Tanner *et al.*
*Diagnostic and Prognostic Research (2023) 7:24*
https://doi.org/10.1186/s41512-023-00163-z

Diagnostic and
Prognostic Research

METHODOLOGY Open Access

# Dynamic updating of clinical survival prediction models in a changing environment

Kamaryn T. Tanner[1*], Ruth H. Keogh[1], Carol A. C. Coupland[2,3], Julia Hippisley-Cox[2] and Karla Diaz-Ordaz[4]

## Abstract

**Background** Over time, the performance of clinical prediction models may deteriorate due to changes in clinical management, data quality, disease risk and/or patient mix. Such prediction models must be updated in order to remain useful. In this study, we investigate dynamic model updating of clinical survival prediction models. In contrast to discrete or one-time updating, dynamic updating refers to a repeated process for updating a prediction model with new data. We aim to extend previous research which focused largely on binary outcome prediction models by concentrating on time-to-event outcomes. We were motivated by the rapidly changing environment seen during the COVID-19 pandemic where mortality rates changed over time and new treatments and vaccines were introduced.

**Methods** We illustrate three methods for dynamic model updating: Bayesian dynamic updating, recalibration, and full refitting. We use a simulation study to compare performance in a range of scenarios including changing mortality rates, predictors with low prevalence and the introduction of a new treatment. Next, the updating strategies were applied to a model for predicting 70-day COVID-19-related mortality using patient data from QResearch, an electronic health records database from general practices in the UK.

**Results** In simulated scenarios with mortality rates changing over time, all updating methods resulted in better calibration than not updating. Moreover, dynamic updating outperformed ad hoc updating. In the simulation scenario with a new predictor and a small updating dataset, Bayesian updating improved the C-index over not updating and refitting. In the motivating example with a rare outcome, no single updating method offered the best performance.

**Conclusions** We found that a dynamic updating process outperformed one-time discrete updating in the simulations. Bayesian updating offered good performance overall, even in scenarios with new predictors and few events. Intercept recalibration was effective in scenarios with smaller sample size and changing baseline hazard. Refitting performance depended on sample size and produced abrupt changes in hazard ratio estimates between periods.

**Keywords** Clinical prediction models, Dynamic model, Model updating, Survival analysis

*Correspondence:
Kamaryn T. Tanner
kamaryn.tanner1@lshtm.ac.uk
Full list of author information is available at the end of the article

## Background

Clinical prediction models are widely used in medicine to provide patients and clinicians with information about the predicted risk of an outcome, to guide treatment plans, and to identify high-risk groups. Although the best of these models go through a thorough development and validation process, performance may deteriorate over time as clinical practices change, mortality or disease risk in the population changes and/or the patient mix shifts [1–3]. After evidence of poor performance, a common response is to repeat the model development by fitting a new model to a new set of data [4]. However, the resulting 'new' model fails to incorporate knowledge learned in the initial development process, may not include the same predictors and could be confusing for end-users of the original model [2]. A technique for updating the existing prediction model rather than redeveloping a new one can alleviate these issues.

A variety of methods have been proposed and studied for updating clinical prediction models with the majority having been applied to binary outcome models based on logistic regression. Previous studies have compared recalibration, refitting (with or without shrinkage), Bayesian methods and testing procedures for selecting the 'best' updating method across a variety of scenarios [3–13]. No single updating method was best across these studies. Rather, the best updating method was found to depend on the sample size of the updating data, event rate, model complexity and whether associations of the predictors with the outcome had changed [4, 6, 7, 9, 12]. In particular, refitting was found to overfit the new data and yield unstable coefficient estimates in parametric models when the updating sample size was small or the number of events was low [3, 6, 7, 11]. Recalibration performed as well as or better than refitting, particularly when predictor relationships were not changing over time [4, 8, 11, 14]. Bayesian updating methods showed good predictive performance and may produce smoother updates to model coefficients than refitting [3, 6, 8].

Once a prediction model has been updated, it becomes susceptible to performance deterioration again due to the evolution of treatments or the disease itself, changes to the affected population, repeated exposures, or the quality of the data available for its implementation. Rather than performing updates as a one-time or discrete update, a strategy for continued updating can combat this. A dynamic updating strategy refers to an approach for updating a model at multiple times in the future when new data becomes available [11, 15]. Although dynamic updating strategies offer many benefits, their use is still limited because of difficulties in implementing and resourcing such a strategy, obtaining the necessary data and communicating the frequent changes to a prediction model [3, 15]. Furthermore, the study of updating strategies for time-to-event outcomes has been limited. Clinical examples of updating survival models include the recalibration of Cox proportional hazards models used to predict 30-day survival and 6-month independence after acute stroke by Sim et al. [5] and the recalibration of Cox models used for predicting survival after a particular treatment for hepatocellular carcinoma by Cucchetti et al. [16].

In this study, our primary aim is to assess methods for dynamic model updating of clinical survival prediction models. We were motivated by the COVID-19 pandemic where mortality rates changed over time and new vaccines were introduced to the population. The "Motivating example: a COVID-19 clinical prediction model" section provides background on the illustrative example, the QCOVID series of survival prediction models [17, 18]. We investigate both one-time and dynamic updating using recalibration, refitting and Bayesian dynamic updating of time-to-event models. We consider data sources where updated data is additional follow-up time on the same individuals and also where updated data refers to data from a new cohort of individuals. These methods are described in the "Methods for model updating and assessment" section. We use a simulation study, presented in the "Simulation study" section, to investigate the performance of multiple updating methods with a focus on calibration, discrimination and variability of hazard ratio estimates to evaluate the performance. In the "Updating a COVID-19 clinical prediction model" section, we illustrate these methods using the QResearch database of electronic health care patient data [19] to predict the risk of contracting and dying from COVID-19. We finish with a discussion in the "Discussion" section.

## Motivating example: a COVID-19 clinical prediction model

Since the beginning of the COVID-19 pandemic in early 2020, numerous clinical prediction models have been proposed to assist both clinical decision-making and policymakers [20]. Among them, the QCOVID series of risk prediction algorithms were developed to help identify those most at risk of getting infected and then dying due to COVID-19 based on individual characteristics such as age, sex and long-standing illnesses [17, 18]. The original QCOVID model was used to prioritise people for vaccination and to inform the government's shielding list [17]. It was later refit with data from the second pandemic wave in the UK (QCOVID2) and extended to account for COVID-19 vaccination (QCOVID3) [18]. These models were developed in a rapidly changing environment in terms of the availability of vaccines, infection prevalence, levels of immunity in the population, new variants and availability of treatments.

Motivated by the need to keep prediction models such as QCOVID up to date to provide accurate predictions of risk, we apply different updating strategies to a model for prediction of the risk of catching and dying from COVID-19 (details found in the "Methods" section). We use a subset of QResearch data, the same database used to develop the QCOVID models. QResearch is an anonymised database of health records from GP practices throughout the UK that contains historical and current information on over 35 million individuals [19].

## Methods for model updating and assessment

### Dynamic model updating

Dynamic model updating is a process whereby a prediction model is repeatedly updated with new information. The process begins during period $u=0$ with an original model $M_0$ fit using a development dataset $D_0$. Dataset $D_0$ contains information from time $t_{-1}$ through time $t_0$. After time $t_0$, new information becomes available during period $u=1$. This new data can be collected to form dataset $D_1$ which contains information from $(t_0, t_1]$. The time period between updates may be fixed or variable and may be as short as the time it takes to receive one new data point or based on a fixed passage of time, e.g. 1 month, 1 quarter or 1 year. Using this new dataset, $D_1$, out-of-sample survival predictions are made at a clinically relevant prediction horizon, $v$, using model $M_0$, and performance is measured (see the "Performance assessment" section). Model $M_0$ is then updated with the information in $D_1$ and the process repeats. New data is collected $D_u$, predictions are made using the model $M_{u-1}$, performance is measured and, finally, model $M_{u-1}$ is updated with data $D_u$ resulting in model $M_u$.

### Performance assessment

The predictive performance of the original and updated models can be assessed on out-of-sample predictions by discrimination, calibration and overall performance [1, 21]. In time-to-event modelling, given a pair of individuals with known survival times, discrimination refers to the model's ability to assign a greater survival probability to the one who survived longer. We use an inverse probability of censoring weighted (IPCW) C-index to account for censoring [22]. Calibration assesses how well predicted outcomes match observed outcomes and we measure weak calibration using calibration intercept and slope [21, 23–25]. We use an IPCW Brier score, which is the mean-squared error for binary outcomes and predictions that are probabilities, to measure overall performance [26].

### Methods for model updating

In this section, we review three clinical prediction model updating methods: intercept recalibration, refitting and Bayesian updating and discuss their use in the context of updating a survival prediction model. We will use the term "no updating" to refer to retaining the original model ($M_0$) without using new data to update it. We assume the original model to be updated is a Cox proportional hazards model [27], which can be written as:

$$h_i(t \mid X_i) = h_0(t)\exp(\beta^{\mathrm{T}}\mathrm{X_i}) \quad (1)$$

where $h_i(t \mid X_i)$ is the hazard for the individual $i$ at time $t$, $h_0(t)$ is the baseline hazard, $X_i$ is a vector of time-fixed covariates and $\beta$ is a vector of parameters to be estimated (the log hazard ratios). An estimate of the predicted survival probability for person $i$ can be computed using:

$$S_i(t \mid X_i) = \exp\left\{-\hat{H}_0(t)\exp(\hat{\beta}^T X_i)\right\} \quad (2)$$

where $\hat{H}_0(t)$ is Breslow's estimate of the cumulative baseline hazard [28]. Note that with each update, we reset the time $t$ to $t = 0$ to compute the survival probabilities.

#### *Intercept recalibration*

For a Cox proportional hazards model, recalibration of the intercept refers to re-estimating the baseline hazard using the new data while holding the log hazard ratios estimated on the original dataset, $\hat{\beta}$, constant. In period $u$, we first calculate the linear predictor, $\eta_{i,u}$, for the new data using $\hat{\beta}$ from the original model and covariates $X_{i,u}$ from the new data. A Cox model is fit with $\eta_{i,u} = \hat{\beta}^T X_{i,u}$ as the only covariate and its coefficient $\beta_\eta$ is fixed at 1. We then estimate the cumulative baseline hazard $H_{0,u}(t)$ at the prediction horizon by setting the value of the linear predictor to zero. Recalibrated survival predictions are computed on the new data as:

$$S_{i,u}(t \mid X_{i,u}) = \exp\left\{-\hat{H}_{0,u}(t)\exp(\eta_{i,u})\right\} \quad (3)$$

Because the predictor coefficients are not updated, recalibration will only affect model performance in terms of calibration; discrimination will not change. This method does not accommodate the addition of new predictors to the model.

#### *Refitting*

Refitting is the most extreme type of updating because the previously estimated predictor coefficients and baseline hazard are discarded. Therefore, it readily accommodates new predictors. In general, a model could be updated by refitting to new data only or to some combination of old and new data [3, 11]. An advantage of refitting to new data only is that the updated model more quickly reflects the new environment but the update may be made on a smaller sample size than the original model development dataset [9]. Also, if the new data reflects a temporary change in the data generating process or contains data from only a particular segment of the population, the updated model may be overfit to this new data and future performance may be poor [29].

#### *Bayesian dynamic updating*

Bayesian model updating is a technique for combining knowledge gained in previous models with information in the new data that arises as time goes on and that is available for making an update to the model. Applying the Bayesian updating technique of McCormick et al. [30], which focused on a logistic outcome model, to proportional hazards regression and assuming exponentially distributed survival times, the model in period $u$ for $u \geq 1$ can be written:

$$\begin{aligned} T_i &\sim Exp(\omega_i) \\ \omega_i &= \lambda_u + \beta_u^T X_i \\ \beta_u &\sim N(\hat{\beta}_{u-1}, \hat{\Sigma}_{u-1}/\xi) \\ \lambda_u &\sim N(0, \sigma_\lambda) \end{aligned} \tag{4}$$

where $T_i$ is an individual $i$'s survival time and $\lambda_u$ is the log baseline hazard in period $u$. $\hat{\beta}_{u-1}$ is the vector of coefficient estimates from the prior period, $\hat{\Sigma}_{u-1}$ is the covariance matrix from the prior period and $\xi \leq 1$ is a 'forgetting factor'. The forgetting factor controls the level of uncertainty in the prior; a smaller $\xi$ yields a less informative prior. $\sigma_\lambda$ was chosen to yield a vague prior. Other distributions of survival times may be used, e.g. Weibull. For further details, see Additional file 1: Appendix A. Survival predictions can be generated from the posterior predictive distribution. Parameter estimates are taken as the median of the posterior distribution.

## Simulation study

### Design

#### *Overview and aims*

The goal of this simulation study is to assess methods for dynamic model updating of clinical survival prediction models in changing environments. Specifically, we aim to identify settings where specific methods outperform others and explore how different types of observational datasets affect the performance of the updating methods. We consider *cohort with replacement* datasets, where the initial cohort is determined at the start of the study but individuals who have an event are replaced in the cohort with a new member, as well as *new cohorts* datasets, in which membership will change each period based on receiving a particular treatment or diagnosis. The simulation scenarios are inspired by our motivating example and thus are characterised by low event rates, predictors with low prevalence in the population, introduction of new treatments and changing baseline risk.

#### *Data generating mechanisms*

Because the goal is to study the process of updating a previously developed 'original' prediction model, we first generated data $D_0$ and fit a model $M_0$ to serve as the starting point. For each individual, 4 covariates were generated: a variable representing age in decades, $X_1 \sim U(1.8, 9.5)$; a continuous risk factor (e.g. representing a biomarker), $X_2 \sim N(1, 1)$; a binary risk factor (e.g. a co-morbidity indicator), $X_3 \sim Bern(p_{X_3})$; and a treatment variable, $X_4 \sim Bern(p_{X_4})$. $p_{X_3}$ and $p_{X_4}$ vary by scenario and, in some cases, depend on age and on each other. A complete list of parameter values used in the simulation can be found in Additional file 1: Table S1. Given log hazard ratios $\beta_1, \beta_2, \beta_3, \beta_4$, we generated survival times assuming an exponential distribution with baseline hazard $e^\lambda$. Survival times were censored at 1 year. A Cox proportional hazards model was then fit to this dataset to create the original model.

We then simulated new data $D_1, \ldots, D_5$, i.e. data arriving in 5 time periods, each of length 3 months, after development of the original model. The $D_u (u = 1, \ldots, 5)$ were generated according to two frameworks which we refer to as: *cohort with replacement* and *new cohorts*. Each $D_u$ contains data for 3 months. In the *cohort with replacement* framework, follow-up time is added to the existing individuals in $D_1$ each month and a small number of new individuals is added to replace those who have had events. We reset time to zero at the start of each quarter. This data generation method aims to mimic the use of electronic health records to assemble a cohort to study the risk of catching an infectious disease and having a related adverse event. There is some churn as new individuals can join the cohort, e.g. new patients in a GP practice.

In contrast, in the *new cohorts* framework, data are simulated for a new (different) group of individuals each month as might be the case if the data source was all individuals with a positive COVID-19 test in that month. Individuals in the first month of the 3-month period are followed for 3 months, while individuals in the second and third months of the period are only followed for 2 and 1 month, respectively. Many individuals are followed up for less time than the prediction horizon (3 months) yielding data with relatively fewer observations at longer times, and, therefore, fewer observed events. Figure 1 illustrates sample follow-up times for individuals under *cohort with replacement* (left) and *new cohorts* (right). Further details of the data generation for both styles are found in Additional file 1: Appendix B and Fig. S1.

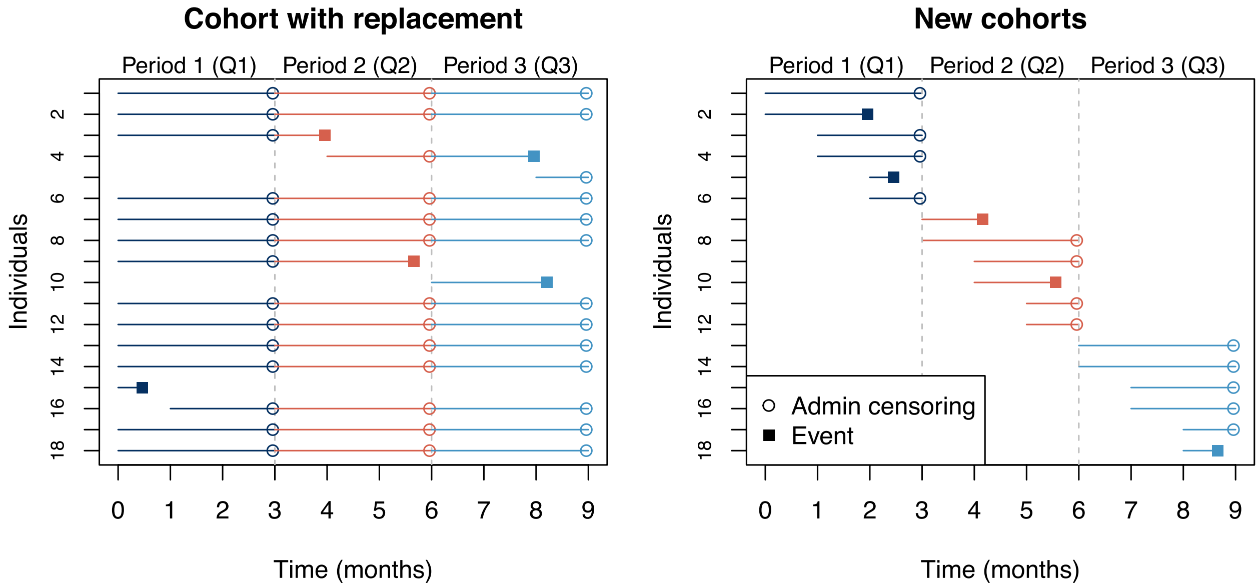


**Fig. 1** Sample follow-up times for 18 simulated individuals under *cohort with replacement* (left) and *new cohorts* (right) frameworks. On the left, most individuals are part of the Q1, Q2 and Q3 cohorts. When an individual has an event (e.g. individual 9 at time 5.7 months), they are replaced by a new person at the start of the following month (e.g. individual 10). On the right, each quarterly dataset consists of different individuals and those joining the new cohort after the first month of the quarter have a shorter follow-up time. For example, individuals 5 and 6 join at the beginning of month 3 and are followed up for one month or until an event occurs

For both data-generating mechanisms, 6 scenarios were studied. Table 1 provides a complete list and the associated parameter values for each scenario can be found in Additional file 1: Table S1. For each scenario and for both frameworks, we generated $n_{sim}$ =600 simulated datasets each with 10,000 individuals for initial model development. The chosen sample size of 10,000 was determined to be a sufficient size for the development of the prediction model [31]. Each updating cohort for *cohort with replacement* consisted of data on 10,000 individuals + 1000 replacements per simulated dataset per simulation scenario. Data on 1000 individuals per month were generated for the *new cohorts* framework per simulated dataset per scenario. In the reference scenario, there were no changes over time. Calibration drift scenarios were characterised by a new baseline hazard in each updated dataset reflecting changes in baseline risk over time. In the "Rare-1%" scenario, we assumed 1% of the total population has a rare risk factor placing them at increased risk of an event and the chance of having this

**Table 1** Listing of all simulation scenarios and the abbreviated name used in the "Results" section

| Group | Description | Scenario name |
|---|---|---|
| Reference scenario | | |
| | Constant baseline hazard; event rate 5% per year | Constant events |
| Scenarios with calibration drift | | |
| | Decreasing baseline hazard; event rate decreased from 5% per year in the first period to 2% in the last period | Decreasing events |
| | Increasing baseline hazard; event rate increased from 5% per year in the first period to 8% in the last period | Increasing events |
| Scenario with rare predictor | | |
| | Rare risk factor for an event found in 1% of people over age 55 | Rare-1% |
| Scenarios with new predictors | | |
| | Treatment introduced in Q2, made available by age group | New treatment |
| | Treatment introduced in Q2, made available by age group and to those with a comorbidity (5% of population) | New treatment + comorbidity |

risk factor increased with age. In the context of our motivating example, this could correspond to an age-related comorbidity such as dementia. The treatment indicator, $X_4$, was a time-varying binary predictor. To simulate the roll-out of a new treatment, such as a vaccine, $p_{X_4}$, the probability of being treated, was increased over time so that an increasingly large segment of the population received treatment. Drawing from the rollout of the COVID-19 vaccine in the UK, the population eligible for treatment was based on age with treatment being available to the elderly first and then to successively younger people. In the "new treatment + comorbidity" scenario, the treatment was made available to those who were immunocompromised ($X_3 = 1$) or met the minimum age criterion ($X_1$).

### *Targets*

The target is the predicted probability of survival to 3 months in the validation data $D_{u+1}$ using a model fit/updated with data from the previous period, $D_u$.

### *Methods*

In each simulation scenario, we fit a Cox proportional hazards model ("original model") to the initial development dataset. This model, $M_0$, is the starting point for all updating methods.

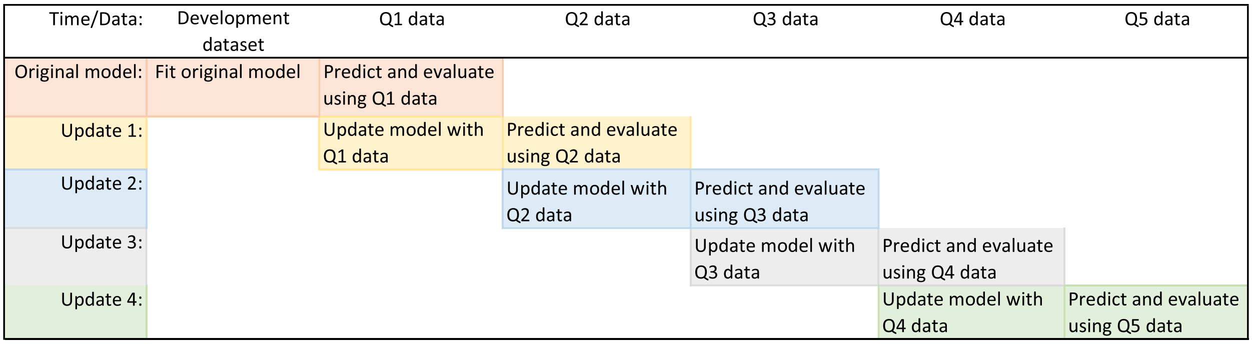
Four strategies, implemented at a single point in time and/or dynamically, were considered for updating the original model over the subsequent 1 year: (1) no update, (2a) refit once at a single time, (2b) refit quarterly, (3a) recalibrate the intercept once at a single time, (3b) recalibrate the intercept quarterly and (4) Bayesian dynamic updating quarterly. In the no update strategy, the original model was retained unchanged and applied to each dataset. The Bayesian model assumed an exponential baseline hazard. For refit quarterly, recalibrate quarterly and Bayesian update quarterly, the updates are performed dynamically each quarter as new data arrives.

The original model was updated with new data per each updating strategy. Each updated model was then used to predict 3-month survival on the subsequent quarter's data (Q2 - Q5) for evaluation (Fig. 2). In this way, we evaluate predictive performance out-of-sample. We used a quarterly updating cycle to mimic the situation where data becomes available for analysts on a quarterly cycle. Note that to fit a model for prediction of 3-month survival, 3 months of follow-up on at least a portion of the new dataset is required.

In addition to these dynamic strategies with quarterly updating, we also investigate one-time model updating that is not performed according to a pre-determined schedule. Model updating on an ad hoc basis when there are resources available or when the model starts performing poorly is common in practice. To simulate this, we imagine 7 different analysts faced with the task of updating the original model one time and having to choose when to do that update. We assume 4 of them select the start of a quarter while the remaining 3 select times in between those quarters (drawn randomly); therefore, the analysts will update with 3 months of data beginning at $t = 0.0$, 0.1, 0.25, 0.46, 0.5, 0.69 and 0.75 months after the start of year 1. Analysts who begin their update in the middle of a month only have access to the prior 3 months of data as we assume data is collected at the end of the month.

### *Performance measures*

The predictive performance of the updating methods was evaluated using calibration intercept and slope, C-index and Brier score as described in the "Performance assessment" section. We also compare the estimated hazard ratios to the true values used to generate the data.



| Time/Data: | Development dataset | Q1 data | Q2 data | Q3 data | Q4 data | Q5 data |
|---|---|---|---|---|---|---|
| Original model: | Fit original model | Predict and evaluate using Q1 data | | | | |
| Update 1: | | Update model with Q1 data | Predict and evaluate using Q2 data | | | |
| Update 2: | | | Update model with Q2 data | Predict and evaluate using Q3 data | | |
| Update 3: | | | | Update model with Q3 data | Predict and evaluate using Q4 data | |
| Update 4: | | | | | Update model with Q4 data | Predict and evaluate using Q5 data |

**Fig. 2** Illustration of the dynamic updating and evaluation process. Beginning at the top left, an original model was fit to the development dataset and evaluated out-of-sample on the Q1 new data. The model was then updated each quarter with new data and evaluated on the subsequent quarter's data. These updates are called Update 1, 2, 3 and 4 where update *u* was performed using data from Quarter *u*. A colour version of this figure can be found in the electronic version of the article

### *Implementation*

All analyses were conducted in R v4.0.2 [32]. Survival times were generated using the R package `simsurv` [33]. We used the `survival` package [34] for Cox proportional hazards regression and the `pec` package [35] to calculate C-index and Brier score. Calibration intercept and slope ("weak calibration" [36]) were computed using the method described by Crowson et al. [24].

Bayesian survival analysis was performed using the `rstanarm` package [37]. Estimation was via Markov chain Monte Carlo, specifically the No-U-Turn Sampler (Hamiltonian Monte Carlo) implemented in Stan [38]. We used 2 chains, each with 7500 iterations of which 1000 were burn-in. This was sufficient to obtain an effective sample size of 10,000 and a Monte Carlo standard error $\approx$ 1% of standard error of the parameter estimates. Convergence was assessed using Gelman and Rubin's Rhat statistic [39] with an Rhat $< 1.1$ required for convergence. For coefficients not present in the original model (e.g. new treatment) and the rate parameter $\lambda$, we set a prior distribution of N(0, 2.5); all other prior distributions were obtained as described in the "Bayesian dynamic updating" section. Typical forgetting factors are just below 1 (e.g. Raftery et al. [40] chose $\xi = 0.99$). McCormick et al. [30] advise that while less volatile processes may be well-represented by $0.99 < \xi < 1$, more volatile ones may require $0.90 < \xi < 1$. We found that performance was not sensitive to the choice of forgetting factor between 0.50 and 0.99 so we used $\xi = 0.9$ (see Additional file 1: Appendix D, Table S8 for a sensitivity analysis of the forgetting factor). If the model could not be refit due to insufficient events/covariate combinations, we retained the model from the previous period to reflect the reality that sometimes it is not possible to refit a model until more data is accrued.

## Results

In the reference scenario with a constant event rate, all methods showed similar discriminative ability (C-index). For calibration intercept and slope (target values of 0 and 1, respectively), results from the updating methods deviated more from the target values than those from no updating. Complete results are available in Additional file 1: Table S2 and Fig. S2.

### *Scenarios with calibration drift (decreasing events, increasing events)*

Discrimination and calibration intercept for the simulation using a decreasing event rate scenario are presented in Fig. 3. Complete results for the calibration drift scenarios can be found in Additional file 1: Tables S3, S4, Figs. S3, S4. For the *cohort with replacement* simulations, in both the increasing and decreasing event rate scenarios, all updating methods (including not updating) produced similar average C-index, Brier score and calibration slope. The average C-index for all methods in the final period was 0.81. Higher values of the C-index indicate better discriminative ability. Although none of the calibration slopes deviated by more than 0.02 from 1.0, the value of the calibration intercept for a model that was not updated moved further away from zero at each prediction time (see Fig. 3). In the decreasing event rate scenario, updating via quarterly Bayesian dynamic updating had the best calibration intercepts but other updating methods were within 0.02 of those values. Best calibration in the increasing event rate scenario was achieved by the quarterly recalibration strategy and, again, other updating methods were close. Models that were recalibrated or refit only once exhibited good calibration at some time points but, overall, quarterly updating strategies produced calibration intercepts closer to 0. Results for the *new cohorts* simulations followed a similar pattern

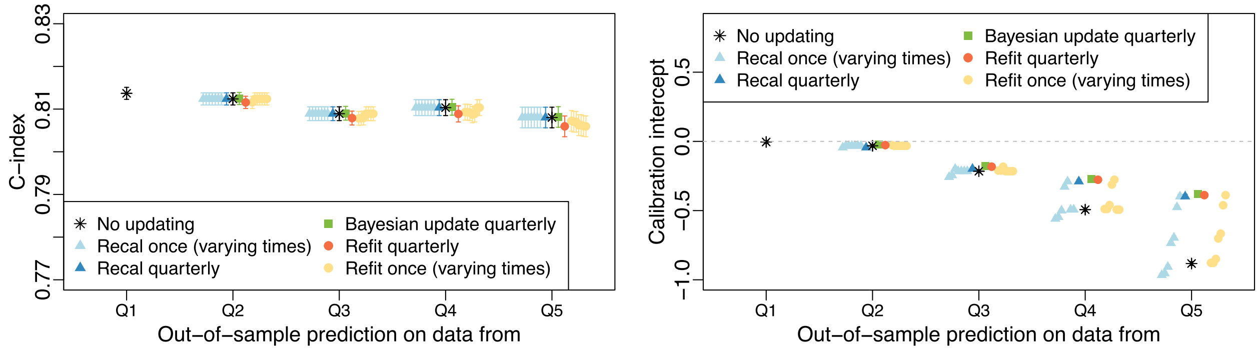


**Fig. 3** *Cohort with replacement* simulation results for a scenario with calibration drift. The left graphic shows the average C-index for each updating method across the 600 simulated datasets at each of the 5 prediction times for a scenario where the event rate decreased over time from 5% per year in Q1 to 2% per year in Q5. On the right, the average calibration intercept is shown for the same scenario. Results for 'Recal once' and 'Refit once' strategies are ordered by update time with the earliest time on the left

but more variability was seen in the estimates of calibration slope from all methods.

### Scenario with rare predictor (Rare-1%)

Predictive performance results for the rare predictor scenario, where 1% of the population had a risk factor for the event and the chance of having the risk factor increased with age are in Additional file 1: Table S5, Figs. S5, S6. In both simulated dataset styles, *cohort with replacement* and *new cohorts*, all updating methods, including no updating, performed similarly for calibration intercept, C-index and Brier score. In the *new cohorts* simulation, the calibration slope of quarterly refitting was 0.94 and 0.95 in the last two quarters compared to 0.99 and 0.99 for the other methods. Differences were also seen in the hazard ratio estimates (see the "Hazard ratio estimates" section below).

### Scenarios with new predictors (new treatment, new treatment + comorbidity)

Full simulation results for the two scenarios with new predictors can be found in Additional file 1: Tables S6, S7, Figs. S7, S8. In these scenarios, a new treatment was introduced at the beginning of Q2 and rolled out to an increasing percentage of the population over time based on age group ("New treatment") or based on age and presence of a comorbidity ("New treatment + comorbidity"). In the new treatment *cohort with replacement* simulation, Bayesian updating, refitting quarterly and one-time refitting after the introduction of the new treatment offered the best calibration intercept and refitting strategies had the best calibration slope. Recalibration strategies generally had a calibration intercept closer to zero than no updating. All methods had a similar Brier score. Refitting and Bayesian updating strategies showed improvements in discrimination over no updating or recalibration strategies with differences in average C-index from 0.03 to 0.07 over no updating in the Q3–Q5 predictions. In the *new cohorts* simulation, quarterly refitting and Bayesian updating offered calibration intercepts ranging from −0.02 to 0.02 in Q3–Q5 compared to −0.58 to −1.18 for no update and −0.25 to −0.33 for quarterly recalibration (see Fig. 4). Bayesian updating produced superior discrimination, with a C-index significantly higher than all other methods in Q3–Q5 (Wilcoxon signed rank test $p$ <0.001).

Similar to the new treatment scenario, in the *cohort with replacement* simulation under the new treatment + comorbidity scenario, Bayesian updating, quarterly refitting and refitting after the introduction of the treatment produced the best calibration intercepts. Brier scores were similar for all updating methods. Bayesian updating and quarterly refitting produced the best average C-index values (0.82, 0.79, 0.76, 0.76) for the four updates compared to (0.82, 0.79, 0.74, 0.72) for no updating and quarterly recalibration. In the *new cohorts* simulation, Bayesian updating had the calibration intercept closest to zero after the introduction of the new treatment (−0.04, −0.01, 0.02). Quarterly recalibration produced better calibration intercepts (−0.29, −0.55, −0.18) than refitting (−0.45, −0.85, −0.76) due to the inability of the model to be refit on some datasets. The highest average C-index came from Bayesian dynamic updating (0.83, 0.83, 0.81, 0.83) (Wilcoxon signed rank test $p$ < 0.05 in Q3–Q5).

As there are small numbers of individuals with the comorbidity who also received the new treatment, in the *new cohorts* datasets it was not always possible to refit the model at each time point: 5% (Q2), 25% (Q3) and 32% (Q4) of models were able to be refit from the 600 simulated datasets. These percentages were higher for *cohort with replacement* datasets: 97% (Q2), 99.5% (Q3) and 65% (Q4). The percentage of simulated data sets in which

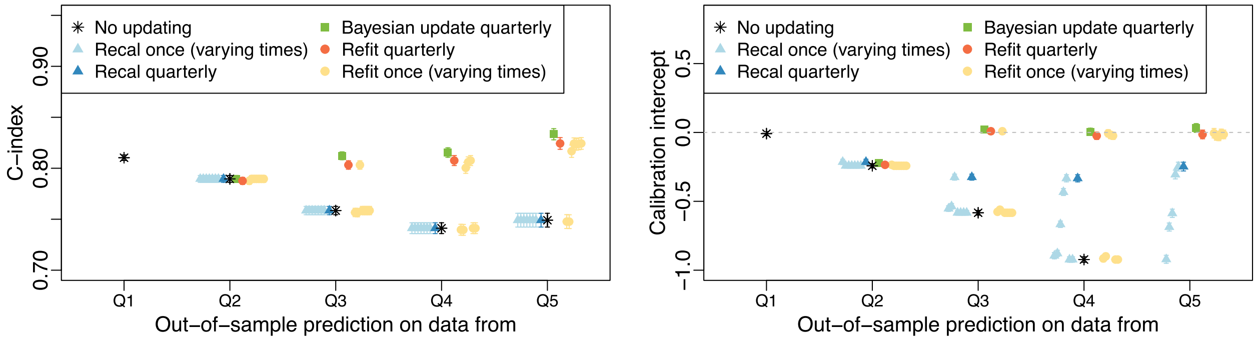


**Fig. 4** *New cohorts* simulation results for the new treatment scenario. The left graphic shows the average C-index for each updating method across the 600 simulated datasets at each of the 5 prediction times for the scenario where a new treatment was introduced at the beginning of Q2. On the right, the average calibration intercept is shown for the same scenario. Results for 'Recal once' and 'Refit once' strategies are ordered by update time with the earliest time on the left

refitting was feasible rises over time in the *new cohorts* datasets as more people are eligible for the treatment and, therefore, it is more likely that there are individuals with the comorbidity who are both treated and have an event. In the *cohort with replacement*, however, the percentage able to be fit in Q4 is lower than the previous quarters because there are very few people who have the comorbidity and are still untreated and even fewer who also have an event.

### *Hazard ratio estimates*

While predictive performance is not measured based on estimates of the parameters of prediction models, (their magnitude or their precision), in the dynamic updating setting users may be interested in the how parameter estimates change after updating. Changes in underlying hazard ratios over time are likely to be reflected in changes in predictive performance and in an updated model providing improved predictive performance. Users of the model will expect a smooth time series of hazard ratio estimates as predictor-outcome relationships are generally not rapidly changing. In this simulation study, across all scenarios, the hazard ratio estimates obtained with refitting strategies showed more variability at each update and across time than those obtained by Bayesian dynamic updating. For example, in the decreasing events scenario, Bayesian dynamic updating log hazard ratio estimates for all three predictors ($\beta_1, \beta_2, \beta_3$) showed less volatility between time points than refitting strategies. Further, Bayesian updating produced log hazard ratio estimates with less bias than the refitting strategies (Fig. 5). Note that hazard ratios are not updated by intercept recalibration.

Looking at the coefficient estimates in the *cohort with replacement* simulation for the rare predictor ($\beta_3$) in the rare-1% scenario, the average of the original model estimates for the log hazard ratio was 0.76 (MCSE 0.01) compared to the true value of 0.8. The average estimated log hazard ratio for Bayesian updating after each of the 4 updates was 0.78, 0.79, 0.80, 0.81 with an MCSE of 0.01 at each time. The refitting strategies showed more variability in the estimates of the log hazard ratio with MCSE of 0.02 at each time and more bias, with log hazard ratio estimates of 0.73, 0.73, 0.75, 0.73 at the 4 update times (see Additional file 1: Fig. S6).

## Updating a COVID-19 clinical prediction model

### Methods

To fit and evaluate a COVID-19 survival prediction model, data on 1,000,000 individuals aged 18 years or older were obtained from the QResearch database (version 46) for the period 24/01/2020 (the first date where cases were reported in the UK) to 30/04/2021. For consistency with Hippisley-Cox et al. [18], our chosen outcome was predicted 70-day survival from COVID-19-related death as determined by death certificate information or death within 28 days of a positive COVID-19 test. Predictors for each individual were: age, body mass index (BMI), sex, type 1 diabetes, chronic obstructive pulmonary disease (COPD), dementia, and UK region. While this represents a subset of the predictors used in the QCOVID models [17, 18], our purpose was not to develop a new COVID-19 prediction model, nor to update an existing one with all their complexities but rather to illustrate methods for model updating. Non-COVID-19-related deaths

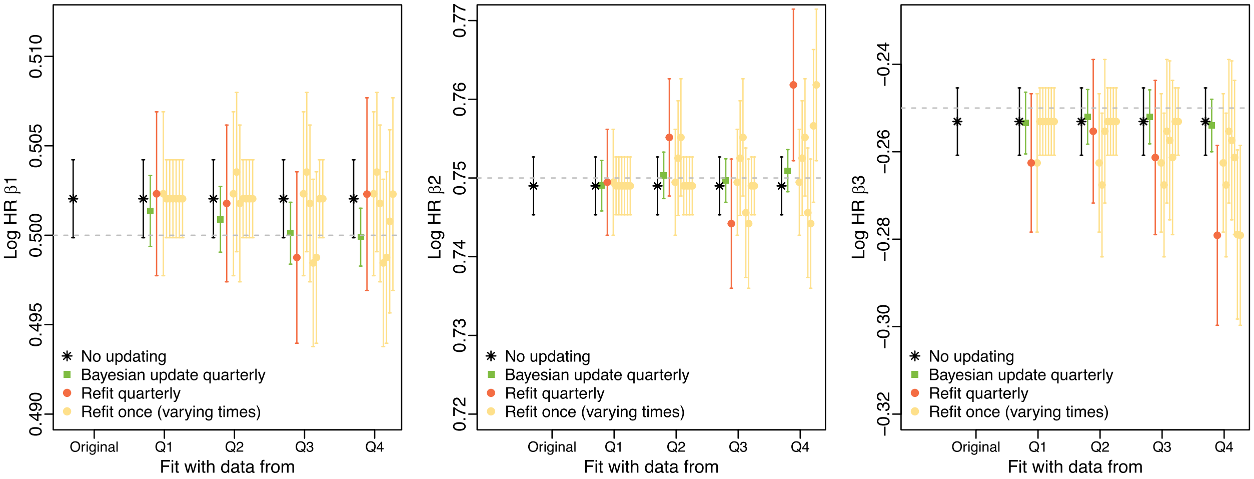


**Fig. 5** *Cohort with replacement* simulation results for the decreasing events scenario. From left to right, the graphics show the estimated log hazard ratios for $\beta_1, \beta_2, \beta_3$ for the original model, Bayesian updating and the refitting strategies. Intercept recalibration strategies are not shown because hazard ratios are not re-estimated

represent a competing risk. Therefore, we fit a sub-distribution hazard model where everyone not experiencing a COVID-19 death, (including those who died of other causes) was censored at the end of the study period [41]. As 18% of baseline BMI values were missing, we added a missing indicator to the model. Both BMI and age were scaled to have a mean of 0 and a standard deviation of 1 and then a natural cubic spline was fitted to each with 3 internal knots. These knots were fixed throughout the updating so that previous model estimates would be valid priors for subsequent models. Additionally, 7-day rolling COVID-19 case rates by region were included as a predictor [42] beginning with the first update. We divided the study period into 5 batches of data to create one development dataset followed by four updating datasets as follows: Period 1 (24 Jan–30 Apr 2020), Period 2 (1 May–31 Jul 2020), Period 3 (1 Aug–31 Oct 2020), Period 4 (1 Nov 2020–31 Jan 2021) and Period 5 (1 Feb–30 Apr 2021). An initial sub-distribution hazard model with main effects only was fit to the period 1 dataset with baseline characteristics measured up to the study period start date of 24 January 2020 and follow-up continuing through 30 April 2020. Each subsequent dataset contained information on co-morbidities and regional case rates updated up to the period start date with follow-up continuing up to the end date of the period.

Intercept recalibration, refitting and Bayesian updating were applied to each updating dataset in succession. The resulting updated prediction models were evaluated out-of-sample on the next period's data for discrimination, calibration and overall predictive performance. For comparison, the original model (without updating) was also evaluated on each updating dataset. Both intercept recalibration and refitting used a proportional hazards model. Bayesian models using a constant baseline hazard were estimated using Stan [43]. We used 6 chains, each with 3000 iterations of which 1000 were burn-in to obtain a Monte Carlo standard error approximately less than or equal to 1% of the standard error of the parameter estimates. The effective sample size was at least 5000 and Gelman and Rubin's Rhat statistic [39] was $< 1.01$ indicating convergence. The forgetting factor was 0.9.

## Results

Additional file 1: Table S9 presents baseline characteristics of the study population. Table 2 presents performance characteristics for each updating method (recalibrate, refit and Bayesian update) and for the original model without any updating. No single updating method gave the best performance across all periods in any of the performance metrics. While refitting produced a higher C-index (0.93) than not updating (0.92) for the first evaluation time, at the second evaluation time refitting had a lower C-index (0.91) than both Bayesian updating (0.92) and no updating (0.94). Refitting was equal to no updating and recalibration (0.91) at the final evaluation time and higher than Bayesian updating (0.90). The C-index for the model after the first Bayesian update (0.76) was the lowest of any method at any time. The Brier scores for all methods in all periods were $< 0.001$. Calibration intercepts from Bayesian updating and refitting were closest to 0 for updates using period 2 and 3 data but the original model, without being updated, showed the best calibration intercept in the final updating period.

**Table 2** Performance of intercept recalibration (Recal), refitting (Refit), and Bayesian dynamic updating (Bayes) methods to update the prediction model for 70-day COVID-19-related death. The original model was fit using data from period 1 and evaluated using data from period 2. The original model was then updated each period with new data and evaluated using the following period's data

| Model fit w/ data from: | Evaluated w/ data from: | Recal | Refit | Bayes | No update$^{\dagger}$ | Recal | Refit | Bayes | No update$^{\dagger}$ |
|---|---|---|---|---|---|---|---|---|---|
| | | C-index | | | | Brier Score | | | |
| Period 1 | Period 2 | | | | 0.95 | | | | 3E−04 |
| Period 2 | Period 3 | 0.92 | 0.93 | 0.76 | 0.92 | 3E−05 | 3E−05 | 3E−05 | 3E−05 |
| Period 3 | Period 4 | 0.94 | 0.91 | 0.92 | 0.94 | 8E−04 | 7E−04 | 8E−04 | 7E−04 |
| Period 4 | Period 5 | 0.91 | 0.91 | 0.90 | 0.91 | 4E−04 | 4E−04 | 4E−04 | 4E−04 |
| | | Calibration intercept | | | | Calibration slope | | | |
| Period 1 | Period 2 | | | | −0.82 | | | | 1.10 |
| Period 2 | Period 3 | -0.79 | 0.30 | 0.15 | −2.12 | 0.92 | 0.82 | 0.86 | 0.92 |
| Period 3 | Period 4 | 3.16 | −0.04 | −0.01 | 0.56 | 0.93 | 0.90 | 0.86 | 0.93 |
| Period 4 | Period 5 | −0.66 | −1.10 | −1.12 | −0.49 | 0.85 | 0.89 | 0.88 | 0.85 |

$^{\dagger}$No update refers to the original model fit in period 1 and evaluated in each subsequent period without any updating

## Discussion

In this study, we investigated the performance of discrete and dynamic updating methods for clinical survival prediction models. Overall, the dynamic updating strategies at regular intervals outperformed a single update. We found that Bayesian dynamic updating offered the best performance in the simulation study in situations with new predictors and less data and that all methods generally improved calibration. In our motivating example, where the environment was changing rapidly and the outcome was rare, no single updating method outperformed the others.

Although no one method performed best in all circumstances, we can draw several conclusions from the study. First, intercept recalibration is an effective tool for calibration maintenance that requires little data, is not computationally intensive and will not change the rank order of predicted survival probabilities between two individuals. This simplifies reporting of the updated model and may be less confusing for users. Recalibration may also be useful when new predictors are introduced but insufficient data has accumulated for a full model refit, as was seen in the new treatment + comorbidity scenario. Second, although refitting can produce a good performing model when adequate data is available, it does not outperform Bayesian updating in general, even when there are new predictors. As in the binary outcome setting, because refitting requires a large number of observations/events and it may produce abruptly changing hazard ratio estimates over time, refit models have the greatest chance of being overfit and are the most likely to produce substantially different predictions for an individual compared to the previous model. Riley et al. [12] caution that attempts to ameliorate this overfitting by applying shrinkage techniques may be unreliable due to estimation uncertainty of the tuning parameters and is best used with a larger sample size. In an unchanging environment, updating may lead to poorer performance than not updating, particularly if the sample size of the updating dataset is small.

Bayesian dynamic updating offers the advantages of both recalibration and refitting and, in the simulation study, was the best performer in the majority of scenarios and time points across the evaluation criteria. However, it is the most computationally intensive updating method we studied with a single update taking 12–24 h using the QResearch dataset of 1,000,000 records. Also, the baseline hazard must be modelled parametrically. We selected an exponential parametrisation for ease of exposition and because over this short horizon, the baseline hazard may be reasonably assumed to be constant. More flexible specifications may be chosen but will come with a higher computational cost. Although the detail of how a Bayesian model is estimated would be complicated to explain to non-statisticians, we believe the concept of a Bayesian update — that it combines knowledge from the current model with information from new data — is intuitively appealing. The relatively stable hazard ratio estimates are a further advantage and may help engender trust amongst users about the updating process. Although our simulation study found that performance was insensitive to choice of the forgetting factor, in other situations this might not be the case and the forgetting parameter could be viewed as a quantity to be tuned. McCormick et al. [30] acknowledge the computational burden of tuning the forgetting factor and propose an approach where two values are considered at each update: forgetting and no-forgetting.

In the COVID-19 application, the discrimination of the first Bayesian updated model was poor. This was primarily due to the method for obtaining the priors for the first update. The original model was fit using a Cox proportional hazards model and the first priors were taken from this model. This mimics the situation where the analyst(s) updating the model are different from the analyst(s) who originally developed it and they may not have access to the original development dataset. In this case, the fitted model is the only source of information for the priors. The poor performance in the first update occurred because the Bayesian updating assumed exponentially distributed survival times whereas the original model made no such assumption and, therefore, the coefficients were estimated with a different baseline hazard. Had the actual survival times been exponentially distributed (as in the simulation study) obtaining the priors from a Cox fit would have produced the same priors as those from an exponential model. We recommend care in obtaining priors when access to the development data is not possible. Also, more complex models of the baseline hazard could be used in the Bayesian model but these come with increased computational cost.

The results from the illustration of updating a model that predicts catching and dying from COVID-19 were inconclusive. Different updating strategies, including no updating, performed well at different times with different metrics. During the study period, the UK experienced multiple waves of COVID-19. Therefore, a model updated during a time of high prevalence could be tested out-of-sample at a time of low prevalence and good calibration was difficult to maintain. We hypothesise that a model predicting risk of COVID-19-related death in those with a positive COVID-19 test would be less susceptible to these cycles. Also, our example was constructed using a 3-month updating period to ensure sufficient events in each dataset to allow for refitting. However, a monthly updating cycle using Bayesian dynamic updating may have performed better. Interesting future work would be to update a COVID-19 prediction model including the period that saw the introduction of vaccines in the UK.

In the simulation study, refitting performed better in the *cohort with replacement-style* data scenarios than in those with *new cohorts*. This result is due to the smaller amount of data available in the *new cohorts* dataset and the fact that, in many cases, we were unable to fit a new model to the new data only. Analysts may consider using some combination of old and new data to overcome the lack of new data but how much old versus new data to include is subjective and the choice can impact the refit prediction model. Schnellinger et al. [11] studied 4 sliding window lengths for updating a logistic regression model and found that although the window length did not affect performance of recalibration techniques, including more old data improved performance of refitting on most metrics. In both binary and time-to-event outcome settings, Bayesian dynamic updating is well-suited to the small sample size case as the model can be successively updated with new evidence without waiting for a dataset as large as the original development dataset to accumulate. The choice of how often to update is a question that will depend upon the availability of new data and the data requirements of the selected updating method. An interesting area for future research is how to adaptively determine when to update the prediction model. For example, an update could be triggered when one or more minimum performance thresholds were crossed such as C-index worsening by more than 5%. Alternatively, algorithms may be implemented to detect changes in the underlying distribution of the data, called concept drift [29, 44]. In the clinical prediction model literature, Davis et al. [10] has proposed an adaptive windowing approach for detecting calibration drift that could be used to inform update timing.

We powered the simulation study to detect a difference of 0.01 in the C-index but it is difficult to know how big a difference in each of the performance criteria would be clinically significant. For example, we found evidence of statistically different values of the C-index and Brier score based on a Wilcoxon signed rank test even when the values themselves were identical to two significant digits. It is unlikely that these differences are relevant in a practical sense. Also, when event rates are very low, as in the COVID-19 application, Brier scores may not be informative for model comparison because all methods are likely to predict a low event probability for most individuals. When averaged, these small differences overwhelm the few cases where predicted risk is higher and the resulting Brier scores will be small.

Although our focus was on time-to-event outcomes, many of our conclusions are applicable to the binary outcome case as well. In particular, the choice of updating method should be carefully selected considering the available data, existence of new predictors and subject matter knowledge. Equally important is the development of a strategy for ongoing dynamic updating including the recurring collection of new data to capture environment changes and distributional shifts. A plan for communicating the updated model and refreshing web pages and calculators is also required when implementing any dynamic strategy. We also wish to caution against over-automating the updating process as clinical input may identify trends and environmental changes before they are evident in the data.

## Supplementary information

The online version contains supplementary material available at https://doi.org/10.1186/s41512-023-00163-z.

**Additional file 1.** Supplementary information for dynamic updating of clinical survival prediction models in a rapidly changing environment. Details of simulation study data generation, full simulation study results, sensitivity analysis of forgetting factor, QResearch study population characteristics.

**Acknowledgements**
We acknowledge the contribution of EMIS practices who contribute to QResearch®and EMIS Health and the the Universities of Nottingham and Oxford for expertise in establishing, developing and supporting the QResearch database. Linked mortality data was provided by the Office for National Statistics. The Office of National Statistics bears no responsibility for the analysis or interpretation of the data.

**Authors' contributions**
JHC secured the funding. All authors contributed to the conceptualisation of this study. KTT, RHK, and KDO designed the methodology with input from CC. JHC provided clinical insights and JHC and CC provided information about the QResearch dataset. KTT conducted the statistical analyses and drafted the initial manuscript. All authors commented on early drafts and approved the final version of the manuscript.

**Funding**
This study is funded by the National Institute for Health Research (NIHR) following a commission by Department of Health and Social Care. RHK is supported by a UKRI Future Leaders Fellowship (MR/S017968/1). KDO is funded by a Royal Society-Wellcome Trust Sir Henry Dale fellowship, grant number 218554/Z/19/Z. The funder had no role in the design of the study, or the collection, analysis, and interpretation of the data, or in the writing of the manuscript.

**Availability of data and materials**
The electronic health record data that support the findings of this study were provided by QResearch but restrictions apply to the availability of these data, which were used under license for the current study, and so are not publicly available. Interested researchers may apply for access to the data directly with QResearch.

## Declarations

**Ethics approval and consent to participate**
This study protocol was reviewed and approved by the London School of Hygiene and Tropical Medicine Observational/Interventions Research Ethics Committee (Reference 26936/RR/27732). The study also received a favourable scientific opinion from the QResearch Scientific Committee. QResearch is a

Research Ethics Approved Research Database, confirmed by the East Midlands - Derby Research Ethics Committee (REC reference 18/EM/0400).

### Consent for publication

Not applicable.

### Competing interests

JHC is an unpaid director of QResearch, a not-for-profit organisation which is a partnership between the University of Oxford and EMIS Health who supply the QResearch database used for this work.

### Author details

[1]Department of Medical Statistics, London School of Hygiene and Tropical Medicine, London WC1E 7HT, UK. [2]Nuffield Department of Primary Health Care Sciences, University of Oxford, Oxford OX2 6HT, UK. [3]Centre for Academic Primary Care, University of Nottingham, Nottingham NG7 2UH, UK. [4]Department of Statistical Science, University College London, London WC1E 6BT, UK.

Received: 16 June 2023 Accepted: 17 October 2023
Published online: 12 December 2023

### References

1. Steyerberg EW. Clinical prediction models: a practical approach to development, validation and updating. Springer; 2009.
2. Moons KGM, Kengne AP, Grobbee DE, Royston P, Vergouwe Y, Altman DG, et al. Risk prediction models: II. External validation, model updating, and impact assessment. Heart. 2012;98:691–698. https://doi.org/10.1136/heartjnl-2011-301247.
3. Hickey GL, Grant SW, Caiado C, Kendall S, Dunning J, Poullis M, et al. Dynamic prediction modeling approaches for cardiac surgery. Circ Cardiovasc Qual Outcome. 2013;6:649–58. https://doi.org/10.1161/CIRCOUTCOMES.111.000012.
4. Janssen KJM, Moons KGM, Kalkman CJ, Grobbee DE, Vergouwe Y. Updating methods improved the performance of a clinical prediction model in new patients. J Clin Epidemiol. 2008;61:76–86. https://doi.org/10.1016/j.jclinepi.2007.04.018.
5. Sim J, Teece L, Dennis MS, Roffe C, Dennis M, Kalra L, et al. Validation and recalibration of two multivariable prognostic models for survival and independence in acute stroke. PLoS ONE. 2016;11:1–17. https://doi.org/10.1371/journal.pone.0153527.
6. Siregar S, Nieboer D, Vergouwe Y, Versteegh MIM, Noyez L, Vonk ABA, et al. Improved Prediction by Dynamic Modeling: An Exploratory Study in the Adult Cardiac Surgery Database of the Netherlands Association for Cardio-Thoracic Surgery. Circ Cardiovasc Qual Outcome. 2016;9:171–81. https://doi.org/10.1161/CIRCOUTCOMES.114.001645.
7. Vergouwe Y, Nieboer D, Oostenbrink R, Debray TPA, Murray GD, Kattan MW, et al. A closed testing procedure to select an appropriate method for updating prediction models. Stat Med. 2017;36:4529–39. https://doi.org/10.1002/sim.7179.
8. Su TL, Jaki T, Hickey GL, Buchan I, Sperrin M. A review of statistical updating methods for clinical prediction models. Stat Methods Med Res. 2018;27:185–97. https://doi.org/10.1177/0962280215626466.
9. Davis SE, Greevy RA, Fonnesbeck C, Lasko TA, Walsh CG, Matheny ME. A nonparametric updating method to correct clinical prediction model drift. J Am Med Inform Assoc. 2019;26:1448–57. https://doi.org/10.1093/jamia/ocz127.
10. Davis SE, Greevy RA, Lasko TA, Walsh CG, Matheny ME. Comparison of Prediction Model Performance Updating Protocols: Using a Data-Driven Testing Procedure to Guide Updating. AMIA Annu Symp Proc. 2020;2019:1002–10.
11. Schnellinger EM, Yang W, Kimmel SE. Comparison of dynamic updating strategies for clinical prediction models. Diagn Prognostic Res. 2021;5:1–10. https://doi.org/10.1186/s41512-021-00110-w.
12. Riley RD, Snell KIE, Martin GP, Whittle R, Archer L, Sperrin M, et al. Penalization and shrinkage methods produced unreliable clinical prediction models especially when sample size was small. J Clin Epidemiol. 2021;132:88–96. https://doi.org/10.1016/j.jclinepi.2020.12.005.
13. Feng J, Gossmann A, Sahiner B, Pirracchio R. Bayesian logistic regression for online recalibration and revision of risk prediction models with performance guarantees. J Am Med Inform Assoc. 2022;29:1–12. https://doi.org/10.1093/jamia/ocab280.
14. Booth S, Riley RD, Ensor J, Lambert PC, Rutherford MJ. Temporal recalibration for improving prognostic model development and risk predictions in settings where survival is improving over time. Int J Epidemiol. 2020;49:1316–25. https://doi.org/10.1093/ije/dyaa030.
15. Jenkins DA, Martin GP, Sperrin M, Riley RD, Debray TPA, Collins GS, et al. Continual updating and monitoring of clinical prediction models: time for dynamic prediction systems? Diagn Prognostic Res. 2021;5:1–7. https://doi.org/10.1186/s41512-020-00090-3.
16. Cucchetti A, Giannini EG, Mosconi C, Torres MCP, Pieri G, Farinati F, et al. Recalibrating survival prediction among patients receiving trans-arterial chemoembolization for hepatocellular carcinoma. Liver Cancer Int. 2021;2:45–53. https://doi.org/10.1002/lci2.33.
17. Clift AK, Coupland CAC, Keogh RH, Diaz-Ordaz K, Williamson E, Harrison EM, et al. Living risk prediction algorithm (QCOVID) for risk of hospital admission and mortality from coronavirus 19 in adults: national derivation and validation cohort study. BMJ. 2020;371. https://doi.org/10.1136/bmj.m3731.
18. Hippisley-Cox J, Coupland CAC, Mehta N, Keogh RH, Diaz-Ordaz K, Khunti K, et al. Risk prediction of covid-19 related death and hospital admission in adults after covid-19 vaccination: National prospective cohort study. BMJ. 2021;374. https://doi.org/10.1136/bmj.n2244.
19. QResearch. 2022. https://www.qresearch.org/. Accessed 16 June 2022
20. Wynants L, Van Calster B, Collins GS, Riley RD, Heinze G, Schuit E, et al. Prediction models for diagnosis and prognosis of covid-19: Systematic review and critical appraisal. BMJ. 2020;369. https://doi.org/10.1136/bmj.m1328.
21. McLernon DJ, Giardiello D, Van Calster B, Wynants L, van Geloven N, van Smeden M, et al. Assessing Performance and Clinical Usefulness in Prediction Models With Survival Outcomes: Practical Guidance for Cox Proportional Hazards Models. Ann Intern Med. 2023;176:105–14. https://doi.org/10.7326/M22-0844.
22. Gerds TA, Kattan MW, Schumacher M, Yu C. Estimating a time-dependent concordance index for survival prediction models with covariate dependent censoring. Stat Med. 2013;32:2173–84. https://doi.org/10.1002/sim.5681.
23. Royston P, Altman DG. External validation of a Cox prognostic model: principles and methods. BMC Med Res Methodol. 2013;13. https://doi.org/10.1186/1471-2288-13-33.
24. Crowson CS, Atkinson EJ, Therneau TM. Assessing Calibration of Prognostic Risk Scores. Stat Methods Med Res. 2016;25:1692–706. https://doi.org/10.1177/0962280213497434.
25. Van Calster B, McLernon DJ, van Smeden MV Wynants L, Steyerberg EW, Bossuyt P, et al. Calibration: The Achilles heel of predictive analytics. BMC Med. 2019;17:1–7. https://doi.org/10.1186/s12916-019-1466-7.
26. Graf E, Schmoor C, Sauerbrei W, Schumacher M. Assessment and Comparison of Prognostic Classification Schemes for Survival Data. Stat Med. 1999;9(18):2529–45. https://doi.org/10.1002/(sici)1097-0258(19990915/30)18:17/18<2529::aid-sim274>3.0.co;2-5.
27. Cox DR. Regression Models and Life-Tables. J R Stat Soc Ser B Methodol. 1972;34:187–220. https://doi.org/10.1007/978-1-4612-4380-9_37.
28. Breslow NE. Contribution to the discussion of paper by D.R. Cox. J R Stat Soc Ser B Methodol. 1972;34:216–7.
29. Gama J, Zliobaite I, Bifet A, Pechenizkiy M, Bouchachia A. A survey on concept drift adaptation. ACM Comput Surv. 2014;46. https://doi.org/10.1145/2523813.
30. McCormick TH, Raftery AE, Madigan D, Burd RS. Dynamic Logistic Regression and Dynamic Model Averaging for Binary Classification. Biometrics. 2012;68:1–19. https://doi.org/10.1111/j.1541-0420.2011.01645.x.Dynamic.
31. Riley RD, Snell KIE, Ensor J, Burke DL, Harrell FE, Moons KGM, et al. Minimum sample size for developing a multivariable prediction model: PART II - binary and time-to-event outcomes. Stat Med. 2019;38:1276–96. https://doi.org/10.1002/sim.7992.
32. R Core Team. R: A language and environment for statistical computing. R Foundation for Statistical Computing; 2020. https://www.r-project.org/. Accessed 1 Aug 2022.
33. Brilleman SL, Wolfe R, Moreno-Betancur M, Crowther MJ. Simulating survival data using the simsurv R package. J Stat Softw. 2021;97:1–27. https://doi.org/10.18637/jss.v097.i03.

34. Therneau T, Grambsch P. Modeling Survival Data: Extending the Cox Model. Springer-Verlag; 2000.
35. Mogensen UB, Ishwaran H, Gerds TA. Evaluating Random Forests for Survival Analysis Using Prediction Error Curves. J Stat Softw. 2012;50. https://doi.org/10.18637/jss.v050.i11.
36. Van Calster B, Nieboer D, Vergouwe Y, De Cock B, Pencina MJ, Steyerberg EW. A calibration hierarchy for risk models was defined: from utopia to empirical data. J Clin Epidemiol. 2016;74:167–76. https://doi.org/10.1016/j.jclinepi.2015.12.005.
37. Brilleman SL, Elci EM, Novik JB, Wolfe R. Bayesian Survival Analysis Using the rstanarm R Package. arXiv:2002.09633. 2020.
38. Hoffman MD, Gelman A. The No-U-Turn Sampler: Adaptively Setting Path Lengths in Hamiltonian Monte Carlo. J Mach Learn Res. 2014;15:1593–623.
39. Gelman A, Rubin DB. Inference from Iterative Simulation Using Multiple Sequences. Stat Sci. 1992;7:457–72.
40. Raftery AE, Karny M, Ettler P. Online Prediction Under Model Uncertainty via Dynamic Model Averaging: Application to a Cold Rolling Mill. Technometrics. 2010;52:52–66. https://doi.org/10.1198/TECH.2009.08104.Online.
41. Fine JP, Gray RJ. A proportional hazards model for the subdistribution of a competing risk. J Am Stat Assoc. 1999;94:496–509. https://doi.org/10.1080/01621459.1999.10474144.
42. UK Health Security Agency. Coronavirus (COVID-19) in the UK. 2022. https://coronavirus.data.gov.uk. Accessed 30 Nov 2022
43. Stan Development Team. RStan: the R interface to Stan. R package version 2.21.2. http://mc-stan.org/. Accessed 18 Oct 2022.
44. Lu J, Liu A, Dong F, Gu F, Gama J, Zhang G. Learning under concept drift: a review. In: IEEE Transactions on Knowledge and Data Engineering, vol. 31, no. 12; 2019. p. 2346–63. https://doi.org/10.1109/TKDE.2018.2876857.

## Publisher's Note

Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.