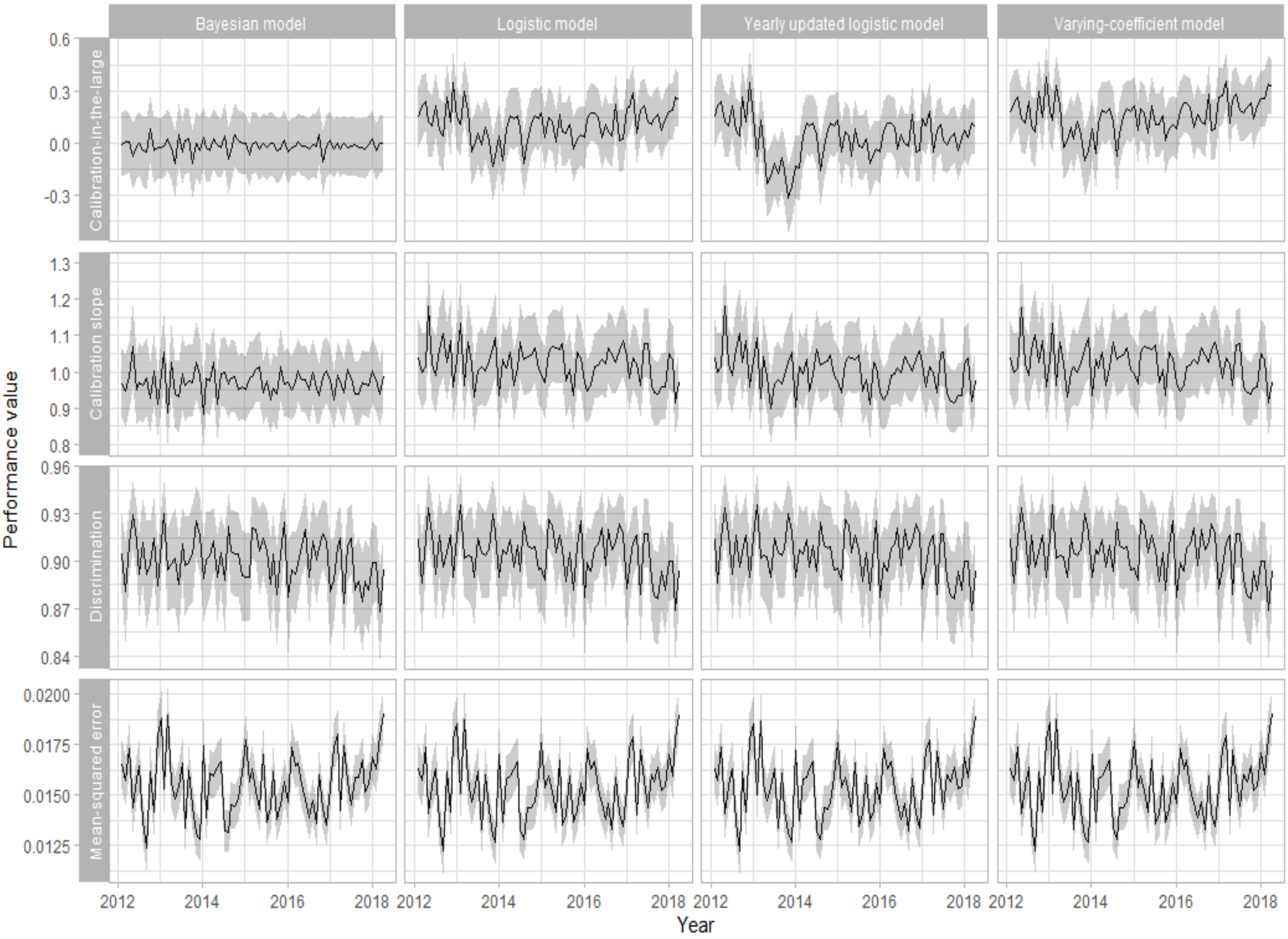


Figure 4.3 - Monthly performance measure for each model between 2012 and 2018

4.6 Discussion

In this study, we have compared the predictive performance of dynamic prediction modelling methods and traditional CPMs under a variety of simulated scenarios and a real-world data set. We found the Bayesian updating model either outperformed or performed as well as the time-invariant methods in all scenarios for the linear outcome. The varying coefficient models had good predictive performance, especially when the changes over time were linear, and the model was correctly specified. In the simulations, the varying coefficient model was the model with the best predictive performance, but for the binary outcome, the confidence intervals were much wider than the Bayesian and logistic regression models, indicating reduced stability[93]. The Bayesian model consistently had the narrowest confidence intervals in the simulation and in the real-world dataset had the best predictive performance, and lowest variability over time, in the calibration-in-the-large. However, the Bayesian model also had the worst calibration slope in the real-world data set. The yearly updated logistic model had the closest calibration slope to one out of all models, 0.995 (95% CI: 0.984 - 1.006), and had better calibration-in-the-large, closer to 0, than the varying coefficient and logistic models.

Although the best performing in the simulation model was the varying coefficient model when we observe linear changes over time, they are less promising when we observe step changes. Needing to correctly specify the functional form of the model with respect to time makes the varying coefficient model less desirable than the Bayesian model because we do not know the ground truth of any changes over time in practice. However, the Bayesian model is capable of reacting to changes and the larger the data, the quicker the response. We therefore advise the use of Bayesian models in the presence of temporal changes in the data, but if using varying coefficient models, take caution in extrapolating into the future as performance is more volatile and consider fixing the time when making predictions to the last time in the development data (or a time soon after the development data).

We have shown Dynamic prediction modelling can reduce the impact of calibration drift; it is therefore essential to increase our understanding of these models and how they can be developed to better facilitate healthcare by providing more stable and precise predictions. Ignoring temporal changes in the data could decrease the predictive performance of a prediction model over time. Hence, when developing and validating prediction models, one should investigate temporal trends in the data. In addition, models used in practice can directly impact the data because users respond to predictions, and this can change distributions of the outcomes and characteristics recorded in the data. For example, a clinician may prescribe a treatment after using the model and this could impact the outcome. Previous work[87,90,94] has suggested continuous surveillance and updating of models may overcome these problems. Our study supports this, showing continuous updating improved performance. However, we do not consider all possible scenarios, rather a set of simplistic simulated scenarios that may not be representative of data in clinical practice.

Furthermore, there are considerable difficulties in implementing dynamic models in practice: including technical and conceptual difficulties. These models would need a continuous flow of data, or at least receive regular data batches, to update. Hence, they require a complex infrastructure to support this data collection and model updating. Also, these approaches are not trivial to understand and the fact a patient's predicted risk from the model can change over time, may make the decision making more challenging and difficult to explain.

4.6.1 Previous literature

Limited research has been conducted in the area of dynamic clinical prediction models[84] and only a small subset of those have compared dynamic Bayesian models to time-invariant models[47,57]. These all compare regression models with Bayesian model updating in cardiac datasets. They showed that dynamic models retain good performance over time and observed improvement in the calibration-in-the-large. However, they found no differences in the discrimination or calibration slope. Our results support their findings. Siregar et al[56] also compared dynamic Bayesian models to time-invariant models and concluded that dynamic models are preferred. However, none of these studies included varying coefficient models or consider choosing forgetting based upon sample size.

Raftery et al and McCormick et al apply Bayesian model updating in a simulation considering a continuous and binary outcome, respectively. A limited number of scenarios are simulated, for example, Raftery et al simulates 3 unique scenarios. Neither of these studies consider the models in the context of clinical prediction and as a result the predictive performance of the models are not evaluated.

4.6.2 Strengths and limitations

The main strength of this work is that we perform a simulation study under a range of scenarios and consider multiple performance measures, thereby allowing a comprehensive and systematic examination of modelling approaches. To our knowledge we are the first to compare time-invariant models, varying coefficient and Bayesian updating models. We are also first to consider sample size and we have shown a way to choose the forgetting factor to satisfy a required sample size. Future applications in dynamic models should consider sample size and the choice of forgetting factor.

Conversely, the main limitation is that we simulate only a crude reflection of the real world and a limited number of possible scenarios. It was not possible to consider the infinite possibilities of reality. However, we evaluated the methods in a real-world dataset and illustrate the use of prequential testing that could be used to monitor the predictive performance of models in numerous scenarios otherwise considered here. Other limitations include: 1) only considering two predictors in the simulation; 2) penalisation was not considered; and 3) we do not address the issue of censoring, or delayed outcome availability, which would require extension to the method.

Although we have outlined an appropriate way to choose a forgetting factor, we only considered one option. This was not the aim of the study but it is likely an influential component in the performance of prediction models. When the healthcare system is changing rapidly, forgetting more quickly could help improve/retain the models performance. Further methodological work is required to optimise this 'forgetting' element of prediction models and devise a way to select the optimal forgetting when developing prediction models to increase performance. Indeed the optimum way to forget data over time might not be by choosing a single value but rather allowing data to be down weighted in a dynamic manner through some data driven optimisation.

4.7 Conclusion

How we acknowledge time in predictive modelling impacts model performance. Dynamic models offer a solution to model coefficient drift and also perform as well as time-invariant CPMs when there are little to no temporal changes. However, further methodological research is needed as well as research to increase utility of these models and implement them for healthcare use. We recommend the use of dynamic prediction models over time-invariant models. Specifically, the use of Bayesian model updating with forgetting and frequent updating, if continual updating is not possible.

4.8 References

1. Murdoch, T. B. & Detsky, A. S. The Inevitable Application of Big Data. *Jama* **309**, 1351–1352 (2014).

2. Wang, Y., Kung, L. A. & Byrd, T. A. Big data analytics: Understanding its capabilities and potential benefits for healthcare organizations. *Technol. Forecast. Soc. Change* **126**, 3–13 (2018).

3. Raghupathi, W. & Raghupathi, V. Big data analytics in healthcare: promise and potential. *Heal. Inf. Sci. Syst.* **2**, 3 (2014).

4. Friedman, C. P., Wong, A. K. & Blumenthal, D. Achieving a Nationwide Learning Health System. *Public Law* **2**, 8–10 (2010).

5. Friedman, C. P., Rubin, J. C. & Sullivan, K. J. Toward an Information Infrastructure for Global Health Improvement. *Yearb. Med. Inform.* **26**, 16–23 (2017).

6. J., A. & I., B. Combining health data uses to ignite health system learning. *Methods Inf. Med.* **54**, 479–487 (2015).

7. Kent, P., Cancelliere, C., Boyle, E., Cassidy, J. D. & Kongsted, A. A conceptual framework for prognostic research. *BMC Med. Res. Methodol.* **7**, 1–13 (2020).

8. Hemingway, H. *et al.* Prognosis research strategy (PROGRESS) 1: A framework for researching clinical outcomes. *BMJ* **346**, 1–11 (2013).

9. Riley, R. D. *et al.* Prognosis Research Strategy (PROGRESS) 2: Prognostic Factor Research. *PLoS Med.* **10**, e1001380 (2013).

10. Steyerberg, E. *et al.* Prognosis research strategy (PROGRESS) 3: prognostic model research. *PLoS Med* **10**, e1001381 (2013).

11. Hingorani, A. D. *et al.* Prognosis research strategy (PROGRESS) 4: Stratified medicine research. *BMJ* **346**, 1–9 (2013).

12. Riley, R. D., Windt, D. Van Der & Moons, K. G. M. Prognosis Research in Health Care. *Progn. Res. Heal. Care* 1–11 (2019). doi:10.1093/med/9780198796619.001.0001

13. Steyerberg, E. W. *Clinical Prediction Models. Statistics for Biology and Health. 2nd edition*. (2019).

14. Hemingway, H. *et al.* Prognosis research strategy (PROGRESS) 1: A framework for researching clinical outcomes. *BMJ* **346**, 1–11 (2013).

15. Gray, L. J. *et al.* The Leicester Risk Assessment score for detecting undiagnosed Type 2 diabetes and impaired glucose regulation for use in a multiethnic UK setting. *Diabet. Med.* **27**, 887–895 (2010).

16. (NICE), N. I. for H. and C. E. Type 2 diabetes: prevention in people at high risk. *Clinical guidelines [PH38]* (2012). Available at: https://www.nice.org.uk/guidance/ph38.

17. McAllister, K. S. L. *et al.* A contemporary risk model for predicting 30-day mortality following percutaneous coronary intervention in England and Wales. *Int. J. Cardiol.* **210**, 125–132 (2016).

18. Damen, J. A. A. G. *et al.* Prediction models for cardiovascular disease risk in the general population: Systematic review. *BMJ* **353**, (2016).

19. D'Agostino, R. B. *et al.* General Cardiovascular Risk Profile for Use in Primary Care: The Framingham Heart Study. *Circulation* **117**, 743–753 (2008).

20. Hippisley-Cox, J. *et al.* Derivation and validation of QRISK, a new cardiovascular disease risk score for the United Kingdom: Prospective open cohort study. *Br. Med. J.* **335**, 136–141 (2007).

21. (NICE), N. I. for H. and C. E. Cardiovascular disease: risk assessment and reduction, including lipid modification. *NICE Guidel. [CG181]* (2014).

22. Hippisley-Cox, J., Coupland, C., Robson, J. & Brindle, P. Derivation, validation, and evaluation of a new QRISK model to estimate lifetime risk of cardiovascular disease: Cohort study using QResearch database. *Bmj* **342**, 93 (2011).

23. Bonnett, L. J., Snell, K. I. E., Collins, G. S. & Riley, R. D. Guide to presenting clinical prediction models for use in clinical settings. *BMJ* **365**, 1–8 (2019).

24. Riley, R. D. *et al.* Minimum sample size for developing a multivariable prediction model: PART II - binary and time-to-event outcomes. *Stat. Med.* **38**, 1276–1296 (2019).

25. Steyerberg, E. W. *Clinical prediction models: a practical approach to development, validation, and updating*. (2008).

26. Harrell Jr, F. E. *Regression Modeling Strategies : With Applications to Linear Models, Logistic and Ordinal Regression, and Survival Analysis in Springer Series in Statistics*. *Springer* (Springer, 2015).

27. Steyerberg, E. W. *et al.* Internal validation of predictive models: Efficiency of some procedures for logistic regression analysis. *J. Clin. Epidemiol.* **54**, 774–781 (2001).

28. Steyerberg, E. W. *et al.* Assessing the performance of prediction models: a framework for some traditional and novel measures. *Epidemiology* **21**, 128–138 (2013).

29. Alba, A. C. *et al.* Discrimination and calibration of clinical prediction models: Users' guides to the medical literature. *JAMA - J. Am. Med. Assoc.* **318**, 1377–1384 (2017).

30. Van Calster, B. *et al.* Calibration: The Achilles heel of predictive analytics. *BMC Med.* **17**, 1–7 (2019).

31. Steyerberg, E. W. & Vergouwe, Y. Towards better clinical prediction models: Seven steps for development and an ABCD for validation. *Eur. Heart J.* **35**, 1925–1931 (2014).

32. Kramer, A. A. & Zimmerman, J. E. Assessing the calibration of mortality benchmarks in critical care: The Hosmer-Lemeshow test revisited. *Crit. Care Med.* **35**, 2052–2056 (2007).

33. Hickey, G. L. *et al.* Dynamic trends in cardiac surgery: Why the logistic euroscore is no longer suitable for contemporary cardiac surgery and implications for future risk models. *Eur. J. Cardio-thoracic Surg.* **43**, 1146–1152 (2013).

34. Davis, S. E., Lasko, T. A., Chen, G., Siew, E. D. & Matheny, M. E. Calibration drift in regression and machine learning models for acute kidney injury. *J. Am. Med. Informatics Assoc.* **24**, 1052–1061 (2017).

35. van Houwelingen, H. C. & Thorogood, J. Construction, validation and updating of a prognostic model for kidney graft survival. *Stat. Med.* **14**, 1999–2008 (1995).

36. Janssen, K. J. M., Moons, K. G. M., Kalkman, C. J., Grobbee, D. E. & Vergouwe, Y. Updating

methods improved the performance of a clinical prediction model in new patients. *J. Clin. Epidemiol.* **61**, 76–86 (2008).

37. Debray, T. P. A., Koffijberg, H., Vergouwe, Y., Moons, K. G. M. & Steyerberg, E. W. Aggregating published prediction models with individual participant data: A comparison of different approaches. *Stat. Med.* **31**, 2697–2712 (2012).

38. Martin, G. P., Mamas, M. A., Peek, N., Buchan, I. & Sperrin, M. Clinical prediction in defined populations: a simulation study investigating when and how to aggregate existing models. *BMC Med. Res. Methodol.* **17**, 1 (2017).

39. Martin, G. P., Mamas, M. A., Peek, N., Buchan, I. & Sperrin, M. A multiple-model generalisation of updating clinical prediction models. *Stat. Med.* **37**, 1343–1358 (2018).

40. Steyerberg, E. W., Borsboom, G. J. J. M., van Houwelingen, H. C., Eijkemans, M. J. C. & Habbema, J. D. F. Validation and updating of predictive logistic regression models: A study on sample size and shrinkage. *Stat. Med.* **23**, 2567–2586 (2004).

41. Toll, D. B., Janssen, K. J. M., Vergouwe, Y. & Moons, K. G. M. Validation, updating and impact of clinical prediction rules: A review. *J. Clin. Epidemiol.* **61**, 1085–1094 (2008).

42. Hippisley-Cox, J., Coupland, C. & Brindle, P. Development and validation of QRISK3 risk prediction algorithms to estimate future risk of cardiovascular disease : prospective cohort study. **2099**, 1–21 (2017).

43. Nashef, S. A. M. *et al.* European system for cardiac operative risk evaluation (EuroSCORE). *Eur. J. Cardio-thoracic Surg.* **16**, 9–13 (1999).

44. Nashef, S. A. M. *et al.* Euroscore II. *Eur. J. Cardio-thoracic Surg.* **41**, 734–745 (2012).

45. Raftery, A. E. & Ettler, P. Online Prediction Under Model Uncertainty via Dynamic Model Averaging : Application to a Cold Rolling Mill. **52**, 52–66 (2010).

46. Bull, L. M., Lunt, M., Martin, G. P., Hyrich, K. & Sergeant, J. C. Harnessing repeated measurements of predictor variables for clinical risk prediction: a review of existing methods. *Diagnostic Progn. Res.* **4**, (2020).

47. Hickey, G. L. *et al.* Dynamic prediction modeling approaches for cardiac surgery. *Circ. Cardiovasc. Qual. Outcomes* **6**, 649–658 (2013).

48. Mccormick, T. H., Raftery, A. E., Madigan, D. & Burd, R. S. Dynamic Logistic Regression and Dynamic Model Averaging for Binary Classification. *Biometrics* **68**, 23–30 (2012).

49. McCormick, T. H., Raftery, A. & Madigan, D. dma: Dynamic Model Averaging. (2018).

50. Five Year Forward View. (2014).

51. Salive, M. E. Multimorbidity in older adults. *Epidemiol. Rev.* **35**, 75–83 (2013).

52. Divo, M. J., Martinez, C. H. & Mannino, D. M. Ageing and the epidemiology of multimorbidity. *Eur. Respir. J.* **44**, 1055–1068 (2014).

53. Watkins, J. *et al.* Effects of health and social care spending constraints on mortality in England: a time trend analysis. *BMJ Open* **7**, e017722 (2017).

54. Abu-Hanna, A. & Lucas, P. J. F. Prognostic Models in Medicine. AI and Statistical Approaches.

Method Inf. Med **40**, 1–5 (2001).

55. Damen, J. A. A. G. *et al.* Prediction models for cardiovascular disease risk in the general population: systematic review. *Bmj* i2416 (2016). doi:10.1136/bmj.i2416

56. Siregar, S. *et al.* Improved Prediction by Dynamic Modeling. *Circ. Cardiovasc. Qual. Outcomes* **9**, 171–181 (2016).

57. Su, T.-L., Jaki, T., Hickey, G. L., Buchan, I. & Sperrin, M. A review of statistical updating methods for clinical prediction models. *Stat. Methods Med. Res.* 1–16 (2016). doi:10.1177/0962280215626466

58. van Houwelingen, H. C. & Thorogood, J. Construction, validation and updating of a prognostic model for kidney graft survival. *Stat. Med.* **14**, 1999–2008 (1995).

59. Draper, N. R., Nostrand, R. C. Van & Draper, N. R. Ridge Regression and James-Stein Estimation : Review and Comments Linked references are available on JSTOR for this article : Ridge Regression and James-Stein Estimation : Review and Comments. **21**, 451–466 (2016).

60. Copas, J. . Regression, Prediction and Shrinkage. *R. Stat. Soc.* **45**, 311–354 (1983).

61. Finkelman, B. S., French, B. & Kimmel, S. E. The prediction accuracy of dynamic mixed-effects models in clustered data. *BioData Min.* **9**, 5 (2016).

62. Fan, J. & Zhang, W. Statistical Methods with Varying Coefficient Models. *Stat Interface* **1**, 179–195 (2008).

63. Hoover, D. R., Rice, J. A., Wu, C. O. & Yang, L.-P. Nonparametric smoothing estimates of time-varying coefficient models with longitudinal data. *Biometrika* **85**, 809–822 (1998).

64. Madigan, D. & Raftery, A. E. Model Selection and Accounting for Model Uncertainty in Graphical Models Using Occam's Window. (1991).

65. Onorante, L. & Raftery, A. E. Dynamic model averaging in large model spaces using dynamic Occam's window. *Eur. Econ. Rev.* **81**, 2–14 (2016).

66. Ohata, T., Kaneko, M., Kuratani, T., Ueda, H. & Shimamura, K. Using the EuroSCORE to assess changes in the risk profiles of the patients undergoing coronary artery bypass grafting before and after the introduction of less invasive coronary surgery. *Ann. Thorac. Surg.* **80**, 131–135 (2005).

67. Ramsay, J. O., Wickham, H., Graves, S. & Hooker, G. fda: Functional Data Analysis. (2017).

68. Altman, D. G. & Royston, P. What do we mean by validating a prognistic model? *Stat. Med.* **19**, 453–473 (2000).

69. Justice, A. C., Covinsky, K. E. & Berlin, J. A. Assessing the generalizability of prognostic information. *Ann. Intern. Med.* **130**, 515–524 (1999).

70. Zimmerman, J. E., Kramer, A. A., McNair, D. S. & Malila, F. M. Acute Physiology and Chronic Health Evaluation (APACHE) IV: Hospital mortality assessment for today's critically ill patients. *Crit. Care Med.* **34**, 1297–1310 (2006).

71. Vergouwe, Y. *et al.* A closed testing procedure to select an appropriate method for updating prediction models. *Stat. Med.* **36**, 4529–4539 (2017).

72. Hafkamp-De Groen, E. *et al.* Predicting asthma in preschool children with asthma-like symptoms: Validating and updating the PIAMA risk score. *J. Allergy Clin. Immunol.* **132**, (2013).

73. Genders, T. S. S. *et al.* A clinical prediction rule for the diagnosis of coronary artery disease: Validation, updating, and extension. *Eur. Heart J.* **32**, 1316–1330 (2011).

74. Moons, K. G. M. *et al.* Risk prediction models: II. External validation, model updating, and impact assessment. *Heart* **98**, 691–698 (2012).

75. Hippisley-Cox, J., Coupland, C. & Brindle, P. Development and validation of QRISK3 risk prediction algorithms to estimate future risk of cardiovascular disease: Prospective cohort study. *BMJ* **357**, 1–21 (2017).

76. Martin, G. P., Sperrin, M. & Sotgiu, G. Performance of Prediction Models for Covid-19: The Caudine Forks of the External Validation. *Eur. Respir. J.* 2003728 (2020). doi:10.1183/13993003.03728-2020

77. Wynants, L. *et al.* Prediction models for diagnosis and prognosis of covid-19: Systematic review and critical appraisal. *BMJ* **369**, (2020).

78. Vergouwe, Y., Moons, K. G. M. & Steyerberg, E. W. External validity of risk models: Use of benchmark values to disentangle a case-mix effect from incorrect coefficients. *Am. J. Epidemiol.* **172**, 971–980 (2010).

79. Debray, T. P. A. *et al.* A guide to systematic review and meta-analysis of prediction model performance. *BMJ* **356**, (2017).

80. Luijken, K. *et al.* Changing predictor measurement procedures affected the performance of prediction models in clinical examples. *J. Clin. Epidemiol.* **119**, 7–18 (2020).

81. Riley, R. D. *et al.* External validation of clinical prediction models using big datasets from e-health records or IPD meta-analysis: Opportunities and challenges. *BMJ* **353**, 27–30 (2016).

82. Debray, T. P. A., Riley, R. D., Rovers, M. M., Reitsma, J. B. & Moons, K. G. M. Individual Participant Data (IPD) Meta-analyses of Diagnostic and Prognostic Modeling Studies: Guidance on Their Use. *PLoS Med.* **12**, 1–12 (2015).

83. Hippisley-Cox, J. *et al.* Predicting cardiovascular risk in England and Wales: Prospective derivation and validation of QRISK2. *Bmj* **336**, 1475–1482 (2008).

84. Jenkins, D. A., Sperrin, M., Martin, G. P. & Peek, N. Dynamic models to predict health outcomes: current status and methodological challenges. *Diagnostic Progn. Res.* **2**, 23 (2018).

85. Halabi, S. *et al.* Updated prognostic model for predicting overall survival in first-line chemotherapy for patients with metastatic castration-resistant prostate cancer. *J. Clin. Oncol.* **32**, 671–677 (2014).

86. Dawid, A. Present Position and Potential Developments : Some Personal Views : Statistical Theory : The Prequential Approach Author (s): A . P . Dawid Source : Journal of the Royal Statistical Society . Series A (General), Vol . 147 , No . 2 , The 150th Annivers. *J. R. Stat. Soc. Ser. A* **147**, 278–292 (1984).

87. Lenert, M. C., Matheny, M. E. & Walsh, C. G. Prognostic models will be victims of their own success, unless…. *J. Am. Med. Inform. Assoc.* **26**, 1645–1650 (2019).

88. Adibi, A., Sadatsafavi, M. & Ioannidis, J. P. A. Validation and Utility Testing of Clinical Prediction

Models. *JAMA* **2004**, (2020).

89. Booth, S., Riley, R. D., Ensor, J., Lambert, P. C. & Rutherford, M. J. Temporal recalibration for improving prognostic model development and risk predictions in settings where survival is improving over time. *Int. J. Epidemiol.* 1–10 (2020). doi:10.1093/ije/dyaa030

90. Jenkins, D. A. *et al.* Continual updating and monitoring of clinical prediction models: time for dynamic prediction systems? *Diagnostic Progn. Res.* **5**, 1–7 (2021).

91. Mccormick, T. H. *et al.* Dynamic Logistic Regression and Dynamic Model Averaging for Binary Classification. 23–30 (2012). doi:10.1111/j.1541-0420.2011.01645.x

92. R Core Team. R: A language and environment for statistical computing. R Foundation for Statistical. (2014).

93. Martin, G. P., Riley, R. D., Collins, G. S. & Sperrin, M. Developing clinical prediction models when adhering to minimum sample size recommendations: The importance of quantifying bootstrap variability in tuning parameters and predictive performance. *Stat. Methods Med. Res.* **30**, 2545–2561 (2021).

94. Sperrin, M., Jenkins, D., Martin, G. P. & Peek, N. Explicit causal reasoning is needed to prevent prognostic models being victims of their own success. *J. Am. Med. Informatics Assoc.* **26**, 1675–1676 (2019).

95. Hickey, G. L. *et al.* Clinical registries: Governance, management, analysis and applications. *Eur. J. Cardio-thoracic Surg.* **44**, 605–614 (2013).

96. Moons, K. G. M. *et al.* Risk prediction models: II. External validation, model updating, and impact assessment. *Heart* **98**, 691–698 (2012).

97. Davis, S. E. *et al.* A nonparametric updating method to correct clinical prediction model drift. *J. Am. Med. Informatics Assoc.* **26**, 1448–1457 (2019).

98. Minne, L. *et al.* Statistical process control for monitoring standardized mortality ratios of a classification tree model. *Methods Inf. Med.* **51**, 353–358 (2012).

99. Statistics, M. SUMS OF NONINDEPENDENT BERNOULLI RANDOM VARIABLES Author (s): Jose M . Gonzalez-Barrios Source : Brazilian Journal of Probability and Statistics , JUNE 1998 , Vol . 12 , No . 1 (JUNE Published by : Institute of Mathematical Statistics Stable URL : http. **12**, 55–64 (1998).

100. Koetsier, A., De Keizer, N. F., De Jonge, E., Cook, D. A. & Peek, N. Performance of risk-adjusted control charts to monitor in-hospital mortality of intensive care unit patients: A simulation study. *Crit. Care Med.* **40**, 1799–1807 (2012).

101. Team, R. core. R: A language and environment for statistical computing. *R Foundation for Statistical Computing, Vienna, Austria.* (2021). Available at: https://www.r-project.org/.

102. Albert, A. A. *et al.* On-line variable live-adjusted displays with internal and external risk-adjusted mortalities. A valuable method for benchmarking and early detection of unfavourable trends in cardiac surgery. *Eur. J. Cardio-thoracic Surg.* **25**, 312–319 (2004).

103. Lovegrove, J., Valencia, O., Treasure, T., Sherlaw-Johnson, C. & Gallivan, S. Monitoring the results of cardiac surgery by variable life-adjusted display. *Lancet* **350**, 1128–1130 (1997).

104. Poloniecki, J., Valencia, O. & Littlejohns, P. Correction: Cumulative risk adjusted mortality chart

for detecting changes in death rate: Observational study of heart surgery (British Medical Journal (1998) (1697-700)). *Br. Med. J.* **316**, 1947 (1998).

105. Minne, L. *et al.* Effect of changes over time in the performance of a customized SAPS-II model on the quality of care assessment. *Intensive Care Med.* **38**, 40–46 (2012).

106. Shi, X., Gallagher, C., Lund, R. & Killick, R. A Comparison of Single and Multiple Changepoint Techniques for Time Series Data. 1–29 (2021). doi:10.1016/j.csda.2022.107433

107. Wittenberg, P., Gan, F. F. & Knoth, S. A simple signaling rule for variable life-adjusted display derived from an equivalent risk-adjusted CUSUM chart. *Stat. Med.* **37**, 2455–2473 (2018).

108. Davis, S. E., Greevy, R. A., Lasko, T. A., Walsh, C. G. & Matheny, M. E. Detection of calibration drift in clinical prediction models to inform model updating. *J. Biomed. Inform.* **112**, 103611 (2020).

109. Riley, R. D. *et al.* Minimum sample size for external validation of a clinical prediction model with a binary outcome. *Stat. Med.* **40**, 4230–4251 (2021).

110. Steyerberg, E. W., Borsboom, G. J. J. M., van Houwelingen, H. C., Eijkemans, M. J. C. & Habbema, J. D. F. Validation and updating of predictive logistic regression models: A study on sample size and shrinkage. *Stat. Med.* **23**, 2567–2586 (2004).

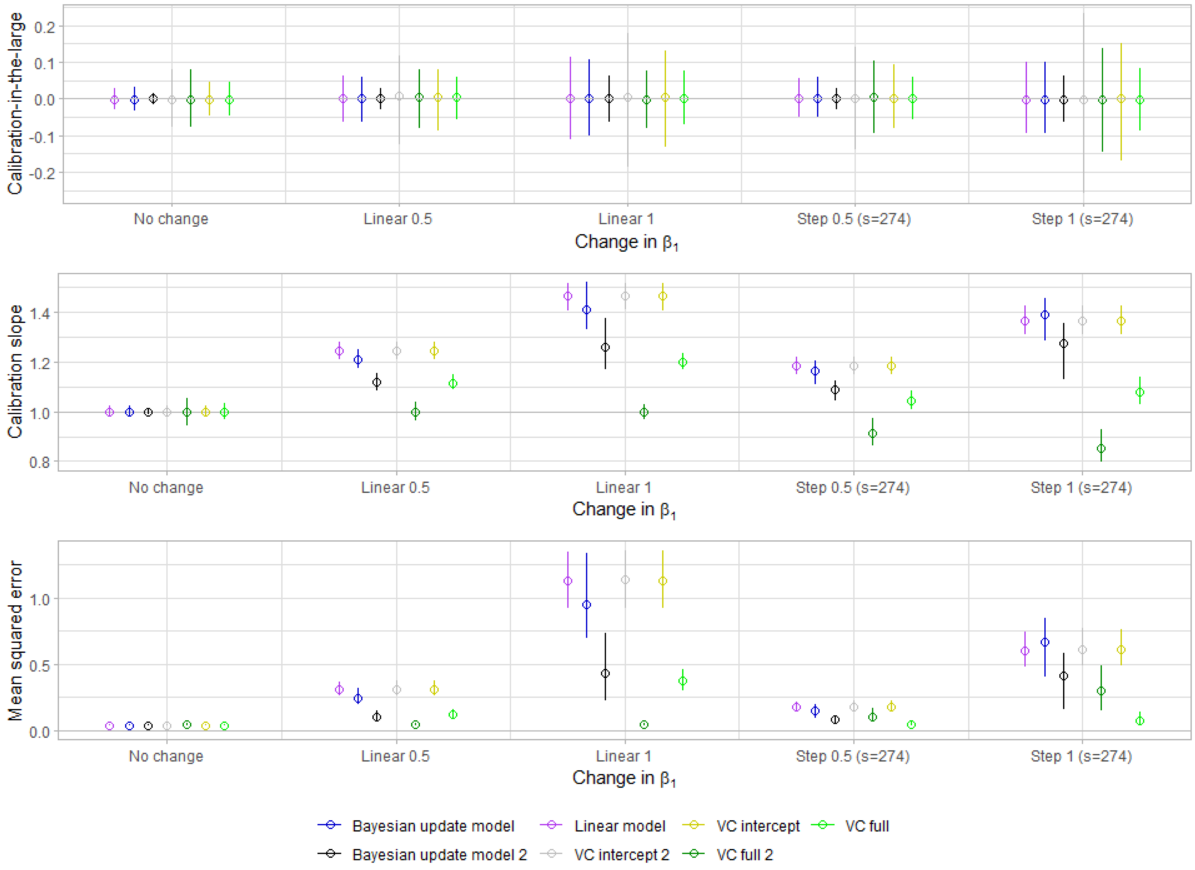
111. Grigg, O. A., Farewell, V. T. & Spiegelhalter, D. J. Use of risk-adjusted CUSUM and RSPRT charts for monitoring in medical contexts. *Stat. Methods Med. Res.* **12**, 147–170 (2003).

112. Pagel, C. *et al.* Real time monitoring of risk-adjusted paediatric cardiac surgery outcomes using variable life-adjusted display: Implementation in three UK centres. *Heart* **99**, 1445–1450 (2013).

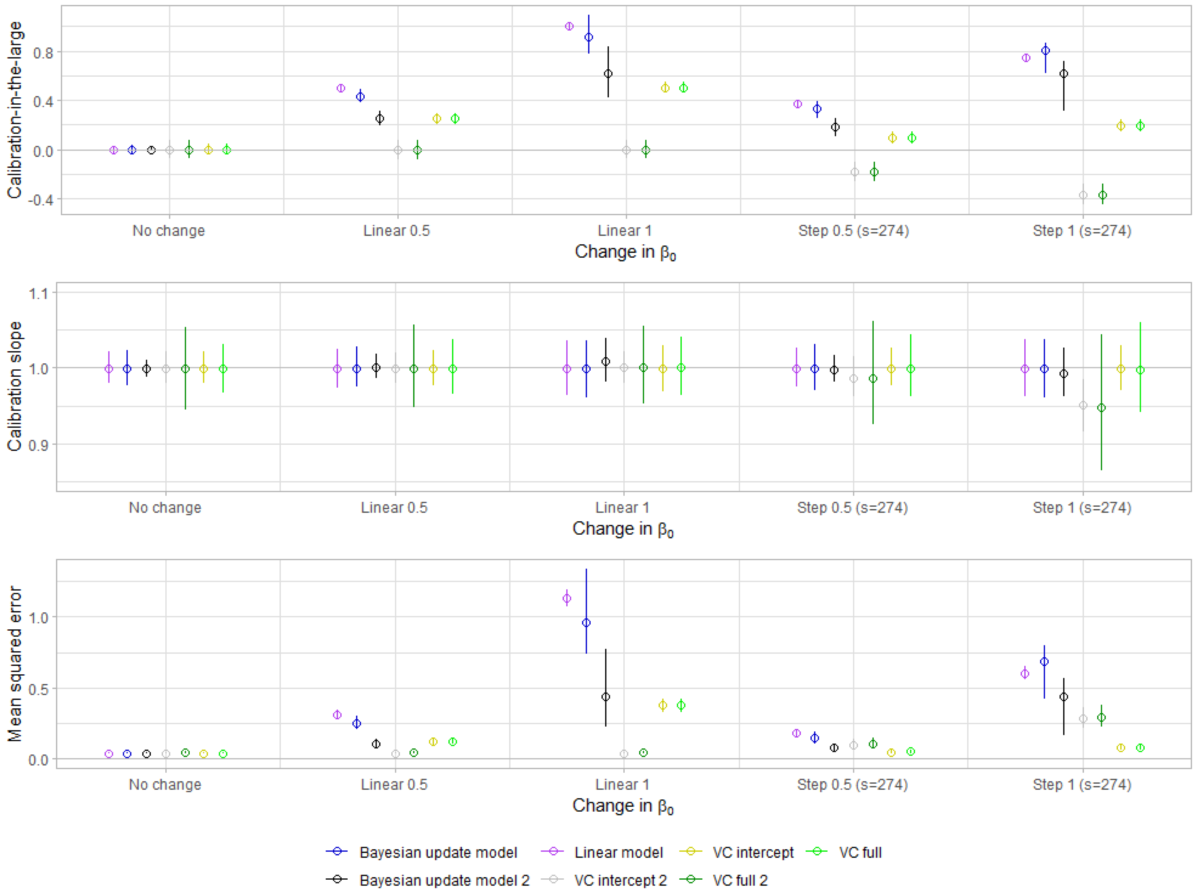
113. Barrett, J. & Su, L. Dynamic predictions using flexible joint models of longitudinal and time-to-event data. *Stat. Med.* **36**, 1447–1460 (2017).

114. van Houwelingen, H., & Putter, H. *Dynamic Prediction in Clinical Survival Analysis*. (2012). doi:https://doi.org/10.1201/b11311

4.9 Supplementary material



Supplementary Figure S4.1 - Predictive performance averaged over the full validation data for each linear model separately and for scenarios when $\beta_0(t)$ is static and $\beta_1(t)$ is the same as the selected scenarios. Model 2 refers to the model updating in the validation data (for the Bayesian model) and the model extrapolated to the validation time (for the varying coefficient model). The circles represent median values across the 1000 iterations and vertical lines represent the 95% quantiles.



Supplementary Figure S4.2 - Predictive performance averaged over the full validation data for each linear model separately and for scenarios when $\beta_0(t)$ is the same as the selected scenarios and $\beta_1(t)$ is static. Model 2 refers to the model updating in the validation data (for the Bayesian model) and the model extrapolated to the validation time (for the varying coefficient model). The circles represent median values across the 1000 iterations and vertical lines represent the 95% quantiles.

Supplementary Table S4.3 - Mean and standard deviation of the monthly performance values between 2012 and 2018 for each model

Models	Calibration-in-the-large	Calibration slope	Discrimination	Mean-squared error
Logistic model	0.107 (0.098)	1.020 (0.051)	0.904 (0.015)	0.015 (0.002)
Yearly updated logistic model	0.026 (0.123)	1.000 (0.053)	0.904 (0.015)	0.015 (0.002)
Bayesian updating model	-0.019 (0.036)	0.972 (0.034)	0.901 (0.014)	0.016 (0.002)
Varying-coefficient model	0.154 (0.101)	1.020 (0.051)	0.904 (0.015)	0.015 (0.002)

Chapter 5 Comparing predictive performance of time invariant and time variant clinical prediction models in a UK cardiac surgery dataset

David A Jenkins, Glen P Martin, Matthew Sperrin, Benjamin Brown, Stuart Grant, and Niels Peek

In preparation for submission

5.1 Abstract

5.1.1 Background

Clinical prediction models (CPMs) are used across healthcare to support clinical decision making. The European System for Cardiac Operative Risk Evaluation (EuroSCORE) is one of these models and is used to assess risk of mortality after cardiac surgery. Existing CPMs, including EuroSCORE, are time invariant: they do not acknowledge temporal changes which can result in worsening of predictive performance over time. However, methods are available for time-variant models. We aimed to compare the performance of time-invariant with time-variant models in a single-centre cardiac surgery dataset over a ten-year period.

5.1.2 Methods

We analysed UK National Adult Cardiac Surgery Audit data from Manchester University NHS Foundation Trust between 2009 and 2019. We fitted 4 models to the data, a (time-invariant) logistic regression model and time-variant logistic regression model with varying coefficients (where the intercept is a function of calendar time) to data from 2009-2011 and validated them in the 2012-2019 data. We also fitted a time-invariant logistic model which was updated every year, validating it in each subsequent year. Finally, a continually updating Bayesian logistic model, updating with each new observation and down-weighting older observations, was fit to the data and continuously validated. We report calibration (observed-expected ratio, calibration-in-the-large and calibration slope) and discrimination (C-statistic) over the complete validation cohort and for each year in the validation data.

5.1.3 Results

The dataset comprised of 10,770 individuals, 3021 in the (initial) development data and 7749 in the validation data. The Bayesian model had the best predictive performance over the complete validation data for calibration-in-the-large, 0.019 (95% CI: -0.107, 0.14), and discrimination with a C-statistic of 0.778 (95% CI: 0.747, 0.809). The yearly updated logistic model was also well calibrated-in-the-large 0.057 (95% CI: -0.059, 0.17) but the varying coefficient and time invariant models consistently had the worst performance measures and were miscalibrated with calibration-in-the-large of -0.688 (95% CI: -0.849, -0.535) and -0.237 (95% CI: -0.398, -0.083), respectively.

5.1.4 Conclusion

The Bayesian and yearly updating models had the best predictive performance throughout the study. Therefore, we advise the use of model updating methods, specifically Bayesian model updating.

5.2 Background

Decision making in healthcare is often aided by tools known as clinical prediction models (CPMs)[12,25]. CPMs use information about a patient to provide risk estimates on a certain outcome for the patient, for example, risk of developing cardiovascular disease. The European System for Cardiac Operative Risk Evaluation[43] (EuroSCORE) is a cardiac risk model for predicting mortality after cardiac surgery. This model is used to aid the clinician's decision on whether or not they should perform surgery. The information from the model therefore needs to be accurate otherwise incorrect decisions could be made, impacting patient care and outcomes. EuroSCORE was published in 1999 and the accuracy of the model was shown to diminish over time[33]. Hickey et al showed the observed-expected ratio of the EuroSCORE model in predicting mortality decreased from 0.73 to 0.37 over a 10-year time period. Subsequently, an updated version of the model[44] (EuroSCORE II) was published in 2012. Degradation of CPMs used in clinical practice is often observed and many models experience diminishing predictive performance over time. The healthcare system is constantly evolving and patient populations are changing while our CPMs remain time-invariant and do not consider this temporal nature of the data.

Periodic updating is now increasingly being used to recalibrate models and overcome model degradation. QRISK[20] is an example of such a model, and is updated yearly[22,42,83], but few models are updated this often, if at all. Recently, other methods, known as dynamic prediction models[47,84], have been discussed to overcome model degradation, such as continuously updating Bayesian models[48] and varying coefficient models[63]. In Chapter 4 we compared these two dynamic modelling approaches with a time-invariant model and this was through a simulation study that included a cardiac data example. In the cardiac data we also compared the methods to yearly recalibration[41]. As far as we know, no other systematic comparisons of all these methods have been conducted.

In this chapter we aim to investigate the predictive performance of time-variant and time-invariant modelling methods using a real-world cardiac dataset. We apply the same four modelling approaches as in Chapter 4 to National Adult Cardiac Surgery Audit (NACSA) data collected at Manchester University NHS Foundation Trust between 2009 and 2019. The NACSA includes baseline data and clinical outcomes on individuals undergoing cardiac surgery. We validate and compare the performance of the methods in the data between 2012 and 2019.

5.3 Methods

5.3.1 Study Population and NACSA Registry

The NACSA registry collects data on major heart operations in the UK. The registry includes information on patient baseline demographics, risk factors for intervention, procedural details and patient outcomes. This study included data on all major heart operations from 1st January 2009 to 30th June 2019 from a single hospital. The outcome was discharge status (alive/died) and all predictors included in EuroSCORE II, except heart failure classification and creatinine, were available in the data. All variables were defined as per EuroSCORE II[44], for example, recent myocardial infarction (MI) was defined as MI within the 90 days prior to surgery. Missing categorical data were imputed based on an assumption that missingness was equal to risk-factor absent, representing a plausible missingness mechanism for the registry data[95]. We choose to use the variables in EuroSCORE because we are predicting a very similar outcome and to ensure the study closely represents current CPM practice. Also, the aim of the study is to compare performance rather than derive new models.

5.3.2 Statistical Analysis

Four models were developed and validated in the data. The first model was a time-invariant logistic regression model fitted to the data collected from 1st January 2009 to 31st December 2011. We applied the sample size calculation by Riley et al[24] to ensure adequate sample size. The second was a yearly updated logistic regression model. This was similar to the first model and used the data between 1st January 2009 and 31st December 2011 to develop the model but was then subsequently recalibrated at the start of each year[96]. This was chosen as it represents how some models in clinical practice, for example QRISK[42], are currently updated. The third model was a Bayesian time-variant model with continual updating[48]. The model was updated at each new observation and we derived the forgetting factor as described in chapter 4 and chose to use the size of the development data as the effective window size, which resulted in a forgetting factor of 0.9997. This was chosen to ensure that the dynamic model weighted individuals over time such that the sample size was comparable to the time-invariant logistic model. Finally, the fourth model was a time-variant logistic model with varying coefficients developed using the data from 1st January 2009 to 31st December 2011. Only the intercept term was dependent on time and the functional form was assumed to be linear. This was chosen as it is the simplest varying coefficient model and chapter 4 showed no improvement when further allowing the other coefficients to be dependent on time. Also, due to the amount of variables in the model, if we modelled each of the coefficients as functions of time this would require at least twice the sample size, meaning we would require a larger sample than was available between 1st January 2009 and 31st December 2011.

The models were then validated in the data collected from 1st January 2012 to 30th June 2019. For each model we calculated the calibration-in-the-large (CITL), calibration slope, discrimination (C-statistic) and the observed-expected (OE) ratio for each year seperatly. Prequential testing[86] was used to validate the continuously updated Bayesian dynamic model as described in chapter 4. Each validation measure was calculated for each year of data in the validation data. Chapter 4 showed that the variation in predictive performance varied between models. Therefore, we performed the Bartletts test to test for variation in performance over time. In addition to this, we computed each performance measure on the data between from 1st January 2012 and 30th June 2019 to quantify each model's overall performance in the data. All analyses were performed using R (version 3.6.2) and the dynamic models were fitted using functions adapted from the dma package[49].

5.4 Results

The final data comprised of 10,770 individuals, 3021 between 1st January 2009 to 31st December 2011, and 7749 between 1st January 2012 to 30th June 2019, and a total of 413 (3.83%) patients died following surgery (Table 5.1). A higher proportion of patients had experienced a myocardial infarction within 90 days prior to surgery in the validation data than the development data, 26.0% vs 17.1%. Between 1st January 2009 and 31st December 2011 3% of patients died prior to discharge compared to 4.1% between 1st January 2012 and 30th June 2019.

Table 5.1 - Baseline data in the development data, from 1st January 2009 to 31st December 2011, and the validation data, from 1st January 2012 to 30th June 2019.

Variable	Development data	Validation data
	n = 3021	n = 7749
Age*	65.8 (11.7)	65.2 (12.2)
Male	2146 (71%)	5629 (72.6%)
Diabetes	529 (17.5%)	1762 (22.7%)
History of Pulmonary Disease	428 (14.2%)	1122 (14.5%)
History of Neurological Dysfunction	101 (3.3%)	236 (3%)
Angina	82 (2.7%)	296 (3.8%)
Extracardiac arteriopathy	358 (11.9%)	759 (9.8%)
Previous surgery	122 (4%)	595 (7.7%)
Recent MI	518 (17.1%)	2011 (26%)
Surgery on thoracic aorta	108 (3.6%)	318 (4.1%)
LV function		
Good	2250 (74.5%)	5017 (64.8%)
Moderate	562 (18.6%)	1762 (22.7%)
Poor	199 (6.6%)	516 (6.7%)
Urgency		
Elective	2133 (70.6%)	4204 (54.3%)
Emergency	123 (4.1%)	303 (3.9%)
Urgent	765 (25.3%)	3242 (41.8%)
Critical preoperative state	140 (4.6%)	792 (10.2%)
Died	92 (3.0%)	321 (4.1%)

*mean (standard deviation)

The minimum sample size required for model development based on 17 candidate predictors, an outcome prevalence of 3.8 and r-squared of 0.056 was 2647. This r-squared was chosen as it is approximately a fifth the maximum possible r-squared as defined by equation 23 in Riley et al[24]. 3021 individuals were observed in data between 2009-2011, showing that the minimum required sample size was achieved.

The coefficients for the logistic regression and varying coefficient model can be found in supplementary table S5.1 and the coefficients for each updated version of the yearly updated logistic model can be found in supplementary table S5.2. A subset of the Bayesian model coefficients are also presented in supplementary material (supplementary table S5.3). The Bayesian model is updated with every observation, so we present the coefficients on the 1st January each year in the validation data.

5.4.1 Model performance

Figure 5.1 displays each of the model's performances separate for each year of data from 2012 to 2019. Evidence of variation in calibration and discrimination over time was observed. The bartlett test showed a significant difference in variance between models for calibration-in-the-large, p-value < 0.001. The p-values for the calibration slope, discrimination and observed-expected ratio from the Bartlett test of variance were 0.439, 0.072 and <0.001, respectively. Hence, there was no evidence of a difference in variance for the yearly calibration slope measure. The Bayesian model calibration-in-the-large, calibration slope and observed-expected ratio remained stable over time and the confidence interval for the calibration-in-the-large and observed-expected ratio includes 0 and 1, respectively, at all times (Figure 5.1). A reduction in the observed-expected ratio was observed in 2014 for the logistic model and varying coefficient model but the Bayesian and yearly updated logistic model retained good performance. The Bayesian model discrimination increases over time, except for 2019 where all four models discrimination estimates are approximately 0.7. The Bayesian model discrimination increases from 0.72 and 0.68 in 2012 and 2013, respectively, to 0.80 and 0.84 in 2017 and 2018. In comparison, the other model's discrimination remained between 0.67 and 0.75 in 2017 and 2018. There is evidence that the logistic model was miscalibrated in 2014 and the yearly updated logistic model was miscalibrated in 2015 as the confidence intervals for the calibration-in the-large do not include 0 (Figure 5.1). The calibration-in-the-large confidence interval for the varying coefficient model only includes 0 in 2019.

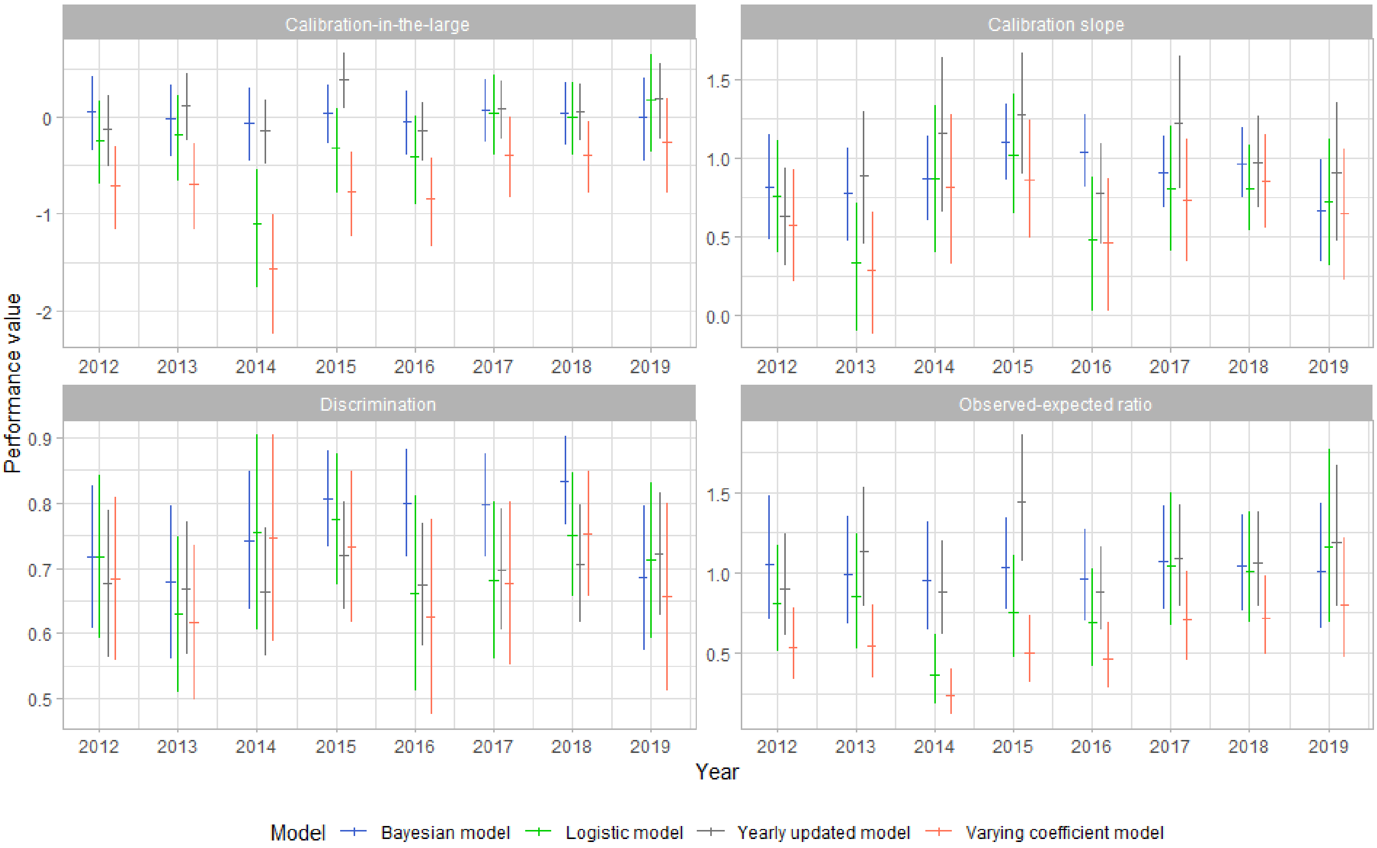


Figure 5.1 - Yearly performance measure for each model between 2012 and 2019

Table 5.2 displays each models performance values, including 95% confidence intervals (CIs), when validating using all of the validation data from 1st January 2012 to 30th June 2019. The varying coefficient model consistently had the worst performance for all validation measures. The confidence intervals of the calibration-in-the large, calibration slope and observed-expected ratio did not include 0, 1 and 1, respectively. The logistic model observed-expected ratio was 0.810 (95% CI: 0.694, 0.935) and was therefore miscalibrated as the confidence intervals did not include 1. However, the Bayesian model and yearly updated logistic model were well calibrated over the validation data and had observed-expected ratios of 1.016 (95% CI: 0.904, 1.134) and 1.054 (95% CI: 0.942, 1.172) over the complete validation data, respectively. The Bayesian model had the highest discrimination of the models over the complete data with a C-statistic of 0.778 (95% CI: 0.747, 0.809).

Table 5.2 – Performance measures and 95% confidence intervals from each model when validating in the complete validation data from 1st January 2012 to 30th June 2019.

Performance measure	Model	Performance value (95% CI)
Calibration-in-the-large	Logistic model	-0.237 (-0.398, -0.083)
	Yearly updated logistic model	0.057 (-0.059, 0.17)
	Bayesian model	0.019 (-0.107, 0.14)
	Varying coefficient model	-0.688 (-0.849, -0.535)
Calibration slope	Logistic model	0.73 (0.602, 0.856)
	Yearly updated logistic model	0.919 (0.794, 1.044)
	Bayesian model	0.925 (0.837, 1.013)
	Varying coefficient model	0.656 (0.525, 0.785)
Observed-expected ratio	Logistic model	0.81 (0.694, 0.935)
	Yearly updated logistic model	1.054 (0.942, 1.172)
	Bayesian model	1.016 (0.904, 1.134)
	Varying coefficient model	0.544 (0.466, 0.628)
Discrimination	Logistic model	0.709 (0.667, 0.751)
	Yearly updated logistic model	0.694 (0.66, 0.727)
	Bayesian model	0.778 (0.747, 0.809)
	Varying coefficient model	0.686 (0.642, 0.73)

5.5 Discussion

In this study we have developed models using four different modelling approaches: logistic regression, yearly updating logistic regression, Bayesian updating and varying coefficient models, and compared the predictive performance of the approaches. Over the complete validation data, the Bayesian model had the best predictive performance for calibration and discrimination, with a calibration-in-the-large of 0.019 (95% CI: -0.107, 0.14) and a C-statistic of 0.778 (95% CI: 0.747, 0.809). The Bayesian model provided the most stable yearly estimates of calibration-in-the-large and, on average, achieved the best discrimination, likely due to the improvement in discrimination between 2015 and 2018. The yearly updated logistic model was also well calibrated over the entire validation data and had a better calibration-in-the-large than the time-invariant logistic model, 0.057 (95% CI: -0.059, 0.17) vs -0.237 (95% CI: -0.398, -0.083). However, the yearly updated logistic model discrimination was worse than the time-invariant logistic model with a C-statistic of 0.694 (95% CI: 0.66, 0.727) vs 0.709 (95% CI: 0.667, 0.751). The varying coefficient model was the worst performing model with the lowest C-statistic of 0.686 (95% CI: 0.642, 0.73) and a calibration-in-the-large value of -0.688 (95% CI: -0.849, -0.535). Little difference was observed between the Bayesian and yearly updating logistic model for the calibration measures, calibration-in-the-large, calibration slope and observed-expected ratio. The Bayesian and yearly updating logistic model consistently outperformed the time-invariant logistic model and varying coefficient model with respect to calibration but the Bayesian model outperformed all models in discrimination

5.5.1 Implications for practice and research

Our work supports the idea that accounting for temporal changes in data improve model performance, but the more flexible and complex modelling approaches may not always be required to ensure models remain calibrated over time. Recalibration is easier to undertake and requires less infrastructure than Bayesian modelling, for example, it does not require continuous data streams. If there is sufficient infrastructure in place, then the results suggest Bayesian modelling should be used for clinical prediction modelling but if the infrastructure is not available then recalibration is sufficient.

While recalibration through periodic updating was shown to be sufficient here, there is no guarantee that this will be true for all prediction models. Prediction models should continuously be assessed when they are used in clinical practice to ensure they remain safe and accurate. For this to be achieved a suitable infrastructure need to be in place that allows for regular monitoring of CPMs[87,90]. Hence, there should be further development of the infrastructure which will enable implementation and monitoring of prediction models. This will also enable implementation of Bayesian and other more complex models to be implemented across healthcare.

5.5.2 Previous work

Chapter 4 showed that varying coefficient models outperformed Bayesian updating models and time-invariant regression models in a simulation study. However, little difference in model performance was found when applying the modelling approaches to a clinical data set, except where they showed the Bayesian model performance was more stable over time. Our results support the finding of less variability in model performance in the Bayesian model. However, the varying coefficient model performance was worse than the time-invariant logistic model and the Bayesian and yearly updating models were preferred in this study. Hickey et al[47] also applied Bayesian modelling to a cardiac data set and compared the performance to a time-invariant model. Sample size was not considered when choosing a forgetting factor. Instead, five choices of forgetting were arbitrarily chosen, and they concluded a forgetting factor of 0.9 was sufficient to yield a smooth model fit. They also compare the model coefficients over time but did not perform model validation and noted the extra complexity of doing this in a dynamic framework. Our study considers sample size and we ensured adequate sample size based upon the criteria outlined by Riley et al[24]. We have also validated each model using prequential validation[86]. Su et al[57] and Siregar et al[56] also compare model updating methods, including Bayesian model updating, to time-invariant model's. Su et el[57] did not find a single method to outperform another but Siregar found model updating to be preferred over time-invariant models. However, neither study consider varying coefficient models and other than the study by Jenkins et al (chapter 4) this is the only other study to compare Bayesian updating, varying coefficient and time-invariant logistic models for use in healthcare.

5.5.3 Strengths and limitations

Our study has several strengths: 1) we had a large real-world data set currently used to inform clinical decisions and closely represent development and validation of the modelling methods in practice; 2) we compare the model performances each year and over the complete validation data where most studies only consider the latter.

We acknowledge some limitations of our work: 1) the cohort consists of data from a single hospital and this could result in selection bias. We only had access to data from a single hospital and the study was designed to compare methods rather than develop generalisable models. However, care needs to be taken when interpreting results as it is unclear the effect this could have on generalisability of findings; 2) although we consider a forgetting factor for the Bayesian model, this might not be the optimum choice. However, we use a novel approach to determine the forgetting factor that meets sample size requirements and increasing forgetting closer to 1 would result in performance closer to the GLM model by definition. Further methodological work is needed to optimise models with respect to forgetting and 3) we choose to include the variables included in EuroSCORE rather than derive models and perform model selection in our data. The models are therefore not likely to be the optimum model for each method. However, we did this to ensure the models were comparable and closely represent an existing CPM used in clinical practice.

5.6 Conclusion

Not considering temporal changes in data when developing a clinical prediction model can lead to suboptimal performance. We found Bayesian updating models to be the best performing model overall, especially with respect to discrimination, but the less complex periodic model recalibration method also outperformed a time-invariant logistic model. Time-variant models should be developed for use in healthcare but the infrastructure needs to be available to implement the more complex methods into systems.

5.7 References

1. Murdoch, T. B. & Detsky, A. S. The Inevitable Application of Big Data. *Jama* **309**, 1351–1352 (2014).

2. Wang, Y., Kung, L. A. & Byrd, T. A. Big data analytics: Understanding its capabilities and potential benefits for healthcare organizations. *Technol. Forecast. Soc. Change* **126**, 3–13 (2018).

3. Raghupathi, W. & Raghupathi, V. Big data analytics in healthcare: promise and potential. *Heal. Inf. Sci. Syst.* **2**, 3 (2014).

4. Friedman, C. P., Wong, A. K. & Blumenthal, D. Achieving a Nationwide Learning Health System. *Public Law* **2**, 8–10 (2010).

5. Friedman, C. P., Rubin, J. C. & Sullivan, K. J. Toward an Information Infrastructure for Global Health Improvement. *Yearb. Med. Inform.* **26**, 16–23 (2017).

6. J., A. & I., B. Combining health data uses to ignite health system learning. *Methods Inf. Med.* **54**, 479–487 (2015).

7. Kent, P., Cancelliere, C., Boyle, E., Cassidy, J. D. & Kongsted, A. A conceptual framework for prognostic research. *BMC Med. Res. Methodol.* **7**, 1–13 (2020).

8. Hemingway, H. *et al.* Prognosis research strategy (PROGRESS) 1: A framework for researching clinical outcomes. *BMJ* **346**, 1–11 (2013).

9. Riley, R. D. *et al.* Prognosis Research Strategy (PROGRESS) 2: Prognostic Factor Research. *PLoS Med.* **10**, e1001380 (2013).

10. Steyerberg, E. *et al.* Prognosis research strategy (PROGRESS) 3: prognostic model research. *PLoS Med* **10**, e1001381 (2013).

11. Hingorani, A. D. *et al.* Prognosis research strategy (PROGRESS) 4: Stratified medicine research. *BMJ* **346**, 1–9 (2013).

12. Riley, R. D., Windt, D. Van Der & Moons, K. G. M. Prognosis Research in Health Care. *Progn. Res. Heal. Care* 1–11 (2019). doi:10.1093/med/9780198796619.001.0001

13. Steyerberg, E. W. *Clinical Prediction Models. Statistics for Biology and Health. 2nd edition*. (2019).

14. Hemingway, H. *et al.* Prognosis research strategy (PROGRESS) 1: A framework for researching clinical outcomes. *BMJ* **346**, 1–11 (2013).

15. Gray, L. J. *et al.* The Leicester Risk Assessment score for detecting undiagnosed Type 2 diabetes and impaired glucose regulation for use in a multiethnic UK setting. *Diabet. Med.* **27**, 887–895 (2010).

16. (NICE), N. I. for H. and C. E. Type 2 diabetes: prevention in people at high risk. *Clinical guidelines [PH38]* (2012). Available at: https://www.nice.org.uk/guidance/ph38.

17. McAllister, K. S. L. *et al.* A contemporary risk model for predicting 30-day mortality following percutaneous coronary intervention in England and Wales. *Int. J. Cardiol.* **210**, 125–132 (2016).

18. Damen, J. A. A. G. *et al.* Prediction models for cardiovascular disease risk in the general population: Systematic review. *BMJ* **353**, (2016).

19. D'Agostino, R. B. *et al.* General Cardiovascular Risk Profile for Use in Primary Care: The Framingham Heart Study. *Circulation* **117**, 743–753 (2008).

20. Hippisley-Cox, J. *et al.* Derivation and validation of QRISK, a new cardiovascular disease risk score for the United Kingdom: Prospective open cohort study. *Br. Med. J.* **335**, 136–141 (2007).

21. (NICE), N. I. for H. and C. E. Cardiovascular disease: risk assessment and reduction, including lipid modification. *NICE Guidel. [CG181]* (2014).

22. Hippisley-Cox, J., Coupland, C., Robson, J. & Brindle, P. Derivation, validation, and evaluation of a new QRISK model to estimate lifetime risk of cardiovascular disease: Cohort study using QResearch database. *Bmj* **342**, 93 (2011).

23. Bonnett, L. J., Snell, K. I. E., Collins, G. S. & Riley, R. D. Guide to presenting clinical prediction models for use in clinical settings. *BMJ* **365**, 1–8 (2019).

24. Riley, R. D. *et al.* Minimum sample size for developing a multivariable prediction model: PART II - binary and time-to-event outcomes. *Stat. Med.* **38**, 1276–1296 (2019).

25. Steyerberg, E. W. *Clinical prediction models: a practical approach to development, validation, and updating*. (2008).

26. Harrell Jr, F. E. *Regression Modeling Strategies : With Applications to Linear Models, Logistic and Ordinal Regression, and Survival Analysis in Springer Series in Statistics*. *Springer* (Springer, 2015).

27. Steyerberg, E. W. *et al.* Internal validation of predictive models: Efficiency of some procedures for logistic regression analysis. *J. Clin. Epidemiol.* **54**, 774–781 (2001).

28. Steyerberg, E. W. *et al.* Assessing the performance of prediction models: a framework for some traditional and novel measures. *Epidemiology* **21**, 128–138 (2013).

29. Alba, A. C. *et al.* Discrimination and calibration of clinical prediction models: Users' guides to the medical literature. *JAMA - J. Am. Med. Assoc.* **318**, 1377–1384 (2017).

30. Van Calster, B. *et al.* Calibration: The Achilles heel of predictive analytics. *BMC Med.* **17**, 1–7 (2019).

31. Steyerberg, E. W. & Vergouwe, Y. Towards better clinical prediction models: Seven steps for development and an ABCD for validation. *Eur. Heart J.* **35**, 1925–1931 (2014).

32. Kramer, A. A. & Zimmerman, J. E. Assessing the calibration of mortality benchmarks in critical care: The Hosmer-Lemeshow test revisited. *Crit. Care Med.* **35**, 2052–2056 (2007).

33. Hickey, G. L. *et al.* Dynamic trends in cardiac surgery: Why the logistic euroscore is no longer suitable for contemporary cardiac surgery and implications for future risk models. *Eur. J. Cardio-thoracic Surg.* **43**, 1146–1152 (2013).

34. Davis, S. E., Lasko, T. A., Chen, G., Siew, E. D. & Matheny, M. E. Calibration drift in regression and machine learning models for acute kidney injury. *J. Am. Med. Informatics Assoc.* **24**, 1052–1061 (2017).

35. van Houwelingen, H. C. & Thorogood, J. Construction, validation and updating of a prognostic model for kidney graft survival. *Stat. Med.* **14**, 1999–2008 (1995).

36. Janssen, K. J. M., Moons, K. G. M., Kalkman, C. J., Grobbee, D. E. & Vergouwe, Y. Updating

methods improved the performance of a clinical prediction model in new patients. *J. Clin. Epidemiol.* **61**, 76–86 (2008).

37. Debray, T. P. A., Koffijberg, H., Vergouwe, Y., Moons, K. G. M. & Steyerberg, E. W. Aggregating published prediction models with individual participant data: A comparison of different approaches. *Stat. Med.* **31**, 2697–2712 (2012).

38. Martin, G. P., Mamas, M. A., Peek, N., Buchan, I. & Sperrin, M. Clinical prediction in defined populations: a simulation study investigating when and how to aggregate existing models. *BMC Med. Res. Methodol.* **17**, 1 (2017).

39. Martin, G. P., Mamas, M. A., Peek, N., Buchan, I. & Sperrin, M. A multiple-model generalisation of updating clinical prediction models. *Stat. Med.* **37**, 1343–1358 (2018).

40. Steyerberg, E. W., Borsboom, G. J. J. M., van Houwelingen, H. C., Eijkemans, M. J. C. & Habbema, J. D. F. Validation and updating of predictive logistic regression models: A study on sample size and shrinkage. *Stat. Med.* **23**, 2567–2586 (2004).

41. Toll, D. B., Janssen, K. J. M., Vergouwe, Y. & Moons, K. G. M. Validation, updating and impact of clinical prediction rules: A review. *J. Clin. Epidemiol.* **61**, 1085–1094 (2008).

42. Hippisley-Cox, J., Coupland, C. & Brindle, P. Development and validation of QRISK3 risk prediction algorithms to estimate future risk of cardiovascular disease : prospective cohort study. **2099**, 1–21 (2017).

43. Nashef, S. A. M. *et al.* European system for cardiac operative risk evaluation (EuroSCORE). *Eur. J. Cardio-thoracic Surg.* **16**, 9–13 (1999).

44. Nashef, S. A. M. *et al.* Euroscore II. *Eur. J. Cardio-thoracic Surg.* **41**, 734–745 (2012).

45. Raftery, A. E. & Ettler, P. Online Prediction Under Model Uncertainty via Dynamic Model Averaging : Application to a Cold Rolling Mill. **52**, 52–66 (2010).

46. Bull, L. M., Lunt, M., Martin, G. P., Hyrich, K. & Sergeant, J. C. Harnessing repeated measurements of predictor variables for clinical risk prediction: a review of existing methods. *Diagnostic Progn. Res.* **4**, (2020).

47. Hickey, G. L. *et al.* Dynamic prediction modeling approaches for cardiac surgery. *Circ. Cardiovasc. Qual. Outcomes* **6**, 649–658 (2013).

48. Mccormick, T. H., Raftery, A. E., Madigan, D. & Burd, R. S. Dynamic Logistic Regression and Dynamic Model Averaging for Binary Classification. *Biometrics* **68**, 23–30 (2012).

49. McCormick, T. H., Raftery, A. & Madigan, D. dma: Dynamic Model Averaging. (2018).

50. Five Year Forward View. (2014).

51. Salive, M. E. Multimorbidity in older adults. *Epidemiol. Rev.* **35**, 75–83 (2013).

52. Divo, M. J., Martinez, C. H. & Mannino, D. M. Ageing and the epidemiology of multimorbidity. *Eur. Respir. J.* **44**, 1055–1068 (2014).

53. Watkins, J. *et al.* Effects of health and social care spending constraints on mortality in England: a time trend analysis. *BMJ Open* **7**, e017722 (2017).

54. Abu-Hanna, A. & Lucas, P. J. F. Prognostic Models in Medicine. AI and Statistical Approaches.

Method Inf. Med **40**, 1–5 (2001).

55. Damen, J. A. A. G. *et al.* Prediction models for cardiovascular disease risk in the general population: systematic review. *Bmj* i2416 (2016). doi:10.1136/bmj.i2416

56. Siregar, S. *et al.* Improved Prediction by Dynamic Modeling. *Circ. Cardiovasc. Qual. Outcomes* **9**, 171–181 (2016).

57. Su, T.-L., Jaki, T., Hickey, G. L., Buchan, I. & Sperrin, M. A review of statistical updating methods for clinical prediction models. *Stat. Methods Med. Res.* 1–16 (2016). doi:10.1177/0962280215626466

58. van Houwelingen, H. C. & Thorogood, J. Construction, validation and updating of a prognostic model for kidney graft survival. *Stat. Med.* **14**, 1999–2008 (1995).

59. Draper, N. R., Nostrand, R. C. Van & Draper, N. R. Ridge Regression and James-Stein Estimation : Review and Comments Linked references are available on JSTOR for this article : Ridge Regression and James-Stein Estimation : Review and Comments. **21**, 451–466 (2016).

60. Copas, J. . Regression, Prediction and Shrinkage. *R. Stat. Soc.* **45**, 311–354 (1983).

61. Finkelman, B. S., French, B. & Kimmel, S. E. The prediction accuracy of dynamic mixed-effects models in clustered data. *BioData Min.* **9**, 5 (2016).

62. Fan, J. & Zhang, W. Statistical Methods with Varying Coefficient Models. *Stat Interface* **1**, 179–195 (2008).

63. Hoover, D. R., Rice, J. A., Wu, C. O. & Yang, L.-P. Nonparametric smoothing estimates of time-varying coefficient models with longitudinal data. *Biometrika* **85**, 809–822 (1998).

64. Madigan, D. & Raftery, A. E. Model Selection and Accounting for Model Uncertainty in Graphical Models Using Occam's Window. (1991).

65. Onorante, L. & Raftery, A. E. Dynamic model averaging in large model spaces using dynamic Occam's window. *Eur. Econ. Rev.* **81**, 2–14 (2016).

66. Ohata, T., Kaneko, M., Kuratani, T., Ueda, H. & Shimamura, K. Using the EuroSCORE to assess changes in the risk profiles of the patients undergoing coronary artery bypass grafting before and after the introduction of less invasive coronary surgery. *Ann. Thorac. Surg.* **80**, 131–135 (2005).

67. Ramsay, J. O., Wickham, H., Graves, S. & Hooker, G. fda: Functional Data Analysis. (2017).

68. Altman, D. G. & Royston, P. What do we mean by validating a prognistic model? *Stat. Med.* **19**, 453–473 (2000).

69. Justice, A. C., Covinsky, K. E. & Berlin, J. A. Assessing the generalizability of prognostic information. *Ann. Intern. Med.* **130**, 515–524 (1999).

70. Zimmerman, J. E., Kramer, A. A., McNair, D. S. & Malila, F. M. Acute Physiology and Chronic Health Evaluation (APACHE) IV: Hospital mortality assessment for today's critically ill patients. *Crit. Care Med.* **34**, 1297–1310 (2006).

71. Vergouwe, Y. *et al.* A closed testing procedure to select an appropriate method for updating prediction models. *Stat. Med.* **36**, 4529–4539 (2017).

72. Hafkamp-De Groen, E. *et al.* Predicting asthma in preschool children with asthma-like symptoms: Validating and updating the PIAMA risk score. *J. Allergy Clin. Immunol.* **132**, (2013).

73. Genders, T. S. S. *et al.* A clinical prediction rule for the diagnosis of coronary artery disease: Validation, updating, and extension. *Eur. Heart J.* **32**, 1316–1330 (2011).

74. Moons, K. G. M. *et al.* Risk prediction models: II. External validation, model updating, and impact assessment. *Heart* **98**, 691–698 (2012).

75. Hippisley-Cox, J., Coupland, C. & Brindle, P. Development and validation of QRISK3 risk prediction algorithms to estimate future risk of cardiovascular disease: Prospective cohort study. *BMJ* **357**, 1–21 (2017).

76. Martin, G. P., Sperrin, M. & Sotgiu, G. Performance of Prediction Models for Covid-19: The Caudine Forks of the External Validation. *Eur. Respir. J.* 2003728 (2020). doi:10.1183/13993003.03728-2020

77. Wynants, L. *et al.* Prediction models for diagnosis and prognosis of covid-19: Systematic review and critical appraisal. *BMJ* **369**, (2020).

78. Vergouwe, Y., Moons, K. G. M. & Steyerberg, E. W. External validity of risk models: Use of benchmark values to disentangle a case-mix effect from incorrect coefficients. *Am. J. Epidemiol.* **172**, 971–980 (2010).

79. Debray, T. P. A. *et al.* A guide to systematic review and meta-analysis of prediction model performance. *BMJ* **356**, (2017).

80. Luijken, K. *et al.* Changing predictor measurement procedures affected the performance of prediction models in clinical examples. *J. Clin. Epidemiol.* **119**, 7–18 (2020).

81. Riley, R. D. *et al.* External validation of clinical prediction models using big datasets from e-health records or IPD meta-analysis: Opportunities and challenges. *BMJ* **353**, 27–30 (2016).

82. Debray, T. P. A., Riley, R. D., Rovers, M. M., Reitsma, J. B. & Moons, K. G. M. Individual Participant Data (IPD) Meta-analyses of Diagnostic and Prognostic Modeling Studies: Guidance on Their Use. *PLoS Med.* **12**, 1–12 (2015).

83. Hippisley-Cox, J. *et al.* Predicting cardiovascular risk in England and Wales: Prospective derivation and validation of QRISK2. *Bmj* **336**, 1475–1482 (2008).

84. Jenkins, D. A., Sperrin, M., Martin, G. P. & Peek, N. Dynamic models to predict health outcomes: current status and methodological challenges. *Diagnostic Progn. Res.* **2**, 23 (2018).

85. Halabi, S. *et al.* Updated prognostic model for predicting overall survival in first-line chemotherapy for patients with metastatic castration-resistant prostate cancer. *J. Clin. Oncol.* **32**, 671–677 (2014).

86. Dawid, A. Present Position and Potential Developments : Some Personal Views : Statistical Theory : The Prequential Approach Author (s): A . P . Dawid Source : Journal of the Royal Statistical Society . Series A (General), Vol . 147 , No . 2 , The 150th Annivers. *J. R. Stat. Soc. Ser. A* **147**, 278–292 (1984).

87. Lenert, M. C., Matheny, M. E. & Walsh, C. G. Prognostic models will be victims of their own success, unless…. *J. Am. Med. Inform. Assoc.* **26**, 1645–1650 (2019).

88. Adibi, A., Sadatsafavi, M. & Ioannidis, J. P. A. Validation and Utility Testing of Clinical Prediction

Models. *JAMA* **2004**, (2020).

89. Booth, S., Riley, R. D., Ensor, J., Lambert, P. C. & Rutherford, M. J. Temporal recalibration for improving prognostic model development and risk predictions in settings where survival is improving over time. *Int. J. Epidemiol.* 1–10 (2020). doi:10.1093/ije/dyaa030

90. Jenkins, D. A. *et al.* Continual updating and monitoring of clinical prediction models: time for dynamic prediction systems? *Diagnostic Progn. Res.* **5**, 1–7 (2021).

91. Mccormick, T. H. *et al.* Dynamic Logistic Regression and Dynamic Model Averaging for Binary Classification. 23–30 (2012). doi:10.1111/j.1541-0420.2011.01645.x

92. R Core Team. R: A language and environment for statistical computing. R Foundation for Statistical. (2014).

93. Martin, G. P., Riley, R. D., Collins, G. S. & Sperrin, M. Developing clinical prediction models when adhering to minimum sample size recommendations: The importance of quantifying bootstrap variability in tuning parameters and predictive performance. *Stat. Methods Med. Res.* **30**, 2545–2561 (2021).

94. Sperrin, M., Jenkins, D., Martin, G. P. & Peek, N. Explicit causal reasoning is needed to prevent prognostic models being victims of their own success. *J. Am. Med. Informatics Assoc.* **26**, 1675–1676 (2019).

95. Hickey, G. L. *et al.* Clinical registries: Governance, management, analysis and applications. *Eur. J. Cardio-thoracic Surg.* **44**, 605–614 (2013).

96. Moons, K. G. M. *et al.* Risk prediction models: II. External validation, model updating, and impact assessment. *Heart* **98**, 691–698 (2012).

97. Davis, S. E. *et al.* A nonparametric updating method to correct clinical prediction model drift. *J. Am. Med. Informatics Assoc.* **26**, 1448–1457 (2019).

98. Minne, L. *et al.* Statistical process control for monitoring standardized mortality ratios of a classification tree model. *Methods Inf. Med.* **51**, 353–358 (2012).

99. Statistics, M. SUMS OF NONINDEPENDENT BERNOULLI RANDOM VARIABLES Author (s): Jose M . Gonzalez-Barrios Source : Brazilian Journal of Probability and Statistics , JUNE 1998 , Vol . 12 , No . 1 (JUNE Published by : Institute of Mathematical Statistics Stable URL : http. **12**, 55–64 (1998).

100. Koetsier, A., De Keizer, N. F., De Jonge, E., Cook, D. A. & Peek, N. Performance of risk-adjusted control charts to monitor in-hospital mortality of intensive care unit patients: A simulation study. *Crit. Care Med.* **40**, 1799–1807 (2012).

101. Team, R. core. R: A language and environment for statistical computing. *R Foundation for Statistical Computing, Vienna, Austria.* (2021). Available at: https://www.r-project.org/.

102. Albert, A. A. *et al.* On-line variable live-adjusted displays with internal and external risk-adjusted mortalities. A valuable method for benchmarking and early detection of unfavourable trends in cardiac surgery. *Eur. J. Cardio-thoracic Surg.* **25**, 312–319 (2004).

103. Lovegrove, J., Valencia, O., Treasure, T., Sherlaw-Johnson, C. & Gallivan, S. Monitoring the results of cardiac surgery by variable life-adjusted display. *Lancet* **350**, 1128–1130 (1997).

104. Poloniecki, J., Valencia, O. & Littlejohns, P. Correction: Cumulative risk adjusted mortality chart

for detecting changes in death rate: Observational study of heart surgery (British Medical Journal (1998) (1697-700)). *Br. Med. J.* **316**, 1947 (1998).

105. Minne, L. *et al.* Effect of changes over time in the performance of a customized SAPS-II model on the quality of care assessment. *Intensive Care Med.* **38**, 40–46 (2012).

106. Shi, X., Gallagher, C., Lund, R. & Killick, R. A Comparison of Single and Multiple Changepoint Techniques for Time Series Data. 1–29 (2021). doi:10.1016/j.csda.2022.107433

107. Wittenberg, P., Gan, F. F. & Knoth, S. A simple signaling rule for variable life-adjusted display derived from an equivalent risk-adjusted CUSUM chart. *Stat. Med.* **37**, 2455–2473 (2018).

108. Davis, S. E., Greevy, R. A., Lasko, T. A., Walsh, C. G. & Matheny, M. E. Detection of calibration drift in clinical prediction models to inform model updating. *J. Biomed. Inform.* **112**, 103611 (2020).

109. Riley, R. D. *et al.* Minimum sample size for external validation of a clinical prediction model with a binary outcome. *Stat. Med.* **40**, 4230–4251 (2021).

110. Steyerberg, E. W., Borsboom, G. J. J. M., van Houwelingen, H. C., Eijkemans, M. J. C. & Habbema, J. D. F. Validation and updating of predictive logistic regression models: A study on sample size and shrinkage. *Stat. Med.* **23**, 2567–2586 (2004).

111. Grigg, O. A., Farewell, V. T. & Spiegelhalter, D. J. Use of risk-adjusted CUSUM and RSPRT charts for monitoring in medical contexts. *Stat. Methods Med. Res.* **12**, 147–170 (2003).

112. Pagel, C. *et al.* Real time monitoring of risk-adjusted paediatric cardiac surgery outcomes using variable life-adjusted display: Implementation in three UK centres. *Heart* **99**, 1445–1450 (2013).

113. Barrett, J. & Su, L. Dynamic predictions using flexible joint models of longitudinal and time-to-event data. *Stat. Med.* **36**, 1447–1460 (2017).

114. van Houwelingen, H., & Putter, H. *Dynamic Prediction in Clinical Survival Analysis*. (2012). doi:https://doi.org/10.1201/b11311

5.8 Supplementary material

Supplementary Table S5.1 - Model coefficients for the logistic regression and varying coefficient models

	Logistic regression model	Varying coefficient model
Intercept	-7.292	-7.943
Age	0.042	0.042
Sex (Male)	0.218	0.339
Diabetes	-0.262	-0.869
History of Pulmonary Disease	0.221	0.339
Neurological Dysfunction	0.563	0.630
Critical preoperative state	0.663	0.128
Angina	-1.381	-0.982
Extracardiac arteriopathy	0.729	0.874
Previous surgery	0.904	1.343
Extracardiac arteriopathy	0.690	0.494
LV function (Moderate)	0.180	0.261
LV function (Poor)	0.254	0.690
MI < 90 days before surgery	-0.122	-0.157
Urgency (Emergency)	2.582	2.108
Urgency (Urgent)	0.905	0.857
Surgery on thoracic aorta	1.051	1.155

Supplementary Table S5.2 - Model coefficients for each version for the yearly updated logistic regression model. The column 2013 represents the model updated using data from 1st January 2012 to 31st December 2012 and was applied to all patients from 1st January 2013 to 31st Dec

	2012	2013	2014	2015	2016	2017	2018	2019
Intercept	-7.292	-6.681	-7.361	-8.133	-7.069	-7.966	-7.739	-7.149
Age	0.042	0.037	-0.237	-0.302	-0.235	-0.292	-0.283	-0.261
Sex (Male)	0.218	0.193	-0.058	-0.074	-0.057	-0.071	-0.069	-0.064
Diabetes	-0.262	-0.231	-0.546	-0.696	-0.541	-0.672	-0.652	-0.603
History of Pulmonary Disease	0.221	0.195	-0.055	-0.070	-0.054	-0.067	-0.065	-0.061
Neurological Dysfunction	0.563	0.497	0.293	0.374	0.291	0.361	0.350	0.324
Critical preoperative state	0.663	0.585	0.394	0.503	0.391	0.485	0.471	0.435
Angina	-1.381	-1.220	-1.685	-2.149	-1.671	-2.075	-2.013	-1.861
Extracardiac arteriopathy	0.729	0.644	0.462	0.590	0.458	0.569	0.552	0.510
Previous surgery	0.904	0.798	0.640	0.816	0.635	0.788	0.764	0.707
Extracardiac arteriopathy	0.690	0.609	0.422	0.538	0.418	0.520	0.504	0.466
LV function (Moderate)	0.180	0.159	-0.096	-0.123	-0.096	-0.119	-0.115	-0.106
LV function (Poor)	0.254	0.224	-0.022	-0.028	-0.022	-0.027	-0.026	-0.024
MI < 90 days before surgery	-0.122	-0.107	-0.404	-0.515	-0.400	-0.497	-0.482	-0.446
Urgency (Emergency)	2.582	2.2800	2.347	2.993	2.327	2.890	2.803	2.592
Urgency (Urgent)	0.905	0.799	0.641	0.818	0.636	0.790	0.766	0.708
Surgery on thoracic aorta	1.051	0.928	0.789	1.006	0.782	0.972	0.942	0.871

Supplementary Table S5.3 – Bayesian updating model coefficients on 1st January each year in the validation data

	2012	2013	2014	2015	2016	2017	2018	2019
Intercept	-7.292	-5.827	-5.582	-4.894	-5.041	-4.92	-5.084	-5.183
Age	0.042	0.024	0.021	0.014	0.016	0.012	0.017	0.019
Sex (Male)	0.218	0.571	0.224	0.152	0.024	0.432	0.465	0.623
Diabetes	-0.262	-0.14	-0.057	-0.322	-0.132	-0.164	-0.265	-0.08
History of Pulmonary Disease	0.221	0.148	0.129	-0.031	-0.1	0.057	-0.036	0.407
Neurological Dysfunction	0.563	0.194	0.347	0.403	0.236	0.165	0.102	-0.06
Critical preoperative state	0.663	1.19	1.711	1.865	2.147	2.45	2.343	2.446
Angina	-1.381	-1.9	0.078	-0.155	0.035	0.154	0.29	0.255
Extracardiac arteriopathy	0.729	0.632	0.484	0.242	0.116	0.11	0.211	0.51
Previous surgery	0.904	0.566	0.505	0.183	0.725	0.754	0.874	0.824
Extracardiac arteriopathy	0.69	1.07	0.643	0.441	0.381	0.228	-0.115	-0.083
LV function (Moderate)	0.18	0.105	0.28	0.103	0.131	0.01	0.175	-0.018
LV function (Poor)	0.254	0.47	0.502	0.372	0.362	0.225	0.341	0.174
MI < 90 days before surgery	-0.122	-0.132	0.018	-0.3	-0.317	-0.431	-0.244	-0.226
Urgency (Emergency)	2.582	2.024	1.392	1.337	1.185	0.679	0.573	0.393
Urgency (Urgent)	0.905	0.55	0.17	0.242	0.238	-0.036	-0.062	-0.066
Surgery on thoracic aorta	1.051	0.853	0.734	0.87	0.625	0.725	0.533	0.406

Chapter 6 Use of statistical process control to monitor the performance of a clinical prediction model

David A Jenkins, Glen P Martin, Niels Peek and Matthew Sperrin

In preparation for submission

6.1 Abstract

6.1.1 Background

Clinical prediction models (CPMs) are useful tools to diagnose current outcomes or predict future outcomes in individuals, based on what is known about that individual and their environment. Once developed, a CPM will remain fixed, meaning that its coefficients are time invariant post development. CPM updating methods, such as the logistic recalibration framework, are well established but a key challenge is that it is unclear when a model should be updated; as such, this is currently done at arbitrary time points. However, there is opportunity to address this problem because infrastructure now exist to collect data in real-time and this provides opportunity to monitor a CPM and determine when it needs updating.

6.1.2 Methods

We analytically describe the use of statistical process control (SPC) to continuously monitor a CPM's predictive performance, specifically the sum of the observed minus expected, and generate an alert when the performance reaches some predefined threshold (control limit). We performed a simulation study based on an existing CPM for 30-day mortality after percutaneous coronary intervention. We used the observed covariate data from a real-world dataset, but simulated binary outcomes according to a prespecified data generating model. We generated simulated outcomes under varying degrees of miscalibration-in-the-large by adding some constant, between 0 and 2, to the model intercept. SPC was then used to monitor the model performance, and this was repeated 1000 times, using bootstrapping, for each miscalibration scenario. The time of an alert, generated from using 3 and 4 standard deviations as the control limits, was recorded for each iteration and degree of miscalibration before being averaged across iterations.

6.1.3 Results

When the intercept of the data generating model was miscalibrated by 0.2 the median and 95% quantile for the 3 and 4SD control limits across the 1000 iterations were 7510 (294 – 27136) and 16022 (1863 – 42160), respectively. Under the scenario when the data generating model was miscalibrated-in-the-large by 2, the median and 95% quantile for the 3 and 4SD control limits across the 1000 iterations were 246 (175 – 320) and 246 (175 – 321), respectively

6.1.4 Conclusion

Arbitrary updating is suboptimal and continual monitoring of CPMs is needed to ensure decisions are not being made based upon a miscalibrated model. SPC is a solution and can provide users an alert when miscalibration is detected.

6.2 Background

Clinical prediction models (CPMs) are tools that estimate the risk of an outcome given a set of patient characteristics that are available at the time of prediction[25]. Once deployed into practice, the model coefficients usually remain fixed. However, healthcare and patient populations are constantly evolving. As a result, predictions based on these models often become less accurate over time. Specifically, agreement between the observed and predicted event rates (i.e. calibration) may worsen over time[33]. This is known as calibration drift and is one of the major pitfalls with CPMs[34,90].

While common practice to address this is to develop another CPM de novo, model updating[35,36,71,97] methods are now well established and preferable because they do not discard historical data/information and previous research efforts[38]. Nonetheless, this updating is still relatively uncommon, often occurs a substantial time after model development and is often undertaken at arbitrary time points[90]. For example, EuroSCORE II[44] was developed 13 years after the original model in 2012, and we are still yet to see another updated version or investigation to determine if the updated EuroSCORE II has suffered from calibration drift. The problem with this approach is that the updates are chosen at arbitrary points in time or performed following a study showing the model has become miscalibrated, which means that incorrect decisions may have already been made as a result of the mis-calibrated model.

Conversely, a model may not need updating and could retain good performance over time. Hence, rather than updating at fixed time points, updating should be data driven so when there are changes in the system updating occurs more frequently, while relatively stable systems (or time periods) can afford less regular updates.

Although model updating methods are established, there remains limited literature on when to update a prediction model. Recent articles discuss the notion of model surveillance, whereby models are monitored continuously as they continue to make predictions on new data. Jenkins et al.[90] (Chapter 3) proposed the use of prequential testing with a feedback loop to monitor model performance in practice and determine when to update a model. However, this has yet to explored beyond the postulation of the idea. One possible solution to implement this idea would be through the use of statistical process control (SPC)[98]. SPC is a method that monitors and controls some process to ensure quality and efficiency and could be used to ensure a model is accurate and efficient (only updating when necessary). The method attempts to distinguish two types of variation: common cause variation and special cause variation. The latter indicates the process is out of statistical control while the former is intrinsic of the process being monitored and will always be present. Hence, SPC could monitor the performance of a prediction model and the feedback loop would be an alert to a user notifying them it has detected miscalibration. The user can then act upon this information.

This study aims to investigate the potential of SPC in clinical prediction modelling as a proof-of-concept. This will be achieved through the following two objectives. The first is to analytically describe the SPC method to monitor the performance of a CPM through time (section 6.3). The second is to investigate the behaviour of the SPC against known behaviour in terms of time-to-alert. We will do this through a simulation study comparing the amount of time taken to trigger an alert when a model is miscalibrated and the number of false positives when there is no miscalibration (section 6.4).

6.3 SPC derivation

For the SPC method to work in real-time, the SPC metric needs to be able to work on each individual new observation. Hence, avoiding metrics that rely on batches of data, for example, calibration-in-the-large, is desirable. We therefore propose to use SPC to monitor the sum of the observed minus expected.

Let 0_i be the observed outcome for an individual *i=(1,…,n)* and π_i be the predicted probability for a binary outcome Y_i, obtained from a prediction model we are monitoring, for individual *i=(1,…,n)* given a set of predictors X_i for each individual. It follows that for each individual, 0_i is a Bernoulli random variable, such that $E[O_i] = p_i$ and π_i is a constant. Suppose we want to monitor the sum of the observed minus expected,

$$\sum_{i=1}^{n} 0_i - \pi_i \, , where \; \pi_i = P(Y_i = 1 \mid X_i)$$

As we observe a new individual, *j*, we can derive the sum of the observed minus expected as,

$$\sum_{i=1}^{j} 0_i - \pi_i = \left(0_j - \pi_j\right) + \sum_{i=1}^{j-1} (0_i - \pi_i) \;\; where \; 0 < j < n$$

We continue to do this for each individual we observe over time, thus continuously monitoring performance of a prediction model as new data become available. As we do this, the SPC checks if the performance measure is within some predefined control limit. Typically, this is chosen to be ± 3 standard deviations away from the process mean. If the performance measure lies outside of this control limit, then the user is alerted, otherwise the SPC will do nothing and continue to monitor new data as they arrive.

For each individual i, the expectation of the observed minus expected equals zero, $E[0_i - \pi_i] = 0$, under the hypothesis that the model is calibrated-in-the-large. Hence, the process mean, for the sum of the observed minus expected, is 0. Also, $Var[0_i - \pi_i] = Var[0_i]$ as π_i is a constant. As 0_i is a Bernoulli random variable, such that $E[O_i] = p_i$, it follows that $Var[0_i] = p_i(1 - p_i)$. Under the null hypothesis that the prediction model we are monitoring in calibrated-in-the-large, $p_i = \pi_i$. Therefore, the variance for the $0_i - \pi_i$ for each individual i is $\pi_i(1 - \pi_i)$. We assume each observation is independent conditional on p_i, and therefore the covariances are zero. Hence, the variance is additive and it follows that,

$$Var\left(\sum_{i=1}^{n} 0_i - \pi_i\right) = \sum_{i=1}^{n} \pi_i \, (1 - \pi_i)$$

Using 3 standard deviations as the control limit to trigger an alert we can calculate control limits for each observation i as $0 \pm 3\sqrt{\sum_{i=1}^{n} \pi_i \, (1 - \pi_i)}$. Note that the control limit is therefore changing as we observe new data.

In addition to this, we propose a burn in period where an alert cannot be triggered. This is to stop unnecessary early alerts occurring and reduce the type I error rate. It is well known that the sum of a Bernoulli random variable with parameter *p* is binomial distributed with parameters *n* and *p*[99]. A common rule of thumb is that the normal approximation works well for a binomial approximation when $np > 5$ and $n(1-p) > 5$. Hence, we suggest waiting until the expected number of events, $\sum_{i=1}^{n} \pi_i$, exceeds 5 before an alert is allowed to occur as the sum of observed minus expected will be approximately normal. Finally, control limits will increase with each new observation and over time could result in miscalibration not being detected, particularly if the model performance measure remains stable for some period of time. Hence, if there is a change in performance it would not be detected for some time, or at all, due to the stability of the model in the past. When the sum of observed minus expected is zero, the model is well calibrated on average for all individuals in the monitoring data. Therefore, we suggest resetting the control limits, such that you forget all past data whenever the sum of observed minus expected crosses zero (sign changes). This will result in the expected number of events being below 5 and prevent an alert occurring for some time after the reset. To overcome this, we calculate control limits based upon all individuals since a previous reset and only using individuals after the reset. Until the expected number of events exceeds 5 after the reset, we continue to use the control limits that use data prior to the reset.

6.4 Simulation

6.4.1 Simulation setup

We followed the approach taken by Koetsier et al[100], where we sampled covariate data from a real world data set and generate outcomes through Monte Carlo simulation. The data set of interest was the British Cardiovascular Intervention Society (BCIS) registry which collects data on percutaneous coronary interventions (PCIs) undertaken across the UK, as described in chapter 4. There exists a model to predict 30-day mortality following PCI, which was developed in the BCIS data, that we used to determine covariate values and baseline risk, prior to adding any miscalibration, to ensure our simulation represents plausible real-world scenarios[17]. This model was used to determine the predicted risks and we simulated outcomes with varying degrees of miscalibration across simulations. For each simulation we recorded when the SPC approach generated an alert.

Data were randomly sampled with replacement from the PCI dataset and given an observation time $t = 1, \dots, n$, meaning we sequentially indexed individuals by the order of sampling. In this study n=1,088,567. The existing BCIS model[17] was used to generate the linear predictor for each observation at a given time t, denoted as LP_t. The linear predictor was then used as the predictor to which synthetic outcomes were generated using Bernoulli trials. We assume that each observation observed at time *t* had a binary outcome, y_t, which is mortality observed 30-days after baseline, which we index by *t* for convenience. We generated these outcomes such that:

$$g\big(E(y_t)\big) = \beta_0 + \beta_1 LP_t$$

Where *g* is the logit link, β_0 is the intercept and β_1 is the regression coefficient for the predictor, LP. We fixed $\beta_1 = 1$ across all simulation scenarios and varied β_0 across simulation scenarios. The value of β_0 represents the degree of miscalibration-in-the-large, corresponding to a change in the prevalence of the outcome between the data the model was developed and validated on, and we chose to vary this across simulations. The coefficients were chosen to be fixed, and not time dependent, for simplicity, making it easier to interpret results. This settup means that a model is either miscalibrated-in-the-large (if $\beta_0 \neq 0$) or is well calibrated (if $\beta_0 = 0$), and any such miscalibration occurs immediately at t=1. Therefore, our primary focus is on the time-until-alert, defined as the time difference between when an alert is triggered and time zero (first observation). Across simulations we chose β_0 to vary from 0 to 2 in increments of 0.1. All analysis was conducted in R[101] (version 3.6.2) and code can be found on Github (https://github.com/David-A-Jenkins/Thesis).

6.4.2 Simulation results

The average linear predictor in the raw BCIS data was -4.8, corresponding to an average risk of 0.8%. Therefore, when $\beta_0 = 2$, the average risk was 5.7%. In the original BCIS data there were 1,088,567 observations in the data with a median number of observations per year and per month of 88,179 and 7,348, respectively.

Figure 6.1 displays the median observation number/time and 95% quantile (shaded region) of when the alerts occurred across the 1000 simulations for each level of miscalibration-in-the-large. The figure also displays the results using the control limits of 3 and 4 standard deviations (SDs), separately. The data used in figure 6.1 is presented in Supplementary table S6.1. The median and 95% quantiles for the SPC alert occurred before the time of the final observation in the data for all levels of miscalibration except 0.1 and no miscalibration. When there was no miscalibration-in-the-large the 4SDs control limit SPC provided one alert out of the 1000 iterations compared to 100 of the 1000 iterations for the 3SDs control limit SPC. We could therefore not obtain median and 95% quantiles for the 4SD control limit simulation when there was no miscalibration ($\beta_0 = 0$), as denoted by the NAs in supplementary table S6.1. Only the lower quantile for the 3SD control limit was obtainable when there was no miscalibration ($\beta_0 = 0$).

The median and 95% quantile for the simulations where miscalibration-in-the-large was above 1.2 were similar. On average, when miscalibration-in-the-large was below 0.3, the 3SD control limit had a wider 95% quantile region and alerted sooner - on average - than the 4SD control limit, but little difference was observed in the width of the intervals for large miscalibration-in-the-large. For miscalibration-in-the-large of 1.4 of above, the median alert time for both the SPCs, with 3SDs and 4SDs, were within 6 observations of each other.

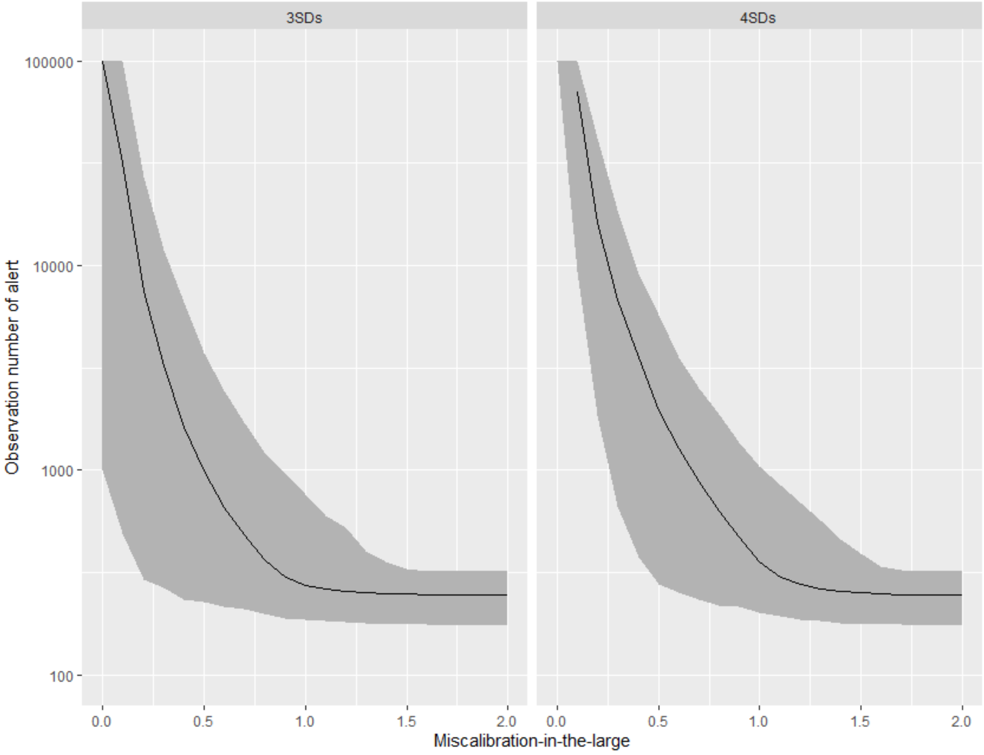


Figure 6.1 - Median and 95% quantile of the observation number an alert was triggered (y-axis) across the 1000 iterations for each value of miscalibration-in-the-large (x-axis). The left and right plot represent the results when the control limit was calculated us

6.5 Discussion

In this study we have described an approach to monitor the overall calibration of a CPM by using SPC to monitor the sum of observed minus expected and generate an alert when it detects miscalibration. We also illustrated the approach in synthetic health data and show the time until an alert for varying degrees of miscalibration-in-the-large. When there was no miscalibration-in-the-large the SPC method with a 4 standard deviation control limit provided only 1 alert before the end of the dataset out of 1000 iterations. When miscalibration-in-the-large was high (above 1) the 3 and 4 SD SPC median and 95% quantiles were below 1000, meaning miscalibration was detected within the first 1000 observations, which correspond to approximately 5 days in the BCIS data. As miscalibration increased to above 1.5, both control limits provided alerts within 400 observations. The 3SD SPC monitoring provides alerts earlier than using 4SDs but also provided more alerts when no miscalibration was present. Therefore, statistical process control shows promise for monitoring CPMs and alerting users of potential miscalibration, providing a data driven approach in place of arbitrary updating, but more research is needed to define the control limits.

In this study, we use SPC to provide an alert to a user and we are not suggesting SPC alone should be used to decide on when to update. Rather the method should be used to alert a user of possible miscalibration which will then need to be assessed and a decision made. It is important to note that different clinical areas, or uses, of a model in healthcare might have different levels of miscalibration they are willing to accept and as seen in the simulation, an alert can still be triggered when no calibration is present. Therefore, there needs to be investigation following an alert, that should include clinical input, and this could result in the decision to update the model, do nothing or monitor the model for longer.

In the simulation, there was little difference in using 3 or 4 SDs to define the control limits when miscalibration-in-the-large was high. The simulation results show a plateau in the time to alert for miscalibration-in-the-large above 1. This is likely due to the burn in period preventing alerts occurring earlier. Once the expectation is above 5 a trigger is occurring immediately as the sum of observed minus expected has been outside the control limit for some period of time already. This burn in prevents unnecessary alerts occurring early but appears to lead to the convergence of the time of an alert from the point of miscalibration above 1 and could result in a delay in an alert of miscalibration.

Under lower levels of miscalibration-in-the-large (less than 0.3), or no miscalibration, the 3SD control limit provided wider confidence intervals. Also, on average, the 3SD control limit alerted users earlier than using 4SDs. This could cause alert fatigue and there might not be enough resources available to keep checking the alerts. Alternatively, increasing the control limit will result in a delay in alerts when a model is miscalibrated-in-the-large. Therefore, care needs to be taken when choosing control limits to ensure alerts are meaningful and can be managed.

6.5.1 Previous literature

SPC has been well established in various healthcare fields, for example, cardiac surgery[102–104]. Albert et al[102] used a control chart known as variable life adjusted display (VLAD) for benchmarking and to detect unfavourable trend. They use the EuroSCORE model[43] as the predicted outcomes and use it to monitor net lives saved for individual surgeons. This monitors individual's surgeon and assumes the model is accurate. This is a typical use of SPC in the literature and limited research has used it to monitor the prediction models.

Minnie et al[105] used SPC to investigate the discrimination of a model that predicts mortality. They split the data into 30 equally sized subsets, computed the c-statistic for each subgroup and used SPC to assess if the model is stable. They used two control limits and if the measure fell outside both control limits they considered this to be 'critical'. If the C-statistic was between the two control limits they called this the 'warning' zone. Other studies using SPC also consider two control limits but these are arbitrarily chosen and it is unclear of what to do when in either of these zones. Although Minnie et al[105] use SPC, they require batches of data and monitoring therefore occurs at arbitrary time points compared to our study that continuously monitors performance. In addition to this, they use what is known as the PreControl chart where the zones are prespecified and a fixed width is used. This is often seen in time-series analysis for monitoring forecast error and detecting change points. SPC methods in time-series to detect change points is well established[106] and it is common to take a set of observations where you assume the process is stable and use this to calculate the control limits/zones. However, in our study we calculate the control limit for every observation that enables the control limit to vary over time. Fixed width zones are not reasonable because the variance of the cumulative sum increases with sample size[107].

Another recent article[108] proposed a method for the detection of calibration drift by deriving dynamic calibration curves using adaptive sliding windows. Davis et al[108] implement their approach in a simulation study and apply the method to a real-world data set. They test for a change in calibration by comparing a recent window with a previous window and do this continuously as new data arrive, whereas we are testing for miscalibration as new data arrive. It is plausible that a model may slowly drift and quickly revert back to being recalibrated. In this scenario the Davis approach could trigger an update when in fact it is not needed because it is looking for change in calibration rather than deciding to update based upon performance directly. Future work should consider comparing Davis' approach with ours and investigate if the methods differ in when they alert the user that a model update may be needed.

6.5.2 Study limitations

This study has several limitations. Recently, external validation sample size calculations[109] have been proposed for CPMs and this was not used. The calculation would almost always result in a longer burn-in period before an alert could be triggered but would ensure a more precise estimate of calibration is used to determine if an alert is triggered. The SPC approach is not intended to provide precise estimates of performance and is designed to detect potential miscalibration as soon as possible, hence we chose to use the expectation of 5. Another limitation is that we only consider observed minus expected. In addition to this, we only consider a limited number of simulated scenarios and all scenarios are miscalibrated at the intercept. Although this simulates a change in the prevalence of an outcome, the model could be miscalibrated in other ways. We also chose to simulate a step change in model intercept at time zero. In practice a model may become miscalibrated gradually over time. We chose this as it means we know the time an alert should occur and it is uncertain when an alert should occur if miscalibration is gradual. Finally, the simulation had a limited number of observations and the SPC did not always trigger an alert. If an alert was not triggered it could have been because there were not enough observations. However, this was a large data set that included many years of data and if an alert was not triggered then it considered the model to be calibrated for more than 1 million patients which in this setting is about 11 years.

6.5.3 Future work

Other performance measures, beyond observed minus expected, are also important[96] when evaluating a clinical prediction model, and depending on the application and clinical setting users may want to monitor these other performance measure, for example, calibration slope or discrimination. Also, calibration drift is often gradual, rather than a sudden step change[33], and does not always occur at the model intercept. Therefore, future work should consider investigating the behaviour of the SPC approach under other miscalibration scenarios and extending the SPC approach to include monitoring of other performance measures.

6.6 Conclusion

Statistical process control offers a way to continuously monitor and test for calibration drift in clinical prediction models. The approach is able to detect miscalibration in a timely manner and can be used to inform users that an update may be unnecessary. However, further work is needed to determine the control limit that should be used and to extend the approach to include monitoring of additional performance measures.

6.7 References

1. Murdoch, T. B. & Detsky, A. S. The Inevitable Application of Big Data. *Jama* **309**, 1351–1352 (2014).

2. Wang, Y., Kung, L. A. & Byrd, T. A. Big data analytics: Understanding its capabilities and potential benefits for healthcare organizations. *Technol. Forecast. Soc. Change* **126**, 3–13 (2018).

3. Raghupathi, W. & Raghupathi, V. Big data analytics in healthcare: promise and potential. *Heal. Inf. Sci. Syst.* **2**, 3 (2014).

4. Friedman, C. P., Wong, A. K. & Blumenthal, D. Achieving a Nationwide Learning Health System. *Public Law* **2**, 8–10 (2010).

5. Friedman, C. P., Rubin, J. C. & Sullivan, K. J. Toward an Information Infrastructure for Global Health Improvement. *Yearb. Med. Inform.* **26**, 16–23 (2017).

6. J., A. & I., B. Combining health data uses to ignite health system learning. *Methods Inf. Med.* **54**, 479–487 (2015).

7. Kent, P., Cancelliere, C., Boyle, E., Cassidy, J. D. & Kongsted, A. A conceptual framework for prognostic research. *BMC Med. Res. Methodol.* **7**, 1–13 (2020).

8. Hemingway, H. *et al.* Prognosis research strategy (PROGRESS) 1: A framework for researching clinical outcomes. *BMJ* **346**, 1–11 (2013).

9. Riley, R. D. *et al.* Prognosis Research Strategy (PROGRESS) 2: Prognostic Factor Research. *PLoS Med.* **10**, e1001380 (2013).

10. Steyerberg, E. *et al.* Prognosis research strategy (PROGRESS) 3: prognostic model research. *PLoS Med* **10**, e1001381 (2013).

11. Hingorani, A. D. *et al.* Prognosis research strategy (PROGRESS) 4: Stratified medicine research. *BMJ* **346**, 1–9 (2013).

12. Riley, R. D., Windt, D. Van Der & Moons, K. G. M. Prognosis Research in Health Care. *Progn. Res. Heal. Care* 1–11 (2019). doi:10.1093/med/9780198796619.001.0001

13. Steyerberg, E. W. *Clinical Prediction Models. Statistics for Biology and Health. 2nd edition*. (2019).

14. Hemingway, H. *et al.* Prognosis research strategy (PROGRESS) 1: A framework for researching clinical outcomes. *BMJ* **346**, 1–11 (2013).

15. Gray, L. J. *et al.* The Leicester Risk Assessment score for detecting undiagnosed Type 2 diabetes and impaired glucose regulation for use in a multiethnic UK setting. *Diabet. Med.* **27**, 887–895 (2010).

16. (NICE), N. I. for H. and C. E. Type 2 diabetes: prevention in people at high risk. *Clinical guidelines [PH38]* (2012). Available at: https://www.nice.org.uk/guidance/ph38.

17. McAllister, K. S. L. *et al.* A contemporary risk model for predicting 30-day mortality following percutaneous coronary intervention in England and Wales. *Int. J. Cardiol.* **210**, 125–132 (2016).

18. Damen, J. A. A. G. *et al.* Prediction models for cardiovascular disease risk in the general population: Systematic review. *BMJ* **353**, (2016).

19. D’Agostino, R. B. *et al.* General Cardiovascular Risk Profile for Use in Primary Care: The Framingham Heart Study. *Circulation* **117**, 743–753 (2008).

20. Hippisley-Cox, J. *et al.* Derivation and validation of QRISK, a new cardiovascular disease risk score for the United Kingdom: Prospective open cohort study. *Br. Med. J.* **335**, 136–141 (2007).

21. (NICE), N. I. for H. and C. E. Cardiovascular disease: risk assessment and reduction, including lipid modification. *NICE Guidel. [CG181]* (2014).

22. Hippisley-Cox, J., Coupland, C., Robson, J. & Brindle, P. Derivation, validation, and evaluation of a new QRISK model to estimate lifetime risk of cardiovascular disease: Cohort study using QResearch database. *Bmj* **342**, 93 (2011).

23. Bonnett, L. J., Snell, K. I. E., Collins, G. S. & Riley, R. D. Guide to presenting clinical prediction models for use in clinical settings. *BMJ* **365**, 1–8 (2019).

24. Riley, R. D. *et al.* Minimum sample size for developing a multivariable prediction model: PART II - binary and time-to-event outcomes. *Stat. Med.* **38**, 1276–1296 (2019).

25. Steyerberg, E. W. *Clinical prediction models: a practical approach to development, validation, and updating*. (2008).

26. Harrell Jr, F. E. *Regression Modeling Strategies : With Applications to Linear Models, Logistic and Ordinal Regression, and Survival Analysis in Springer Series in Statistics*. *Springer* (Springer, 2015).

27. Steyerberg, E. W. *et al.* Internal validation of predictive models: Efficiency of some procedures for logistic regression analysis. *J. Clin. Epidemiol.* **54**, 774–781 (2001).

28. Steyerberg, E. W. *et al.* Assessing the performance of prediction models: a framework for some traditional and novel measures. *Epidemiology* **21**, 128–138 (2013).

29. Alba, A. C. *et al.* Discrimination and calibration of clinical prediction models: Users’ guides to the medical literature. *JAMA - J. Am. Med. Assoc.* **318**, 1377–1384 (2017).

30. Van Calster, B. *et al.* Calibration: The Achilles heel of predictive analytics. *BMC Med.* **17**, 1–7 (2019).

31. Steyerberg, E. W. & Vergouwe, Y. Towards better clinical prediction models: Seven steps for development and an ABCD for validation. *Eur. Heart J.* **35**, 1925–1931 (2014).

32. Kramer, A. A. & Zimmerman, J. E. Assessing the calibration of mortality benchmarks in critical care: The Hosmer-Lemeshow test revisited. *Crit. Care Med.* **35**, 2052–2056 (2007).

33. Hickey, G. L. *et al.* Dynamic trends in cardiac surgery: Why the logistic euroscore is no longer suitable for contemporary cardiac surgery and implications for future risk models. *Eur. J. Cardio-thoracic Surg.* **43**, 1146–1152 (2013).

34. Davis, S. E., Lasko, T. A., Chen, G., Siew, E. D. & Matheny, M. E. Calibration drift in regression and machine learning models for acute kidney injury. *J. Am. Med. Informatics Assoc.* **24**, 1052–1061 (2017).

35. van Houwelingen, H. C. & Thorogood, J. Construction, validation and updating of a prognostic model for kidney graft survival. *Stat. Med.* **14**, 1999–2008 (1995).

36. Janssen, K. J. M., Moons, K. G. M., Kalkman, C. J., Grobbee, D. E. & Vergouwe, Y. Updating

methods improved the performance of a clinical prediction model in new patients. *J. Clin. Epidemiol.* **61**, 76–86 (2008).

37. Debray, T. P. A., Koffijberg, H., Vergouwe, Y., Moons, K. G. M. & Steyerberg, E. W. Aggregating published prediction models with individual participant data: A comparison of different approaches. *Stat. Med.* **31**, 2697–2712 (2012).

38. Martin, G. P., Mamas, M. A., Peek, N., Buchan, I. & Sperrin, M. Clinical prediction in defined populations: a simulation study investigating when and how to aggregate existing models. *BMC Med. Res. Methodol.* **17**, 1 (2017).

39. Martin, G. P., Mamas, M. A., Peek, N., Buchan, I. & Sperrin, M. A multiple-model generalisation of updating clinical prediction models. *Stat. Med.* **37**, 1343–1358 (2018).

40. Steyerberg, E. W., Borsboom, G. J. J. M., van Houwelingen, H. C., Eijkemans, M. J. C. & Habbema, J. D. F. Validation and updating of predictive logistic regression models: A study on sample size and shrinkage. *Stat. Med.* **23**, 2567–2586 (2004).

41. Toll, D. B., Janssen, K. J. M., Vergouwe, Y. & Moons, K. G. M. Validation, updating and impact of clinical prediction rules: A review. *J. Clin. Epidemiol.* **61**, 1085–1094 (2008).

42. Hippisley-Cox, J., Coupland, C. & Brindle, P. Development and validation of QRISK3 risk prediction algorithms to estimate future risk of cardiovascular disease : prospective cohort study. **2099**, 1–21 (2017).

43. Nashef, S. A. M. *et al.* European system for cardiac operative risk evaluation (EuroSCORE). *Eur. J. Cardio-thoracic Surg.* **16**, 9–13 (1999).

44. Nashef, S. A. M. *et al.* Euroscore II. *Eur. J. Cardio-thoracic Surg.* **41**, 734–745 (2012).

45. Raftery, A. E. & Ettler, P. Online Prediction Under Model Uncertainty via Dynamic Model Averaging : Application to a Cold Rolling Mill. **52**, 52–66 (2010).

46. Bull, L. M., Lunt, M., Martin, G. P., Hyrich, K. & Sergeant, J. C. Harnessing repeated measurements of predictor variables for clinical risk prediction: a review of existing methods. *Diagnostic Progn. Res.* **4**, (2020).

47. Hickey, G. L. *et al.* Dynamic prediction modeling approaches for cardiac surgery. *Circ. Cardiovasc. Qual. Outcomes* **6**, 649–658 (2013).

48. Mccormick, T. H., Raftery, A. E., Madigan, D. & Burd, R. S. Dynamic Logistic Regression and Dynamic Model Averaging for Binary Classification. *Biometrics* **68**, 23–30 (2012).

49. McCormick, T. H., Raftery, A. & Madigan, D. dma: Dynamic Model Averaging. (2018).

50. Five Year Forward View. (2014).

51. Salive, M. E. Multimorbidity in older adults. *Epidemiol. Rev.* **35**, 75–83 (2013).

52. Divo, M. J., Martinez, C. H. & Mannino, D. M. Ageing and the epidemiology of multimorbidity. *Eur. Respir. J.* **44**, 1055–1068 (2014).

53. Watkins, J. *et al.* Effects of health and social care spending constraints on mortality in England: a time trend analysis. *BMJ Open* **7**, e017722 (2017).

54. Abu-Hanna, A. & Lucas, P. J. F. Prognostic Models in Medicine. AI and Statistical Approaches.

Method Inf. Med **40**, 1–5 (2001).

55. Damen, J. A. A. G. *et al.* Prediction models for cardiovascular disease risk in the general population: systematic review. *Bmj* i2416 (2016). doi:10.1136/bmj.i2416

56. Siregar, S. *et al.* Improved Prediction by Dynamic Modeling. *Circ. Cardiovasc. Qual. Outcomes* **9**, 171–181 (2016).

57. Su, T.-L., Jaki, T., Hickey, G. L., Buchan, I. & Sperrin, M. A review of statistical updating methods for clinical prediction models. *Stat. Methods Med. Res.* 1–16 (2016). doi:10.1177/0962280215626466

58. van Houwelingen, H. C. & Thorogood, J. Construction, validation and updating of a prognostic model for kidney graft survival. *Stat. Med.* **14**, 1999–2008 (1995).

59. Draper, N. R., Nostrand, R. C. Van & Draper, N. R. Ridge Regression and James-Stein Estimation : Review and Comments Linked references are available on JSTOR for this article : Ridge Regression and James-Stein Estimation : Review and Comments. **21**, 451–466 (2016).

60. Copas, J. . Regression, Prediction and Shrinkage. *R. Stat. Soc.* **45**, 311–354 (1983).

61. Finkelman, B. S., French, B. & Kimmel, S. E. The prediction accuracy of dynamic mixed-effects models in clustered data. *BioData Min.* **9**, 5 (2016).

62. Fan, J. & Zhang, W. Statistical Methods with Varying Coefficient Models. *Stat Interface* **1**, 179–195 (2008).

63. Hoover, D. R., Rice, J. A., Wu, C. O. & Yang, L.-P. Nonparametric smoothing estimates of time-varying coefficient models with longitudinal data. *Biometrika* **85**, 809–822 (1998).

64. Madigan, D. & Raftery, A. E. Model Selection and Accounting for Model Uncertainty in Graphical Models Using Occam's Window. (1991).

65. Onorante, L. & Raftery, A. E. Dynamic model averaging in large model spaces using dynamic Occam's window. *Eur. Econ. Rev.* **81**, 2–14 (2016).

66. Ohata, T., Kaneko, M., Kuratani, T., Ueda, H. & Shimamura, K. Using the EuroSCORE to assess changes in the risk profiles of the patients undergoing coronary artery bypass grafting before and after the introduction of less invasive coronary surgery. *Ann. Thorac. Surg.* **80**, 131–135 (2005).

67. Ramsay, J. O., Wickham, H., Graves, S. & Hooker, G. fda: Functional Data Analysis. (2017).

68. Altman, D. G. & Royston, P. What do we mean by validating a prognistic model? *Stat. Med.* **19**, 453–473 (2000).

69. Justice, A. C., Covinsky, K. E. & Berlin, J. A. Assessing the generalizability of prognostic information. *Ann. Intern. Med.* **130**, 515–524 (1999).

70. Zimmerman, J. E., Kramer, A. A., McNair, D. S. & Malila, F. M. Acute Physiology and Chronic Health Evaluation (APACHE) IV: Hospital mortality assessment for today's critically ill patients. *Crit. Care Med.* **34**, 1297–1310 (2006).

71. Vergouwe, Y. *et al.* A closed testing procedure to select an appropriate method for updating prediction models. *Stat. Med.* **36**, 4529–4539 (2017).

72. Hafkamp-De Groen, E. *et al.* Predicting asthma in preschool children with asthma-like symptoms: Validating and updating the PIAMA risk score. *J. Allergy Clin. Immunol.* **132**, (2013).

73. Genders, T. S. S. *et al.* A clinical prediction rule for the diagnosis of coronary artery disease: Validation, updating, and extension. *Eur. Heart J.* **32**, 1316–1330 (2011).

74. Moons, K. G. M. *et al.* Risk prediction models: II. External validation, model updating, and impact assessment. *Heart* **98**, 691–698 (2012).

75. Hippisley-Cox, J., Coupland, C. & Brindle, P. Development and validation of QRISK3 risk prediction algorithms to estimate future risk of cardiovascular disease: Prospective cohort study. *BMJ* **357**, 1–21 (2017).

76. Martin, G. P., Sperrin, M. & Sotgiu, G. Performance of Prediction Models for Covid-19: The Caudine Forks of the External Validation. *Eur. Respir. J.* 2003728 (2020). doi:10.1183/13993003.03728-2020

77. Wynants, L. *et al.* Prediction models for diagnosis and prognosis of covid-19: Systematic review and critical appraisal. *BMJ* **369**, (2020).

78. Vergouwe, Y., Moons, K. G. M. & Steyerberg, E. W. External validity of risk models: Use of benchmark values to disentangle a case-mix effect from incorrect coefficients. *Am. J. Epidemiol.* **172**, 971–980 (2010).

79. Debray, T. P. A. *et al.* A guide to systematic review and meta-analysis of prediction model performance. *BMJ* **356**, (2017).

80. Luijken, K. *et al.* Changing predictor measurement procedures affected the performance of prediction models in clinical examples. *J. Clin. Epidemiol.* **119**, 7–18 (2020).

81. Riley, R. D. *et al.* External validation of clinical prediction models using big datasets from e-health records or IPD meta-analysis: Opportunities and challenges. *BMJ* **353**, 27–30 (2016).

82. Debray, T. P. A., Riley, R. D., Rovers, M. M., Reitsma, J. B. & Moons, K. G. M. Individual Participant Data (IPD) Meta-analyses of Diagnostic and Prognostic Modeling Studies: Guidance on Their Use. *PLoS Med.* **12**, 1–12 (2015).

83. Hippisley-Cox, J. *et al.* Predicting cardiovascular risk in England and Wales: Prospective derivation and validation of QRISK2. *Bmj* **336**, 1475–1482 (2008).

84. Jenkins, D. A., Sperrin, M., Martin, G. P. & Peek, N. Dynamic models to predict health outcomes: current status and methodological challenges. *Diagnostic Progn. Res.* **2**, 23 (2018).

85. Halabi, S. *et al.* Updated prognostic model for predicting overall survival in first-line chemotherapy for patients with metastatic castration-resistant prostate cancer. *J. Clin. Oncol.* **32**, 671–677 (2014).

86. Dawid, A. Present Position and Potential Developments : Some Personal Views : Statistical Theory : The Prequential Approach Author (s): A . P . Dawid Source : Journal of the Royal Statistical Society . Series A (General), Vol . 147 , No . 2 , The 150th Annivers. *J. R. Stat. Soc. Ser. A* **147**, 278–292 (1984).

87. Lenert, M. C., Matheny, M. E. & Walsh, C. G. Prognostic models will be victims of their own success, unless…. *J. Am. Med. Inform. Assoc.* **26**, 1645–1650 (2019).

88. Adibi, A., Sadatsafavi, M. & Ioannidis, J. P. A. Validation and Utility Testing of Clinical Prediction

Models. *JAMA* **2004**, (2020).

89. Booth, S., Riley, R. D., Ensor, J., Lambert, P. C. & Rutherford, M. J. Temporal recalibration for improving prognostic model development and risk predictions in settings where survival is improving over time. *Int. J. Epidemiol.* 1–10 (2020). doi:10.1093/ije/dyaa030

90. Jenkins, D. A. *et al.* Continual updating and monitoring of clinical prediction models: time for dynamic prediction systems? *Diagnostic Progn. Res.* **5**, 1–7 (2021).

91. Mccormick, T. H. *et al.* Dynamic Logistic Regression and Dynamic Model Averaging for Binary Classification. 23–30 (2012). doi:10.1111/j.1541-0420.2011.01645.x

92. R Core Team. R: A language and environment for statistical computing. R Foundation for Statistical. (2014).

93. Martin, G. P., Riley, R. D., Collins, G. S. & Sperrin, M. Developing clinical prediction models when adhering to minimum sample size recommendations: The importance of quantifying bootstrap variability in tuning parameters and predictive performance. *Stat. Methods Med. Res.* **30**, 2545–2561 (2021).

94. Sperrin, M., Jenkins, D., Martin, G. P. & Peek, N. Explicit causal reasoning is needed to prevent prognostic models being victims of their own success. *J. Am. Med. Informatics Assoc.* **26**, 1675–1676 (2019).

95. Hickey, G. L. *et al.* Clinical registries: Governance, management, analysis and applications. *Eur. J. Cardio-thoracic Surg.* **44**, 605–614 (2013).

96. Moons, K. G. M. *et al.* Risk prediction models: II. External validation, model updating, and impact assessment. *Heart* **98**, 691–698 (2012).

97. Davis, S. E. *et al.* A nonparametric updating method to correct clinical prediction model drift. *J. Am. Med. Informatics Assoc.* **26**, 1448–1457 (2019).

98. Minne, L. *et al.* Statistical process control for monitoring standardized mortality ratios of a classification tree model. *Methods Inf. Med.* **51**, 353–358 (2012).

99. Statistics, M. SUMS OF NONINDEPENDENT BERNOULLI RANDOM VARIABLES Author (s): Jose M . Gonzalez-Barrios Source : Brazilian Journal of Probability and Statistics , JUNE 1998 , Vol . 12 , No . 1 (JUNE Published by : Institute of Mathematical Statistics Stable URL : http. **12**, 55–64 (1998).

100. Koetsier, A., De Keizer, N. F., De Jonge, E., Cook, D. A. & Peek, N. Performance of risk-adjusted control charts to monitor in-hospital mortality of intensive care unit patients: A simulation study. *Crit. Care Med.* **40**, 1799–1807 (2012).

101. Team, R. core. R: A language and environment for statistical computing. *R Foundation for Statistical Computing, Vienna, Austria.* (2021). Available at: https://www.r-project.org/.

102. Albert, A. A. *et al.* On-line variable live-adjusted displays with internal and external risk-adjusted mortalities. A valuable method for benchmarking and early detection of unfavourable trends in cardiac surgery. *Eur. J. Cardio-thoracic Surg.* **25**, 312–319 (2004).

103. Lovegrove, J., Valencia, O., Treasure, T., Sherlaw-Johnson, C. & Gallivan, S. Monitoring the results of cardiac surgery by variable life-adjusted display. *Lancet* **350**, 1128–1130 (1997).

104. Poloniecki, J., Valencia, O. & Littlejohns, P. Correction: Cumulative risk adjusted mortality chart

for detecting changes in death rate: Observational study of heart surgery (British Medical Journal (1998) (1697-700)). *Br. Med. J.* **316**, 1947 (1998).

105. Minne, L. *et al.* Effect of changes over time in the performance of a customized SAPS-II model on the quality of care assessment. *Intensive Care Med.* **38**, 40–46 (2012).

106. Shi, X., Gallagher, C., Lund, R. & Killick, R. A Comparison of Single and Multiple Changepoint Techniques for Time Series Data. 1–29 (2021). doi:10.1016/j.csda.2022.107433

107. Wittenberg, P., Gan, F. F. & Knoth, S. A simple signaling rule for variable life-adjusted display derived from an equivalent risk-adjusted CUSUM chart. *Stat. Med.* **37**, 2455–2473 (2018).

108. Davis, S. E., Greevy, R. A., Lasko, T. A., Walsh, C. G. & Matheny, M. E. Detection of calibration drift in clinical prediction models to inform model updating. *J. Biomed. Inform.* **112**, 103611 (2020).

109. Riley, R. D. *et al.* Minimum sample size for external validation of a clinical prediction model with a binary outcome. *Stat. Med.* **40**, 4230–4251 (2021).

110. Steyerberg, E. W., Borsboom, G. J. J. M., van Houwelingen, H. C., Eijkemans, M. J. C. & Habbema, J. D. F. Validation and updating of predictive logistic regression models: A study on sample size and shrinkage. *Stat. Med.* **23**, 2567–2586 (2004).

111. Grigg, O. A., Farewell, V. T. & Spiegelhalter, D. J. Use of risk-adjusted CUSUM and RSPRT charts for monitoring in medical contexts. *Stat. Methods Med. Res.* **12**, 147–170 (2003).

112. Pagel, C. *et al.* Real time monitoring of risk-adjusted paediatric cardiac surgery outcomes using variable life-adjusted display: Implementation in three UK centres. *Heart* **99**, 1445–1450 (2013).

113. Barrett, J. & Su, L. Dynamic predictions using flexible joint models of longitudinal and time-to-event data. *Stat. Med.* **36**, 1447–1460 (2017).

114. van Houwelingen, H., & Putter, H. *Dynamic Prediction in Clinical Survival Analysis*. (2012). doi:https://doi.org/10.1201/b11311

6.8 Supplementary material

Supplementary Table S6.1 - Median and 95% quantile for the observation number an alert was triggered across the 1000 iterations for each value of miscalibration, split by the statistical process control limits of 3 and 4 standard deviations

	3SDs control limit	**4SDs control limit**
Miscalibration	**Median (95% quantile)**	**Median (95% quantile)**
0	NA* (1006 - NA)	NA (NA - NA)
0.1	31847 (494 - NA)	NA (9344 - NA)
0.2	7510 (294 - 27136)	16022 (1863 - 42160)
0.3	3216 (267 - 11953)	6789 (665 - 18565)
0.4	1624 (235 - 6613)	3559 (377 - 9103)
0.5	990 (229 - 3752)	1959 (279 - 5756)
0.6	662 (216 - 2445)	1273 (253 - 3538)
0.7	485 (209 - 1715)	878 (234 - 2500)
0.8	367 (198 - 1225)	629 (219 - 1887)
0.9	301 (189 - 969)	474 (215 - 1368)

1	275 (186 - 761)	360 (203 - 1041)
1.1	264 (184 - 597)	303 (195 - 842)
1.2	258 (182 - 523)	278 (187 - 695)
1.3	253 (180 - 403)	263 (185 - 571)
1.4	250 (177 - 358)	256 (180 - 458)
1.5	249 (177 - 330)	253 (178 - 392)
1.6	248 (176 - 323)	249 (178 - 340)
1.7	246 (175 - 321)	248 (176 - 325)
1.8	246 (175 - 321)	247 (175 - 321)
1.9	246 (175 - 320)	246 (175 - 321)
2	246 (175 - 320)	246 (175 - 321)

*NA in the table represents no alert.

Chapter 7 General Discussion

At the end of each chapter there has been a chapter specific discussion and conclusion and not everything that has already been said in those sections will be repeated here. Rather, this chapter aims to briefly summarize each chapter and relate the findings of the thesis back to the initial research objectives (Section 7.1), and discuss unanswered questions and provide ideas of further work (Section 7.2).

7.1 Summary of findings

A representation of the relationship between the chapters and objectives is given in Figure 7.1 and discussed in this section.

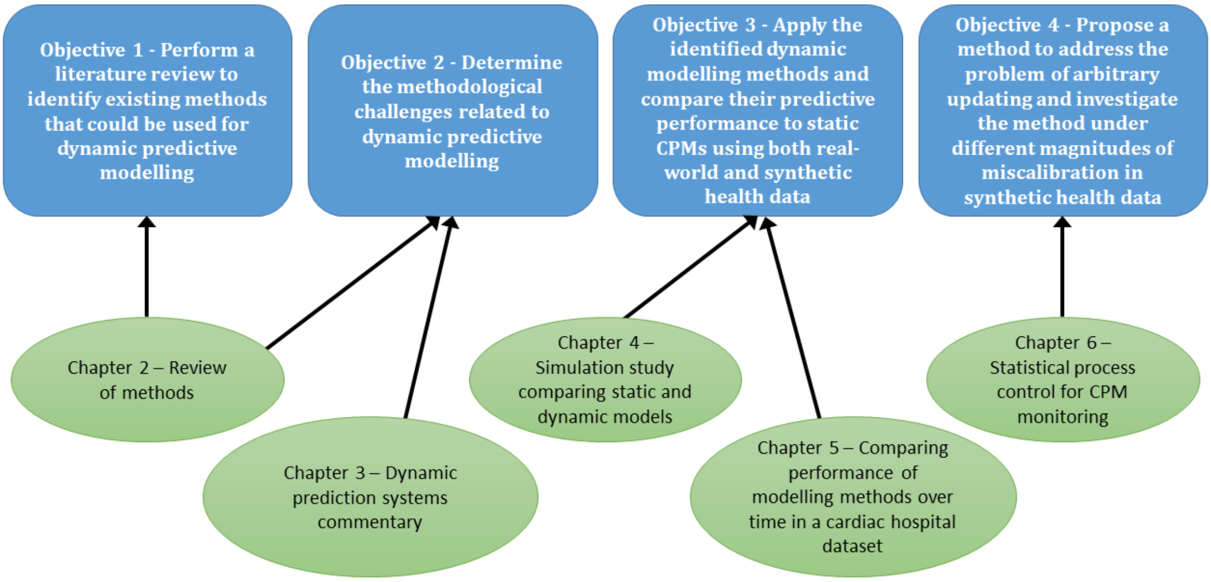

Objective 1 - Perform a literature review to identify existing methods that could be used for dynamic predictive modelling
Objective 2 - Determine the methodological challenges related to dynamic predictive modelling
Objective 3 - Apply the identified dynamic modelling methods and compare their predictive performance to static CPMs using both real-world and synthetic health data
Objective 4 - Propose a method to address the problem of arbitrary updating and investigate the method under different magnitudes of miscalibration in synthetic health data
Chapter 2 – Review of methods
Chapter 3 – Dynamic prediction systems commentary
Chapter 4 – Simulation study comparing static and dynamic models
Chapter 5 – Comparing performance of modelling methods over time in a cardiac hospital dataset
Chapter 6 – Statistical process control for CPM monitoring

Figure 7.1 - Pictorial representation of how the chapters link with the research objectives

7.1.1 Objective 1 - Perform a literature review to identify existing methods that could be used for dynamic predictive modelling

Chapter 2 of this thesis addresses Objective 1 by identifying methods for dynamic prediction modelling through a review of the literature. Eleven papers were included after screening and seven modelling methods to address calibration drift were identified. These were split into three categories: discrete model updating, Bayesian model updating and varying coefficient modelling. Discrete model updating[35,36,41,110] uses batches of new data to recalibrate or revise the model and four discrete modelling methods were identified. Two Bayesian model updating methods[45,48] were also found where the information obtained from past data is used as prior information and combined with the new data to obtain updated estimates. These methods can down weight historical observations and either update with every new data point, in an online learning manner, or with batches of data. Finally, the varying coefficient modelling[63] approach uses the data up to a given time point to estimate the relationship between the predictor and outcome as a function of time. This can be simply including time into the model so the intercept is time varying or including interactions and non-linear function of time to model more complex data structures.

Bayesian model updating has been used for dynamic predictive modelling for healthcare in the past but varying coefficient models have yet to be explored for this purpose. The review goes beyond solely identifying the methods by describing the methods in detail and the current use of the methods in healthcare literature.

7.1.2 Objective 2 - Determine the methodological challenges related to dynamic predictive modelling

As well as identifying methods for dynamic prediction modelling, Chapter 2 also discusses the methodological challenges related to dynamic predictive modelling described in the literature[84]. Methodology for validation has not been established for dynamic clinical prediction models (CPMs) and it is unclear how past observations should be dealt with over time when developing CPMs. Approaches such as windowing or forgetting have been used but the methodological challenge on how to handle past observations most appropriately to improve prediction remains. Chapter 3 further highlights both of these challenges and suggests that the current approach to static CPM validation is also flawed. Once a model has been validated it does not follow that the model will continue to perform well in the future and Chapter 3 highlights the need for a dynamic/continuous validation approach. In addition to this, uncertainty on when to update a model was another methodological challenge discussed in Chapter 3[90]. It may not be possible, or necessary, to continuously update prediction models and instead models should be updated via a data driven process and when needed, not at arbitrary time points. Model surveillance and the use of prequential testing was proposed as a solution to explore this but the methodological advances have yet to be developed and subsequently tested. Furthermore, results from the surveillance will need to be transported back to the model to enable the model to learn and ensure the model continually provides accurate predictions. This resulted in the suggestion of a system based approach that combines the current CPM methods and pipeline with the learning health system to generate a learning prediction system[90].

Together, Chapters 2 and 3 determine the methodological challenges related to dynamic prediction modelling and propose a solution, dynamic prediction systems, to overcome many of the challenges. However, methodological development is still needed to implement this system-based approach.

7.1.3 Objective 3 - Apply the identified dynamic modelling methods and compare their predictive performance to static CPMs using both real-world and synthetic health data.

Chapter 4 was a simulation study comparing the methods identified in Chapter 2 under different scenarios for a binary and continuous response variable. The models fitted were: one frequentist time-invariant model, four frequentist time-variant models with varying coefficients, and two Bayesian time-variant models with continual updating. From the varying coefficient models, two only allowed the intercept to depend on time and the other two considered all coefficients as linear functions of time. For the Bayesian and both types of varying coefficient models we predicted the outcomes using the model at the end of the development data and at the time of each prediction. Calibration and discrimination (for the logistic outcome) were calculated for each simulation scenario. Following this, Chapter 4 also compared the performance of the modelling approaches in a cardiac data set.

Chapter 5 was a retrospective cohort study that developed dynamic and non-dynamic CPMs in a single-centre cardiac surgery data set and compared the performance of the models from 2012 to 2019. The models developed were a time-invariant logistic regression model, a yearly updated logistic regression model, a time-variant logistic model with varying coefficients and Bayesian time-variant model with continual updating.

The results discussed in Chapters 4 and 5 have shown the benefit of dynamic, compared to static, CPMs in terms of predictive performance. In the simulated data the varying coefficient model had the best predictive performance, compared to Bayesian updating and static (time-invariant) regression, but Bayesian updating had less variability in predictive performance measures for the logistic scenarios. In the real-world data, from both chapters, the Bayesian updating models outperformed the varying coefficient models and less variability was observed in predictive performance over time. Also, the varying coefficient model was the worst performing model, dynamic or otherwise, in the National Adult Cardiac Surgery Audit data. Periodic (e.g. yearly) updating was also considered in the real-world data sets and outperformed the varying coefficient model and the static (time-invariant) regression models. Periodic updating performed as well as Bayesian modelling over the complete validation data but was more variable over time. Overall, accounting for temporal changes in data are important and how we do this can impact the performance of our models. Hence, Chapters 4 and 5 have addressed Objective 3 and compared dynamic modelling methods to static CPMs in synthetic (simulated) data and two real-world data sets.

7.1.4 Objective 4 - Propose a method to address the problem of arbitrary updating and investigate the method under different magnitudes of miscalibration in synthetic health data.

The uncertainty of when to update a model was discussed as a methodological challenge in Chapter 3. The use of statistical process control (SPC) to detect calibration drift and trigger an alert is a possible solution to address this challenge. Grigg et al.[111] previously discussed the potential of SPC using control charts for monitoring in medical contexts following the initial proposal by Pagel[112]. Chapter 6 outlines the statistical process control equations for monitoring the cumulative sum of the expected minus observed performance. The chapter builds upon the existing literature by proposing a resetting process for the control limits that ensure the control limits do not continue to increase while the process being monitored remains stable. Thus, ensuring miscalibration is detected in a timely manner following a stable process. Alternatively, to decrease type I error and ensure alerts are not falsely occurring at the start of monitoring, a threshold was proposed such that alerts could not occur until a given condition was met, the expectation of the process mean for the cumulative sum of the observed minus expected must be above 5. Following the novel description of this method a simulation study was undertaken to demonstrate the feasibility and accuracy of the proposed approach and investigate the time to alert under different magnitudes of miscalibration in synthetic health data.

Beyond derivation and illustration of the method, the results showed the approach was able to detect miscalibration in a timely manner but when the magnitude of the miscalibration was small (or zero) there was large variation in drift detection which could result in unnecessary updates. As miscalibration increased the time to detection plateaued and similar time-to-alert was observed for all values when the model was miscalibrated by a shift in the model intercept by 1 or more. Hence, Chapter 6 has achieved Objective 4 and SPC is a promising approach to continuously monitor prediction models and provide feedback on when a model may need to be updated.

7.2 Unanswered questions and future work

Although all of the objectives have been achieved, there are elements that have not fully been achieved and there remain some unanswered questions that still require further research. Also, while this thesis presents important advances in dynamic prediction modelling for healthcare use, by addressing the outlined aims and objectives, several areas warrant further investigation.

Firstly, the thesis focuses on short term outcomes and as a result we consider logistic and linear regression throughout the thesis. However, many prediction models consider long term outcomes, for example, QRISK[20] predicts 10 year risk of cardiovascular disease. Often these models consider the outcome as a time-to-event outcome and are developed using survival analysis. Research on dynamic survival analysis is sparse. Where "dynamic survival analysis" is mentioned in the literature it typically refers to the problem of handling multiple observations per individual rather than updating over time as new individuals are observed[113,114]. However, Booth et al[89] recently proposed an approach to update survival prediction models over time but this only updated the baseline hazard and updating occurs at arbitrary time points. The method proposed in Chapter 6 for determining when to update a model could be used for survival models but further work is needed to consider updating of the linear predictor (not just the baseline hazard), and how to deal with delay in outcome measures. Temporal changes in the data are not observed until the outcome measure has been recorded and so survival models are subject to increased latency. Model updating can only go so far to address this and extension of methods, such as extrapolation, are needed.

Second, while this thesis discussed an approach to choose the forgetting factor for dynamic Bayesian CPMs based upon sample size criteria, refinement of this approach is needed to optimise the performance of dynamic CPMs. The method discussed simply ensures an adequate sample size and does not consider optimisation. Further methodological work is required to optimise this 'forgetting' element of dynamic prediction. Some key areas of development should include: 1) some parts of the model adapting to recent data more quickly: e.g. the intercept may need to adapt quicker than predictor-outcome relationships (e.g., to ensure that the overall event rate is estimated correctly); and 2) the 'forgetting' may need to be itself dynamic – e.g. forgetting historical data more quickly when the healthcare system is changing rapidly, such as in a pandemic. In addition to Bayesian models, 'forgetting' needs further consideration in all model updating methods. Different methods exist to update a model and they all handle past data differently, for example, model recalibration uses only the new data available to update, while revision will use a window of data and this could be all of the data available, including the data originally used to develop the original model, or only the most recent data, say a month. Vergouwe et al[71] developed a closed test procedure to determine which discrete updating method to use and this is linked to dealing with past data, but there is no research on how to weight individuals over time to improve the models. One of the main reasons to update a model is calibration drift. So, if we know there are likely changes over time then why do we consider all data equal in model development, or at best, only consider a subset of data where we weight all data points equally? Further research should consider how to weight observations to optimise CPMs.

Finally, while Chapter 6 showed promise in a novel method to provide an alert that a CPM may need to be updated, this is the first study investigating the method. Rarely, if at all, are new methods accepted and adopted by stakeholders without multiple studies, considering different scenarios, to instil trust in the method. Investigation of the method in real-word data, and variety of clinical areas, is therefore needed. Recently, Davis et al.[108] proposed a method for the detection of calibration drift by deriving dynamic calibration curves using adaptive sliding windows. They implement their approach in a simulation study and apply the method to a real-world data set, testing for miscalibration by comparing a recent window with a previous window and do this continuously as new data arrive. It is therefore worth comparing the Davis approach with the proposed method in Chapter 6, for example, through a simulation study to compare the time both methods detect calibration drift. In addition to this, comparison of both methods impact on the long-term predictive performance of models would be worthwhile to investigate. However, before this is to be achieved, further consideration of the method proposed in Chapter 6 is needed. The proposed method currently only monitors the observed-expected value but CPM performance is typically evaluated using discrimination and calibration. Although this approach is to monitor a model and not to directly validate the model's performance, monitoring model validation measure could be a more robust approach to monitoring. Expanding the method to monitor other performance measures, for example, calibration in the large and discrimination would be useful. The point at which the method provides an alert could differ depending on the performance measure it is monitoring and different applications, or healthcare settings, may consider a different performance measure to be more relevant. Upon doing this, one could envisage two uses of the method: 1) the method monitors all performance measures and provides feedback to update at the first time point that any measure determines the model needs to be updated; or 2) a user could request to only update the models upon certain performance measures and thresholds. Regarding thresholds, the thesis considers control limit thresholds of 3 and 4 standard deviations as this is whats commonly used in the literature. However, in medical statistics we often set a threshold, a significance level, when performing a hypothesis test, known as type I error. When using statistical process control for monitoring a model, we are performing hypothesis tests at each time point and testing if the sum of the observed minus expected is different to zero. Hence, rather than setting a threshold of standard deviation, type I error could be used to ensure the same error rate is used over time. This is not elementary as tests are correlated over time and further methodological work is needed to advance the method. Therefore, future work should consider extension of the method to include other performance measures and to use type I error for determining the threshold. Following this, comparison of the method to the dynamic calibration curve method by Davis et al[108] will be needed. Subsequently, following an alert, the question of how to update and down weight past observations in a model update remain. Hence, the second discussion point for further work is also linked to this and to achieve the dynamic prediction systems, discussed in Chapter 3, more time, funding and research is required.

7.3 Conclusion

In conclusion, this thesis has: 1) provided an overview of dynamic modelling methods for developing clinical prediction models and compared the methods to non-dynamic (static) models; 2) identified challenges associated with dynamic model development and validation; and 3) proposed a novel approach to address the problem of arbitrary updating. Generally, the thesis has shown the value of dynamic modelling methods for clinical prediction model updating and monitoring for healthcare use and illustrated the, existing and newly proposed, methods in real-world and simulated health data.

7.4 References

1. Murdoch, T. B. & Detsky, A. S. The Inevitable Application of Big Data. *Jama* **309**, 1351–1352 (2014).

2. Wang, Y., Kung, L. A. & Byrd, T. A. Big data analytics: Understanding its capabilities and potential benefits for healthcare organizations. *Technol. Forecast. Soc. Change* **126**, 3–13 (2018).

3. Raghupathi, W. & Raghupathi, V. Big data analytics in healthcare: promise and potential. *Heal. Inf. Sci. Syst.* **2**, 3 (2014).

4. Friedman, C. P., Wong, A. K. & Blumenthal, D. Achieving a Nationwide Learning Health System. *Public Law* **2**, 8–10 (2010).

5. Friedman, C. P., Rubin, J. C. & Sullivan, K. J. Toward an Information Infrastructure for Global Health Improvement. *Yearb. Med. Inform.* **26**, 16–23 (2017).

6. J., A. & I., B. Combining health data uses to ignite health system learning. *Methods Inf. Med.* **54**, 479–487 (2015).

7. Kent, P., Cancelliere, C., Boyle, E., Cassidy, J. D. & Kongsted, A. A conceptual framework for prognostic research. *BMC Med. Res. Methodol.* **7**, 1–13 (2020).

8. Hemingway, H. *et al.* Prognosis research strategy (PROGRESS) 1: A framework for researching clinical outcomes. *BMJ* **346**, 1–11 (2013).

9. Riley, R. D. *et al.* Prognosis Research Strategy (PROGRESS) 2: Prognostic Factor Research. *PLoS Med.* **10**, e1001380 (2013).

10. Steyerberg, E. *et al.* Prognosis research strategy (PROGRESS) 3: prognostic model research. *PLoS Med* **10**, e1001381 (2013).

11. Hingorani, A. D. *et al.* Prognosis research strategy (PROGRESS) 4: Stratified medicine research. *BMJ* **346**, 1–9 (2013).

12. Riley, R. D., Windt, D. Van Der & Moons, K. G. M. Prognosis Research in Health Care. *Progn. Res. Heal. Care* 1–11 (2019). doi:10.1093/med/9780198796619.001.0001

13. Steyerberg, E. W. *Clinical Prediction Models. Statistics for Biology and Health. 2nd edition*. (2019).

14. Hemingway, H. *et al.* Prognosis research strategy (PROGRESS) 1: A framework for researching clinical outcomes. *BMJ* **346**, 1–11 (2013).

15. Gray, L. J. *et al.* The Leicester Risk Assessment score for detecting undiagnosed Type 2 diabetes and impaired glucose regulation for use in a multiethnic UK setting. *Diabet. Med.* **27**, 887–895 (2010).

16. (NICE), N. I. for H. and C. E. Type 2 diabetes: prevention in people at high risk. *Clinical guidelines [PH38]* (2012). Available at: https://www.nice.org.uk/guidance/ph38.

17. McAllister, K. S. L. *et al.* A contemporary risk model for predicting 30-day mortality following percutaneous coronary intervention in England and Wales. *Int. J. Cardiol.* **210**, 125–132 (2016).

18. Damen, J. A. A. G. *et al.* Prediction models for cardiovascular disease risk in the general population: Systematic review. *BMJ* **353**, (2016).

19. D'Agostino, R. B. *et al.* General Cardiovascular Risk Profile for Use in Primary Care: The Framingham Heart Study. *Circulation* **117**, 743–753 (2008).

20. Hippisley-Cox, J. *et al.* Derivation and validation of QRISK, a new cardiovascular disease risk score for the United Kingdom: Prospective open cohort study. *Br. Med. J.* **335**, 136–141 (2007).

21. (NICE), N. I. for H. and C. E. Cardiovascular disease: risk assessment and reduction, including lipid modification. *NICE Guidel. [CG181]* (2014).

22. Hippisley-Cox, J., Coupland, C., Robson, J. & Brindle, P. Derivation, validation, and evaluation of a new QRISK model to estimate lifetime risk of cardiovascular disease: Cohort study using QResearch database. *Bmj* **342**, 93 (2011).

23. Bonnett, L. J., Snell, K. I. E., Collins, G. S. & Riley, R. D. Guide to presenting clinical prediction models for use in clinical settings. *BMJ* **365**, 1–8 (2019).

24. Riley, R. D. *et al.* Minimum sample size for developing a multivariable prediction model: PART II - binary and time-to-event outcomes. *Stat. Med.* **38**, 1276–1296 (2019).

25. Steyerberg, E. W. *Clinical prediction models: a practical approach to development, validation, and updating*. (2008).

26. Harrell Jr, F. E. *Regression Modeling Strategies : With Applications to Linear Models, Logistic and Ordinal Regression, and Survival Analysis in Springer Series in Statistics*. *Springer* (Springer, 2015).

27. Steyerberg, E. W. *et al.* Internal validation of predictive models: Efficiency of some procedures for logistic regression analysis. *J. Clin. Epidemiol.* **54**, 774–781 (2001).

28. Steyerberg, E. W. *et al.* Assessing the performance of prediction models: a framework for some traditional and novel measures. *Epidemiology* **21**, 128–138 (2013).

29. Alba, A. C. *et al.* Discrimination and calibration of clinical prediction models: Users' guides to the medical literature. *JAMA - J. Am. Med. Assoc.* **318**, 1377–1384 (2017).

30. Van Calster, B. *et al.* Calibration: The Achilles heel of predictive analytics. *BMC Med.* **17**, 1–7 (2019).

31. Steyerberg, E. W. & Vergouwe, Y. Towards better clinical prediction models: Seven steps for development and an ABCD for validation. *Eur. Heart J.* **35**, 1925–1931 (2014).

32. Kramer, A. A. & Zimmerman, J. E. Assessing the calibration of mortality benchmarks in critical care: The Hosmer-Lemeshow test revisited. *Crit. Care Med.* **35**, 2052–2056 (2007).

33. Hickey, G. L. *et al.* Dynamic trends in cardiac surgery: Why the logistic euroscore is no longer suitable for contemporary cardiac surgery and implications for future risk models. *Eur. J. Cardio-thoracic Surg.* **43**, 1146–1152 (2013).

34. Davis, S. E., Lasko, T. A., Chen, G., Siew, E. D. & Matheny, M. E. Calibration drift in regression and machine learning models for acute kidney injury. *J. Am. Med. Informatics Assoc.* **24**, 1052–1061 (2017).

35. van Houwelingen, H. C. & Thorogood, J. Construction, validation and updating of a prognostic

model for kidney graft survival. *Stat. Med.* **14**, 1999–2008 (1995).

36. Janssen, K. J. M., Moons, K. G. M., Kalkman, C. J., Grobbee, D. E. & Vergouwe, Y. Updating methods improved the performance of a clinical prediction model in new patients. *J. Clin. Epidemiol.* **61**, 76–86 (2008).

37. Debray, T. P. A., Koffijberg, H., Vergouwe, Y., Moons, K. G. M. & Steyerberg, E. W. Aggregating published prediction models with individual participant data: A comparison of different approaches. *Stat. Med.* **31**, 2697–2712 (2012).

38. Martin, G. P., Mamas, M. A., Peek, N., Buchan, I. & Sperrin, M. Clinical prediction in defined populations: a simulation study investigating when and how to aggregate existing models. *BMC Med. Res. Methodol.* **17**, 1 (2017).

39. Martin, G. P., Mamas, M. A., Peek, N., Buchan, I. & Sperrin, M. A multiple-model generalisation of updating clinical prediction models. *Stat. Med.* **37**, 1343–1358 (2018).

40. Steyerberg, E. W., Borsboom, G. J. J. M., van Houwelingen, H. C., Eijkemans, M. J. C. & Habbema, J. D. F. Validation and updating of predictive logistic regression models: A study on sample size and shrinkage. *Stat. Med.* **23**, 2567–2586 (2004).

41. Toll, D. B., Janssen, K. J. M., Vergouwe, Y. & Moons, K. G. M. Validation, updating and impact of clinical prediction rules: A review. *J. Clin. Epidemiol.* **61**, 1085–1094 (2008).

42. Hippisley-Cox, J., Coupland, C. & Brindle, P. Development and validation of QRISK3 risk prediction algorithms to estimate future risk of cardiovascular disease : prospective cohort study. **2099**, 1–21 (2017).

43. Nashef, S. A. M. *et al.* European system for cardiac operative risk evaluation (EuroSCORE). *Eur. J. Cardio-thoracic Surg.* **16**, 9–13 (1999).

44. Nashef, S. A. M. *et al.* Euroscore II. *Eur. J. Cardio-thoracic Surg.* **41**, 734–745 (2012).

45. Raftery, A. E. & Ettler, P. Online Prediction Under Model Uncertainty via Dynamic Model Averaging : Application to a Cold Rolling Mill. **52**, 52–66 (2010).

46. Bull, L. M., Lunt, M., Martin, G. P., Hyrich, K. & Sergeant, J. C. Harnessing repeated measurements of predictor variables for clinical risk prediction: a review of existing methods. *Diagnostic Progn. Res.* **4**, (2020).

47. Hickey, G. L. *et al.* Dynamic prediction modeling approaches for cardiac surgery. *Circ. Cardiovasc. Qual. Outcomes* **6**, 649–658 (2013).

48. Mccormick, T. H., Raftery, A. E., Madigan, D. & Burd, R. S. Dynamic Logistic Regression and Dynamic Model Averaging for Binary Classification. *Biometrics* **68**, 23–30 (2012).

49. McCormick, T. H., Raftery, A. & Madigan, D. dma: Dynamic Model Averaging. (2018).

50. Five Year Forward View. (2014).

51. Salive, M. E. Multimorbidity in older adults. *Epidemiol. Rev.* **35**, 75–83 (2013).

52. Divo, M. J., Martinez, C. H. & Mannino, D. M. Ageing and the epidemiology of multimorbidity. *Eur. Respir. J.* **44**, 1055–1068 (2014).

53. Watkins, J. *et al.* Effects of health and social care spending constraints on mortality in England:

a time trend analysis. *BMJ Open* **7**, e017722 (2017).

54. Abu-Hanna, A. & Lucas, P. J. F. Prognostic Models in Medicine. AI and Statistical Approaches. *Method Inf. Med* **40**, 1–5 (2001).

55. Damen, J. A. A. G. *et al.* Prediction models for cardiovascular disease risk in the general population: systematic review. *Bmj* i2416 (2016). doi:10.1136/bmj.i2416

56. Siregar, S. *et al.* Improved Prediction by Dynamic Modeling. *Circ. Cardiovasc. Qual. Outcomes* **9**, 171–181 (2016).

57. Su, T.-L., Jaki, T., Hickey, G. L., Buchan, I. & Sperrin, M. A review of statistical updating methods for clinical prediction models. *Stat. Methods Med. Res.* 1–16 (2016). doi:10.1177/0962280215626466

58. van Houwelingen, H. C. & Thorogood, J. Construction, validation and updating of a prognostic model for kidney graft survival. *Stat. Med.* **14**, 1999–2008 (1995).

59. Draper, N. R., Nostrand, R. C. Van & Draper, N. R. Ridge Regression and James-Stein Estimation : Review and Comments Linked references are available on JSTOR for this article : Ridge Regression and James-Stein Estimation : Review and Comments. **21**, 451–466 (2016).

60. Copas, J. . Regression, Prediction and Shrinkage. *R. Stat. Soc.* **45**, 311–354 (1983).

61. Finkelman, B. S., French, B. & Kimmel, S. E. The prediction accuracy of dynamic mixed-effects models in clustered data. *BioData Min.* **9**, 5 (2016).

62. Fan, J. & Zhang, W. Statistical Methods with Varying Coefficient Models. *Stat Interface* **1**, 179–195 (2008).

63. Hoover, D. R., Rice, J. A., Wu, C. O. & Yang, L.-P. Nonparametric smoothing estimates of time-varying coefficient models with longitudinal data. *Biometrika* **85**, 809–822 (1998).

64. Madigan, D. & Raftery, A. E. Model Selection and Accounting for Model Uncertainty in Graphical Models Using Occam's Window. (1991).

65. Onorante, L. & Raftery, A. E. Dynamic model averaging in large model spaces using dynamic Occam's window. *Eur. Econ. Rev.* **81**, 2–14 (2016).

66. Ohata, T., Kaneko, M., Kuratani, T., Ueda, H. & Shimamura, K. Using the EuroSCORE to assess changes in the risk profiles of the patients undergoing coronary artery bypass grafting before and after the introduction of less invasive coronary surgery. *Ann. Thorac. Surg.* **80**, 131–135 (2005).

67. Ramsay, J. O., Wickham, H., Graves, S. & Hooker, G. fda: Functional Data Analysis. (2017).

68. Altman, D. G. & Royston, P. What do we mean by validating a prognistic model? *Stat. Med.* **19**, 453–473 (2000).

69. Justice, A. C., Covinsky, K. E. & Berlin, J. A. Assessing the generalizability of prognostic information. *Ann. Intern. Med.* **130**, 515–524 (1999).

70. Zimmerman, J. E., Kramer, A. A., McNair, D. S. & Malila, F. M. Acute Physiology and Chronic Health Evaluation (APACHE) IV: Hospital mortality assessment for today's critically ill patients. *Crit. Care Med.* **34**, 1297–1310 (2006).

71. Vergouwe, Y. *et al.* A closed testing procedure to select an appropriate method for updating prediction models. *Stat. Med.* **36**, 4529–4539 (2017).

72. Hafkamp-De Groen, E. *et al.* Predicting asthma in preschool children with asthma-like symptoms: Validating and updating the PIAMA risk score. *J. Allergy Clin. Immunol.* **132**, (2013).

73. Genders, T. S. S. *et al.* A clinical prediction rule for the diagnosis of coronary artery disease: Validation, updating, and extension. *Eur. Heart J.* **32**, 1316–1330 (2011).

74. Moons, K. G. M. *et al.* Risk prediction models: II. External validation, model updating, and impact assessment. *Heart* **98**, 691–698 (2012).

75. Hippisley-Cox, J., Coupland, C. & Brindle, P. Development and validation of QRISK3 risk prediction algorithms to estimate future risk of cardiovascular disease: Prospective cohort study. *BMJ* **357**, 1–21 (2017).

76. Martin, G. P., Sperrin, M. & Sotgiu, G. Performance of Prediction Models for Covid-19: The Caudine Forks of the External Validation. *Eur. Respir. J.* 2003728 (2020). doi:10.1183/13993003.03728-2020

77. Wynants, L. *et al.* Prediction models for diagnosis and prognosis of covid-19: Systematic review and critical appraisal. *BMJ* **369**, (2020).

78. Vergouwe, Y., Moons, K. G. M. & Steyerberg, E. W. External validity of risk models: Use of benchmark values to disentangle a case-mix effect from incorrect coefficients. *Am. J. Epidemiol.* **172**, 971–980 (2010).

79. Debray, T. P. A. *et al.* A guide to systematic review and meta-analysis of prediction model performance. *BMJ* **356**, (2017).

80. Luijken, K. *et al.* Changing predictor measurement procedures affected the performance of prediction models in clinical examples. *J. Clin. Epidemiol.* **119**, 7–18 (2020).

81. Riley, R. D. *et al.* External validation of clinical prediction models using big datasets from e-health records or IPD meta-analysis: Opportunities and challenges. *BMJ* **353**, 27–30 (2016).

82. Debray, T. P. A., Riley, R. D., Rovers, M. M., Reitsma, J. B. & Moons, K. G. M. Individual Participant Data (IPD) Meta-analyses of Diagnostic and Prognostic Modeling Studies: Guidance on Their Use. *PLoS Med.* **12**, 1–12 (2015).

83. Hippisley-Cox, J. *et al.* Predicting cardiovascular risk in England and Wales: Prospective derivation and validation of QRISK2. *Bmj* **336**, 1475–1482 (2008).

84. Jenkins, D. A., Sperrin, M., Martin, G. P. & Peek, N. Dynamic models to predict health outcomes: current status and methodological challenges. *Diagnostic Progn. Res.* **2**, 23 (2018).

85. Halabi, S. *et al.* Updated prognostic model for predicting overall survival in first-line chemotherapy for patients with metastatic castration-resistant prostate cancer. *J. Clin. Oncol.* **32**, 671–677 (2014).

86. Dawid, A. Present Position and Potential Developments : Some Personal Views : Statistical Theory : The Prequential Approach Author (s): A . P . Dawid Source : Journal of the Royal Statistical Society . Series A (General), Vol . 147 , No . 2 , The 150th Annivers. *J. R. Stat. Soc. Ser. A* **147**, 278–292 (1984).

87. Lenert, M. C., Matheny, M. E. & Walsh, C. G. Prognostic models will be victims of their own

success, unless…. *J. Am. Med. Inform. Assoc.* **26**, 1645–1650 (2019).

88. Adibi, A., Sadatsafavi, M. & Ioannidis, J. P. A. Validation and Utility Testing of Clinical Prediction Models. *JAMA* **2004**, (2020).

89. Booth, S., Riley, R. D., Ensor, J., Lambert, P. C. & Rutherford, M. J. Temporal recalibration for improving prognostic model development and risk predictions in settings where survival is improving over time. *Int. J. Epidemiol.* 1–10 (2020). doi:10.1093/ije/dyaa030

90. Jenkins, D. A. *et al.* Continual updating and monitoring of clinical prediction models: time for dynamic prediction systems? *Diagnostic Progn. Res.* **5**, 1–7 (2021).

91. Mccormick, T. H. *et al.* Dynamic Logistic Regression and Dynamic Model Averaging for Binary Classification. 23–30 (2012). doi:10.1111/j.1541-0420.2011.01645.x

92. R Core Team. R: A language and environment for statistical computing. R Foundation for Statistical. (2014).

93. Martin, G. P., Riley, R. D., Collins, G. S. & Sperrin, M. Developing clinical prediction models when adhering to minimum sample size recommendations: The importance of quantifying bootstrap variability in tuning parameters and predictive performance. *Stat. Methods Med. Res.* **30**, 2545–2561 (2021).

94. Sperrin, M., Jenkins, D., Martin, G. P. & Peek, N. Explicit causal reasoning is needed to prevent prognostic models being victims of their own success. *J. Am. Med. Informatics Assoc.* **26**, 1675–1676 (2019).

95. Hickey, G. L. *et al.* Clinical registries: Governance, management, analysis and applications. *Eur. J. Cardio-thoracic Surg.* **44**, 605–614 (2013).

96. Moons, K. G. M. *et al.* Risk prediction models: II. External validation, model updating, and impact assessment. *Heart* **98**, 691–698 (2012).

97. Davis, S. E. *et al.* A nonparametric updating method to correct clinical prediction model drift. *J. Am. Med. Informatics Assoc.* **26**, 1448–1457 (2019).

98. Minne, L. *et al.* Statistical process control for monitoring standardized mortality ratios of a classification tree model. *Methods Inf. Med.* **51**, 353–358 (2012).

99. Statistics, M. SUMS OF NONINDEPENDENT BERNOULLI RANDOM VARIABLES Author (s): Jose M . Gonzalez-Barrios Source : Brazilian Journal of Probability and Statistics , JUNE 1998 , Vol . 12 , No . 1 (JUNE Published by : Institute of Mathematical Statistics Stable URL : http. **12**, 55–64 (1998).

100. Koetsier, A., De Keizer, N. F., De Jonge, E., Cook, D. A. & Peek, N. Performance of risk-adjusted control charts to monitor in-hospital mortality of intensive care unit patients: A simulation study. *Crit. Care Med.* **40**, 1799–1807 (2012).

101. Team, R. core. R: A language and environment for statistical computing. *R Foundation for Statistical Computing, Vienna, Austria.* (2021). Available at: https://www.r-project.org/.

102. Albert, A. A. *et al.* On-line variable live-adjusted displays with internal and external risk-adjusted mortalities. A valuable method for benchmarking and early detection of unfavourable trends in cardiac surgery. *Eur. J. Cardio-thoracic Surg.* **25**, 312–319 (2004).

103. Lovegrove, J., Valencia, O., Treasure, T., Sherlaw-Johnson, C. & Gallivan, S. Monitoring the

results of cardiac surgery by variable life-adjusted display. *Lancet* **350**, 1128–1130 (1997).

104. Poloniecki, J., Valencia, O. & Littlejohns, P. Correction: Cumulative risk adjusted mortality chart for detecting changes in death rate: Observational study of heart surgery (British Medical Journal (1998) (1697-700)). *Br. Med. J.* **316**, 1947 (1998).

105. Minne, L. *et al.* Effect of changes over time in the performance of a customized SAPS-II model on the quality of care assessment. *Intensive Care Med.* **38**, 40–46 (2012).

106. Shi, X., Gallagher, C., Lund, R. & Killick, R. A Comparison of Single and Multiple Changepoint Techniques for Time Series Data. 1–29 (2021). doi:10.1016/j.csda.2022.107433

107. Wittenberg, P., Gan, F. F. & Knoth, S. A simple signaling rule for variable life-adjusted display derived from an equivalent risk-adjusted CUSUM chart. *Stat. Med.* **37**, 2455–2473 (2018).

108. Davis, S. E., Greevy, R. A., Lasko, T. A., Walsh, C. G. & Matheny, M. E. Detection of calibration drift in clinical prediction models to inform model updating. *J. Biomed. Inform.* **112**, 103611 (2020).

109. Riley, R. D. *et al.* Minimum sample size for external validation of a clinical prediction model with a binary outcome. *Stat. Med.* **40**, 4230–4251 (2021).

110. Steyerberg, E. W., Borsboom, G. J. J. M., van Houwelingen, H. C., Eijkemans, M. J. C. & Habbema, J. D. F. Validation and updating of predictive logistic regression models: A study on sample size and shrinkage. *Stat. Med.* **23**, 2567–2586 (2004).

111. Grigg, O. A., Farewell, V. T. & Spiegelhalter, D. J. Use of risk-adjusted CUSUM and RSPRT charts for monitoring in medical contexts. *Stat. Methods Med. Res.* **12**, 147–170 (2003).

112. Pagel, C. *et al.* Real time monitoring of risk-adjusted paediatric cardiac surgery outcomes using variable life-adjusted display: Implementation in three UK centres. *Heart* **99**, 1445–1450 (2013).

113. Barrett, J. & Su, L. Dynamic predictions using flexible joint models of longitudinal and time-to-event data. *Stat. Med.* **36**, 1447–1460 (2017).

114. van Houwelingen, H., & Putter, H. *Dynamic Prediction in Clinical Survival Analysis*. (2012). doi:https://doi.org/10.1201/b11311