MEDINFO 2023 — The Future Is Accessible
J. Bichel-Findlay et al. (Eds.)

doi:10.3233/SHTI231120

Comparing Predictive Performance of Time Invariant and Time Variant Clinical Prediction Models in Cardiac Surgery

David A JENKINS[a,b], Glen P MARTIN[a], Matthew SPERRIN[a], Benjamin BROWN[b], Linda KIMANI[c], Stuart GRANT[c] and Niels PEEK[a,b,1]
[a]*Division of Informatics, Imaging and Data Science, Faculty of Biology, Medicine and Health, University of Manchester, Manchester, UK*
[b]*NIHR Greater Manchester Patient Safety Translational Research Centre, University of Manchester, Manchester, UK*
[c]*Manchester University Hospital NHS Foundation Trust, Manchester, UK*

Abstract. Clinical prediction models are increasingly used across healthcare to support clinical decision making. Existing methods and models are time-invariant and thus ignore the changes in populations and healthcare practice that occur over time. We aimed to compare the performance of time-invariant with time-variant models in UK National Adult Cardiac Surgery Audit data from Manchester University NHS Foundation Trust between 2009 and 2019. Data from 2009-2011 were used for initial model fitting, and data from 2012-2019 for validation and updating. We fitted four models to the data: a time-invariant logistic regression model (not updated), a logistic model which was updated every year and validated it in each subsequent year, a logistic regression model where the intercept is a function of calendar time (not updated), and a continually updating Bayesian logistic model which was updated with each new observation and continuously validated. We report predictive performance over the complete validation cohort and for each year in the validation data. Over the complete validation data, the Bayesian model had the best predictive performance.

Keywords. Clinical prediction models, dynamic model, validation, model updating, model development

1. Introduction

Clinical prediction models (CPMs) are increasingly used to assist in preventative decision making in healthcare [1]. CPMs use information about a patient to provide risk estimates for a certain outcome for the patient. For example, the European System for Cardiac Operative Risk Evaluation [2] (EuroSCORE) is a cardiac risk model for predicting mortality after cardiac surgery. This model is used to aid the clinician's decision on whether 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 has diminished over time [3]. Degradation of CPMs used in clinical practice is

[1] Corresponding Author: Niels Peek, email: niels.peek@manchester.ac.uk.

often observed and many models experience diminishing predictive performance over time [3]. 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.

A common approach to overcome this is periodic updating, where a CPM is revised based on most recent data [4]. Recently, other methods, known as dynamic prediction models [5], have been discussed to overcome model degradation, such as continuously updating Bayesian models [6] and varying coefficient models [7]. In this study we aim to investigate the predictive performance of time-variant and time-invariant modelling methods using a real-world cardiac dataset.

2. Methods

The National Adult Cardiac Surgery Audit (NACSA) registry collects data on major heart operations in the UK. It includes information on patient baseline demographics, risk factors for intervention, procedural details and patient outcomes. This study included NACSA data on all major heart operations from 1st January 2009 to 30th June 2019 from Wythenshawe hospital (Manchester University NHS Foundation Trust). The outcome was hospital mortality 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 [8], 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 [9]. 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.

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. The second was a yearly updated logistic regression model. This was identical to begin with as the first model but was then subsequently recalibrated at the start of each year [10]. The third model was a Bayesian time-variant model with continual updating [6]. The model was updated at each new observation and data were down weighted over time, by a 'forgetting' factor. We chose the effective window size [11] to be the same as the development set which resulted in a forgetting factor of 0.9997. This was chosen to ensure that the Bayesian model weighted individuals over time such that the sample size for each iteration 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 if we modelled each of the coefficients as functions of time this would require a much larger sample size.

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 separately. 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] Prequential testing [12] was used to validate the continuously

updated Bayesian dynamic model. Each validation measure was calculated for each year of data in the validation data and over the complete validation data. All analyses were performed using R (version 3.6.2) and the dynamic models were fitted using functions adapted from the dma package [13]. All code and supplementary material can be found on github (https://github.com/David-A-Jenkins/MedInfo-2023).

3. Results

The final data comprised 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, 92 (3.0%) of those were in the development data. Online supplementary table 1 displays baseline characteristics of the development and validation cohort.

The coefficients for the logistic regression and varying coefficient model can be found in supplementary tables 2,3 and 4. Figure 1 displays each of the model's performances separately for each year of data from 2012 to 2019. 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. A reduction in the observed-expected ratio was observed in 2014 for the logistic model and varying coefficient model. The Bayesian model discrimination increases from 0.72 in 2012 to 0.84 in 2018. In comparison, the other models' 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. Online Supplementary table 5 displays each model's performance values 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 and 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).

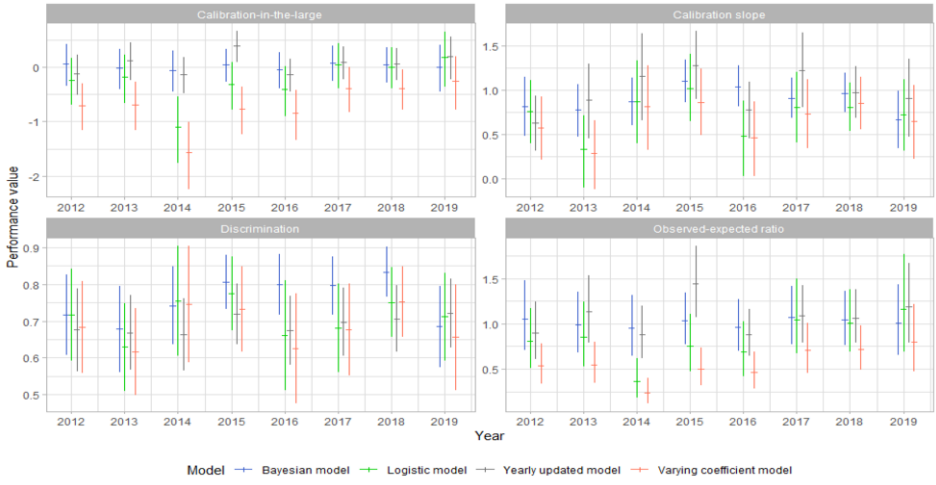

Figure 1. Yearly performance measure for each model between 2012 and 2019.

4. 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 their predictive performance. Over the complete validation data, the Bayesian model had the best predictive performance for calibration and discrimination. The Bayesian model provided the most stable yearly estimates of calibration-in-the-large and, on average, achieved the best discrimination. 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. 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 and both models consistently outperformed the time-invariant logistic model and varying coefficient model with respect to calibration but the Bayesian model outperformed all models in discrimination.

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 likely 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 [14, 15]. 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.

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, 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. Conclusions

CPMs are increasingly used to support clinical decision making, but typically ignore data shift over time. This can lead to suboptimal performance and thus impact the quality of clinical decisions. In a comparison of time-invariant logistic regression, periodic model recalibration, and continuous Bayesian updating, we found Bayesian updating models to

be the best performing model overall, but the less complex periodic model recalibration method also outperformed the time-invariant model. We recommend that time-variant models be considered when developing CPMs for assisting clinical decision making. The infrastructure needs to be available to implement these more complex methods.

Acknowledgments

DAJ's and NP's time are partly funded by the National Institute for Health Research Greater Manchester Patient Safety Translational Research Centre (NIHR Greater Manchester PSTRC). The views expressed are those of the author(s) and not necessarily those of the NHS, the NIHR or the Department of Health and Social Care.

References

[1] Riley RD, Moons KG, Hayden J, Sauerbrei W, Altman DG. Prognostic factor research. In: Prognosis research in health care: concepts, methods, and impact. 2019 Jan;24:107-38, doi: 10.1093/med/9780198796619.003.0007.

[2] Nashef SA, Roques F, Michel P, Gauducheau E, Lemeshow S, Salamon R, EuroSCORE Study Group. European system for cardiac operative risk evaluation (Euro SCORE). Eur J Cardio-Thoracic Surg. 1999 Jul;16(1):9-13, doi: 10.1016/S1010-7940(99)00134-7.

[3] Hickey GL, Grant SW, Murphy GJ, Bhabra M, Pagano D, McAllister K, Buchan I, Bridgewater B. 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. 2013 Jun;43(6):1146-52, doi: 10.1093/ejcts/ezs584.

[4] Vergouwe Y, Nieboer D, Oostenbrink R, Debray TP, Murray GD, Kattan MW, Koffijberg H, Moons KG, Steyerberg EW. A closed testing procedure to select an appropriate method for updating prediction models. Statistics Med. 2017 Dec;36(28):4529-39, doi: 10.1002/sim.7179.

[5] Jenkins DA, Sperrin M, Martin GP, Peek N. Dynamic models to predict health outcomes: current status and methodological challenges. Diagn Progn Res. 2018 Dec;2(1):23, doi: 10.1186/s41512-018-0045-2.

[6] McCormick TH, Raftery AE, Madigan D, Burd RS. Dynamic logistic regression and dynamic model averaging for binary classification. Biometrics. 2012 Mar;68(1):23-30, doi: 10.1111/j.1541-0420.2011.01645.x.

[7] Hoover DR, Rice JA, Wu CO, Yang LP. Nonparametric smoothing estimates of time-varying coefficient models with longitudinal data. Biometrika. 1998 Dec;85(4):809-22, doi: 10.1093/biomet/85.4.809.

[8] Nashef SA, Roques F, Sharples LD, Nilsson J, Smith C, Goldstone AR, Lockowandt U. Euroscore ii. Eur J Cardio-Thoracic Surg. 2012 Apr;41(4):734-45, doi: 10.1093/ejcts/ezs043.

[9] Hickey GL, Grant SW, Cosgriff R, Dimarakis I, Pagano D, Kappetein AP, Bridgewater B. Clinical registries: governance, management, analysis and applications. Eur J Cardio-Thoracic Surg. 2013 Oct;44(4):605-14, doi: 10.1093/ejcts/ezt018.

[10] Moons KG, Kengne AP, Grobbee DE, Royston P, Vergouwe Y, Altman DG, Woodward M. Risk prediction models: II. external validation, model updating, and impact assessment. Heart. 2012 May;98(9):691-8, doi: 10.1136/heartjnl-2011-301246.

[11] Raftery AE, Kárný M, Ettler P. Online prediction under model uncertainty via dynamic model averaging: application to a cold rolling mill. Technometrics. 2010 Feb;52(1):52-66, doi: 10.1198/TECH.2009.08104.

[12] Dawid AP. Present position and potential developments: some personal views statistical theory the prequential approach. J R Stat Soc Ser A Stat Soc: Ser A (General). 1984 Mar;147(2):278-90, doi: 10.2307/2981683.

[13] McCormick TH, Raftery A, Madigan D. dma: Dynamic Model Averaging. 2018.

[14] Jenkins DA, Martin GP, Sperrin M, Riley RD, Debray TP, Collins GS, Peek N. Continual updating and monitoring of clinical prediction models: time for dynamic prediction systems?. Diagnostic Progn Res. 2021 Dec;5:1-7, doi: 10.1186/s41512-020-00090-3.

[15] Lenert MC, Matheny ME, Walsh CG. Prognostic models will be victims of their own success, unles... J Am Med Inform Assoc. 2019 Dec;26(12):1645-50, doi: 10.1093/jamia/ocz145.