Statistics in Medicine

WILEY

Statistics in Medicine

RESEARCH ARTICLE OPEN ACCESS

The Impact of Violation of the Proportional Hazards Assumption on the Calibration of the Cox Proportional Hazards Model

Peter C. Austin[1,2,3] | Daniele Giardiello[4]

[1]ICES, Toronto, Ontario, Canada | [2]Institute of Health Policy, Management and Evaluation, University of Toronto, Ontario, Canada | [3]Sunnybrook Research Institute, Toronto, Ontario, Canada | [4]Bicocca Bioinformatics Biostatistics and Bioimaging B4 Center, School of Medicine and Surgery, University of Milano-Bicocca, Monza, Italy

Correspondence: Peter C. Austin (peter.austin@ices.on.ca)

Received: 20 November 2024 | **Revised:** 23 April 2025 | **Accepted:** 27 May 2025

Funding: This work was supported by the Canadian Institutes of Health Research (PJT 166161).

Keywords: calibration | Cox regression | model validation | Monte Carlo simulations | proportional hazards

ABSTRACT

The Cox proportional hazards regression model is frequently used to develop clinical prediction models for time-to-event outcomes, allowing clinicians to estimate an individual's risk of experiencing the outcome within specified time horizons (e.g., estimate an individual's 10-year risk of death). The Cox regression model models the association between covariates and the hazard of the outcome. A key assumption of the Cox model is the proportional hazards assumption: the ratio of the hazard function for any two individuals is constant over time, and the ratio is a function of only their covariates and the regression coefficients. Calibration is an important aspect of the validation of clinical prediction models. Calibration refers to the concordance between predicted and observed risk. The impact of the violation of the proportional hazards assumption on the calibration of clinical prediction models developed using the Cox model has not been examined. We conducted a set of Monte Carlo simulations to assess the impact of the magnitude of the violation of the proportional hazards assumption on the calibration of the Cox model. We compared the calibration of predictions obtained using a Cox regression model that ignored the violation of the proportional hazards assumption with those obtained using accelerated failure time (AFT) models, Royston and Parmar's spline-based parametric survival models, and generalized linear models using pseudo-observations. We found that violation of the proportional hazards assumption had negligible impact on the calibration of predictions obtained using a Cox model.

1 | Introduction

Clinical prediction models are increasingly being developed and used to estimate an individual's risk of having a disease or experiencing an outcome. Time-to-event or survival outcomes are common in clinical and epidemiological research. Regression models from survival analysis allow investigators to estimate the absolute risk of an event within specified prediction horizons (e.g., 10-year risk of death). The most popular regression model for the analysis of time-to-event outcomes in settings with censoring appears to be the Cox proportional hazards regression model [1]. The Cox regression model models the association between covariates and the hazard of the outcome. A key assumption of the model is the proportional hazards assumption: the ratio of the hazard function for any two individuals is constant over time, and this ratio is a function of only their covariates and the regression coefficients.

Violation of the proportional hazards assumption can have effects on inferences about the estimated model coefficients. The estimated hazard ratio can be biased [2]. This is because the estimated hazard ratio will be a weighted average of the time-varying hazard ratios [3]. Furthermore, the magnitude of the estimated hazard ratio will depend on the length of follow-up [4]. Thus, the estimated hazard ratios could differ across studies if the duration of follow-up varied across studies. Another consequence of the violation of the proportional hazards assumption is that, in multivariable models, the estimated standard errors can be incorrect [3]. A consequence of the bias in estimated standard errors is that estimated confidence intervals may be too wide or too narrow and that significance levels of the associated hypothesis tests may be incorrect. Estimates of the risk of the occurrence of the outcome of interest within a specified duration of time (i.e., of the complement of the estimated survival function) are, in part, a function of the estimated regression coefficients. Thus, if the estimated regression coefficients are biased in settings in which the proportional hazards assumption is violated, it would be natural to anticipate that estimates of risk are also subject to bias.

Calibration is an important aspect of assessing the accuracy of a clinical prediction model. Calibration refers to the concordance between predicted and observed risk. The impact of the violation of the proportional hazards assumption on calibration has received little attention. The objective of the current paper was to examine the impact of the violation of the proportional hazards assumption on the calibration of prediction models developed using the Cox proportional hazards regression model. The paper is structured as follows: In Section 2, we describe a series of Monte Carlo simulations that were designed to address this question. In Section 3, we report the results of these simulations. Finally, in Section 4, we summarize our findings and place them in the context of the existing literature.

2 | Monte Carlo Simulation Methods

We conducted a set of Monte Carlo simulations to examine the impact of violation of the proportional hazards assumption on the calibration of predictions obtained from a Cox proportional hazards model. We compared the calibration of the Cox model under violation of the proportional hazards assumption with alternative methods for estimating risk with time-to-event outcomes: a Cox model that stratified on the variable for which the proportional hazard assumption was violated, parametric accelerated failure time (AFT) survival models, Royston and Parmar's spline-based parametric survival models, and generalized linear models based on pseudo-observations.

2.1 | Factors in the Monte Carlo Simulations

We allowed two factors to vary in the Monte Carlo simulations: (i) the magnitude of the violation of the proportional hazards assumption; (ii) whether the variable for which the proportional hazards assumption was violated was binary or continuous. The first factor took on nine values while the second factor took on two values. In each of the 18 scenarios, we simulated 1000 datasets.

2.2 | Empirical Analyses to Inform the Data-Generating Process

We used data on 19 559 patients who were hospitalized with an acute myocardial infarction (AMI or heart attack) in 2016 in Ontario, Canada. These data were obtained from the Ontario Myocardial Infarction Database (OMID) [5]. We extracted the age and sex of each individual. We standardized age so that it had a mean zero and a standard deviation of one. We followed each individual from the time of hospital admission until the time of death (including out-of-hospital deaths), censoring patients after 5 years if they were still alive after 5 years.

We used a Weibull parametric survival model to regress time to death on age and sex (observed survival times of zero were changed to 0.5 to avoid the removal of individuals who died on the day of hospital admission). The estimated shape and scale parameters of the underlying Weibull distribution were 0.4836 and 0.0073, respectively. We then fit a Cox proportional hazards model in which we regressed the hazard of death on age and sex, assuming proportional hazards for both variables. The regression coefficients for age and female sex were 1.084 and −0.018, respectively. Thus, a one standard deviation increase in age was associated with a 2.96-fold increase in the hazard of death. Similarly, female sex was associated with a 2% decrease in the hazard of death compared to males.

These analyses were conducted using the R statistical programming language (version 3.6.3). The Weibull parametric survival model was fit using the weibreg function from the eha package (version 2.10.1). The Cox regression model was fit using the coxph function from the survival package (version 3.2-11).

2.3 | Data-Generating Process

We describe the data-generating process for the scenarios in which the proportional hazards assumption was violated for a continuous variable. In each simulation iteration, we simulated a sample of size 2000. For each individual, we generated two baseline covariates: a continuous baseline variable (X_1) and a binary baseline variable (X_2). We assumed that the former followed a standard normal distribution (reflecting the fact that in the empirical analyses described in the previous section, we had standardized age so that it had mean zero and standard deviation one). We simulated these two variables so that there they were correlated. We generated two continuous variables from a multivariate normal distribution with a mean vector (0,0) and variance–covariance matrix $\Sigma = \begin{pmatrix} 1 & 0.2 \\ 0.2 & 1 \end{pmatrix}$. We retained the first variable as continuous. We then dichotomized the second variable such that it was set equal to 1 if it was less than zero and set equal to 0 otherwise.

We assumed that the underlying model relating the hazard of the outcome to the two baseline covariates was of the form: $\log(h(t)) = \log\big(h_0(t)\big) + \big(\beta_0 + \beta_1 t\big)X_1 + \beta_2 X_2$, where $h(t)$ denotes the hazard function and $h_0(t)$ denotes the baseline hazard function. Thus, the regression coefficient for X_1 is a linear function of time and the proportional hazards assumption was violated for this coefficient. In simulating event times, we assumed that

the maximum observed event time was 1825 days (i.e., 5 years), to reflect what was observed in the empirical analyses described in the previous section. We allowed the value of β_1 to vary from 0 to 0.0008 in increments of 0.0001. For a given value of β_1, we determined the value of β_0 so that the average regression coefficient across the 1825 days would be equal to 0.4836 (i.e., it was equal to the time-invariant regression coefficient for age estimated in the previous section). We thus examined nine different scenarios, with the first scenario ($\beta_1 = 0$) denoting a scenario in which the proportional hazards assumption was satisfied. Increasing values of β_1 denote stronger violations of the proportional hazards assumption. The regression coefficient for the binary variable (β_2) was set to equal the regression coefficient for female sex, which was estimated in the empirical analyses described above. Time-to-event outcomes were simulated from a Weibull model using methods described by Crowther and Lambert [6]. This procedure was repeated 1000 times so that for each simulation scenario, we created 1000 simulated datasets.

The time-varying hazard ratios under different magnitudes of violation of the proportional hazards assumption are described in Figure 1. The left panel describes the time-varying hazard ratios for the binary variable when the proportional hazards assumption is violated for the binary variable. The right panel describes the corresponding information for the continuous variable. On each panel, we have superimposed a horizontal line denoting the average hazard ratio over the 1825 days of follow-up. The figure illustrates that we considered scenarios in which the magnitude of violation of the proportional hazards assumption was small and scenarios in which it was large.

2.4 | Analyses in the Simulated Datasets

Each simulated dataset was split into a derivation sample and a validation sample, each of size 1000. In each simulated derivation sample, we regressed the hazard of the outcome on the two baseline covariates using a Cox regression model in which we assumed that the proportional hazards assumption was satisfied (i.e., we fit a misspecified model of the form: $\log(h(t)) = \log(h_0(t)) + \beta_1 X_1 + \beta_2 X_2$). The fitted model was then applied to the validation sample, and a predicted probability of the outcome at 1, 2, 3, 4, and 5 years (i.e., at 365, 730, 1095, 1460, and 1825 days) was obtained for each individual in the validation sample.

Calibration of the model in the validation sample was assessed at each of the five time points (1–5 years) using four quantitative metrics: the integrated calibration index (ICI), E50, E90, and the ratio of observed-to-predicted risk. The ICI, E50, and E90 are the mean, median, and 90th percentile, respectively, of the absolute difference between predicted survival probabilities and smoothed survival frequencies [7]. Smoothed survival frequencies were obtained using Kooperberg's flexible adaptive hazard regression model with the complementary log–log transformation of the predicted probabilities as the sole predictor variable [8]. The ratio of observed-to-predicted risk was computed by

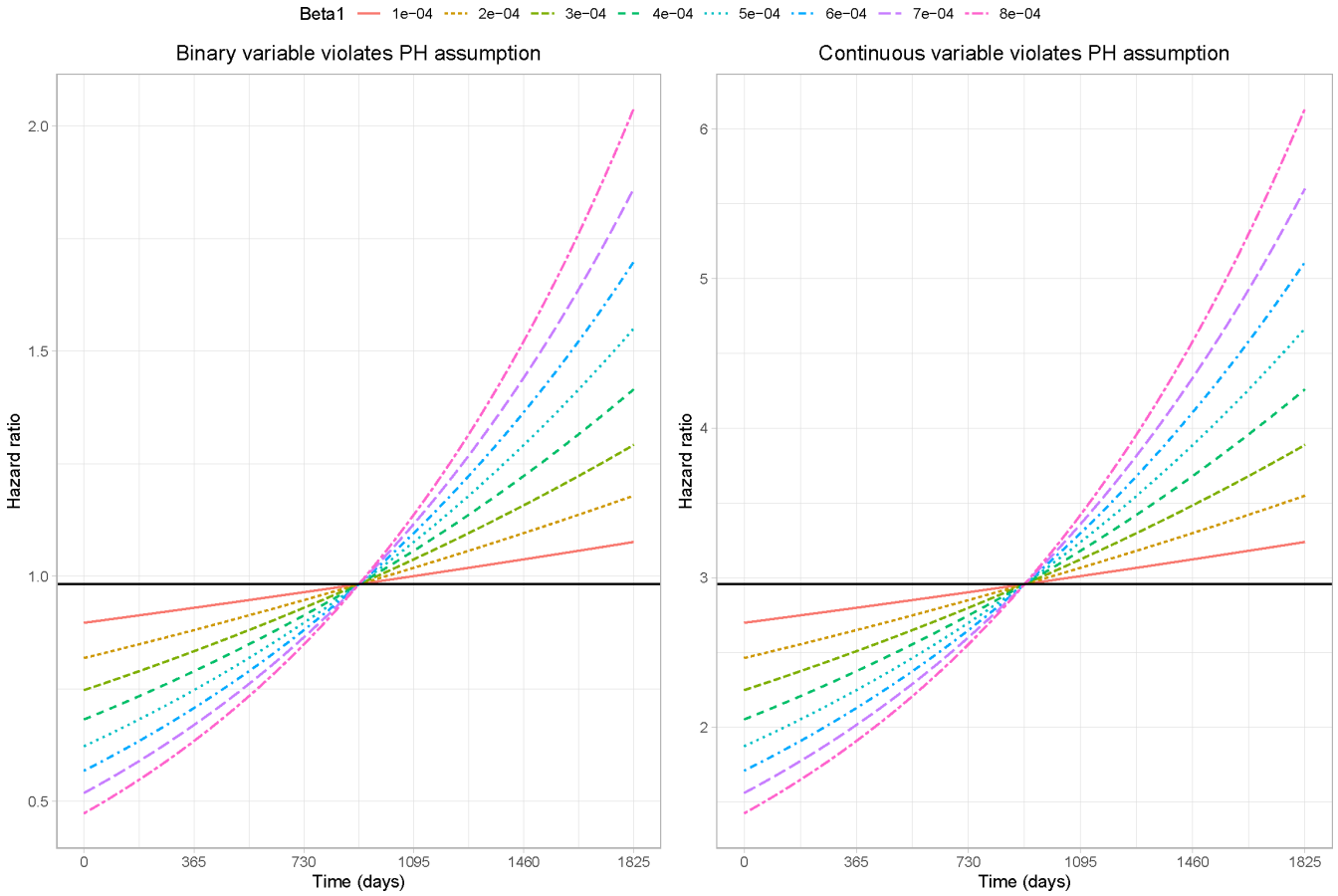

FIGURE 1 | Time-varying hazard ratio under different scenarios.

dividing the observed risk of the outcome at the specified time by the mean predicted risk of the outcome derived from the fitted Cox regression model. The observed risk of the outcome was determined using the Kaplan–Meier estimate of risk at the given duration of time. The above four methods for assessing calibration are quantitative metrics. We also computed graphical calibration curves to allow for a graphical assessment of calibration [7]. Thus, in each simulated validation sample, we computed the ICI, E50, E90, the ratio of observed-to-predicted risk, and a graphical calibration curve. The mean of each of these was then computed across the 1000 simulated samples for each simulation scenario.

We also examined the calibration of three other methods for obtaining estimates of risk at specified durations of time: an AFT parametric survival model, Royston and Parmar's spline-based parametric survival models, and generalized linear models that used pseudo-observations. For the AFT parametric survival model, we fit a generalized gamma model in the derivation sample. We used the generalized gamma model because it is not a proportional hazards model. We then applied the fitted AFT model to the validation sample and computed the probability of the outcome occurring within 1, 2, 3, 4, and 5 years for each individual. Royston and Parmar's spline-based parametric survival model uses natural cubic splines to model the baseline log-cumulative hazard function [9]. We used splines with four degrees of freedom (this decision was made through analyses conducted in one simulated dataset in which we fit models with one through four degrees of freedom and selected the model with the lowest AIC value). For the pseudo-observation-based method, we computed pseudo-observations for each individual in the derivation and validation samples [10]. In the derivation sample, we regressed the pseudo-observations at 1 year on the two baseline covariates using a generalized linear model with a normal distribution and a bounded logit link function (we then repeated this method using a bounded complementary log–log link function) [11]. We then applied the fitted linear model to the validation sample to obtain the probability of the occurrence of the outcome within 1 year. This was repeated for Years 2–5.

In those settings in which the proportional hazards assumption was violated for the binary variable, we considered an additional method in which we fit a Cox regression that stratified on the binary variable, thereby allowing a separate baseline hazard function for each of the two levels of the binary variable. This stratified model had a single predictor variable: the continuous baseline variable. The stratum-specific baseline hazard function was used to estimate the risk of the outcome at each of the prediction horizons.

2.5 | Software

The simulations were conducted using the R statistical programming language (version 3.6.3) [12]. Time-to-event outcomes were simulated using the simsurv function from the simsurv package (version 1.0.0). The Cox model was fit using the coxph function from the survival package (version 3.2-11). Observed risk for the observed-predicted ratio was computed using the survfit function from the survival package. The predicted risk from the fitted Cox model was estimated using the predictSurvProb function from the pec package (version 2019.11.03). Graphical calibration curves were computed using the hare function from the polspline package (version 1.1.19), while ICI, E50, and E90 were computed from smoothed survival frequencies generated using the hare function. The generalized gamma AFT model was fit using the flexsurvreg function from the flexsurv package (version 2.3). Royston and Parmar's spline-based parametric survival model was fit using the stpm2 function from the rstpm2 package (version 1.5.2). Pseudo-observations were computed using the pseudosurv function from the pseudo package (version 1.4.3). The generalized linear model with the pseudo-observations was fit using the glm function using the bounded logit link function blogit from the survival package. The bounded complementary log–log link function was implemented using the cloglog function from the survival package. Simulation results were summarized using the simsum function from the rsimsurv package (version 0.13.0).

3 | Monte Carlo Simulation Results

We report our results separately for the scenarios when the variable for which the proportional hazards assumption was violated was binary and when it was continuous.

3.1 | Proportional Hazards Assumption Violated for a Binary Variable

Results for the time prediction horizons of 1 and 5 years are reported in Figures 2 and 3, respectively, while those for Years 2–4 are reported in Figures A1–A3, respectively, in the online supplemental material. Each of these figures consists of four panels, one for each of the quantitative calibration metrics (ICI, E50, E90, and the observed-predicted ratio). For a given metric, the scale of the vertical axis is the same across the five figures, allowing results to be compared across prediction horizons.

Across all five prediction horizons, the use of pseudo-observations with the bounded complementary log–log link function resulted in predictions that displayed the worst calibration across all four calibration metrics. Among the remaining methods, the use of pseudo-observations with the bounded logit link function tended to have worse performance than the other methods across the five prediction horizons and the four calibration metrics. The performance of the AFT (generalized gamma) model varied across the five prediction horizons. At shorter prediction horizons (e.g., 1 and 2 years), the AFT model systematically underestimated risk when the proportional hazards assumption was not violated or when the magnitude of its violation was low, while it systematically overestimated risk when the magnitude of the violation of the proportional hazards assumption was high. However, the magnitude of underestimation and overestimation diminished as the length of the prediction horizon increased, such that with the 5-year prediction horizon, the observed-to-predicted ratio was very close to one, regardless of the magnitude of the violation of the proportional hazards assumption. The misspecified Cox model that ignored the violation of the proportional hazards assumption, the Cox model that stratified on the binary variable for which the proportional hazards assumption was violated, and Royston

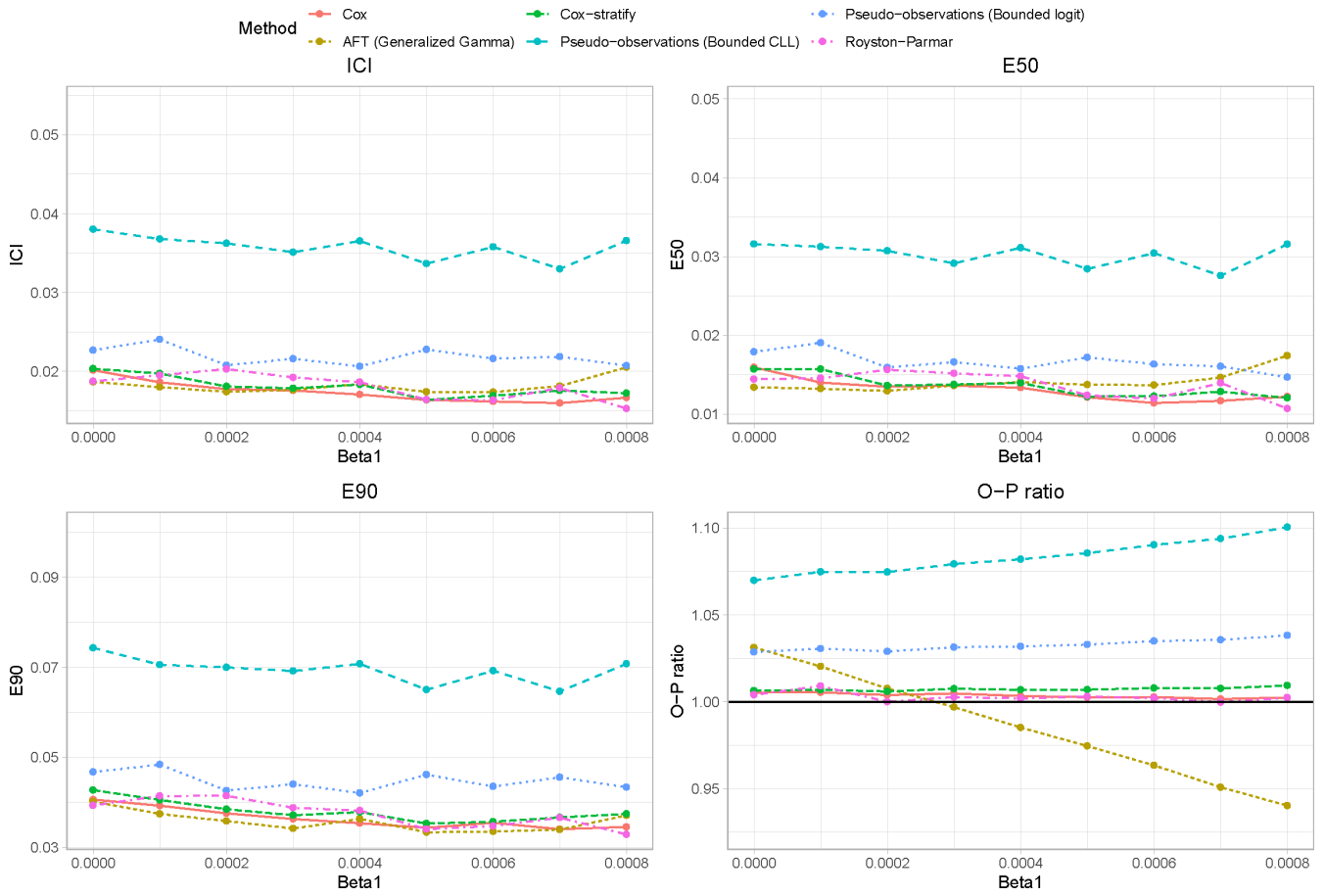

FIGURE 2 | Relationship between β_1 and calibration at 1 year (binary covariate).

and Parmar's spline-based parametric survival model had comparable performance and displayed the least lack of calibration of all the prediction methods. Furthermore, the magnitude of the violation of the proportional hazards assumption tended to have no meaningful impact on the magnitude of miscalibration for these three methods. The Monte Carlo standard errors of the mean estimated ICI are described in Figure A4.

Graphical calibration curves are described in Figures 4 and 5 for the 1-year prediction horizon and 5-year prediction horizon, respectively, and in Figures A5–A7 for the 2-year prediction horizon, the 3-year prediction horizon, and the 4-year prediction horizon, respectively. Each figure consists of nine panels, one for each of the values of β_1 (with $\beta_1 = 0$ denoting that the proportional hazards assumption was valid). We have superimposed a diagonal line on each panel, denoting the line of perfect calibration. Deviation from this line denotes a lack of calibration. Across all five prediction horizons, the use of pseudo-observations with a generalized linear model with a bounded complementary log–log link function displayed a minor lack of calibration, particularly when the predicted risk of the outcome was high. The other five methods tended to display very good calibration across all five prediction horizons. The calibration curve for the conventional Cox model that ignored the violation of the proportional hazards assumption and the calibration curve for Royston and Parmar's spline-based parametric survival model were essentially indistinguishable from one another.

3.2 | Proportional Hazards Assumption Violated for a Continuous Variable

Results for the time prediction horizons of 1 and 5 years are reported in Figures 6 and 7, respectively, while those for Years 2–4 are reported in Figures A8–A10, respectively.

When assessing calibration using ICI, E50, and E90, the primary observation was that, when using the conventional misspecified Cox regression model, increasing magnitude of the violation of the proportional hazards assumption did not result in a meaningful increase in the lack of calibration for predictions at longer prediction horizons (i.e., at 3, 4, and 5 years). However, at shorter prediction horizons (i.e., 1 and 2 years), the magnitude of miscalibration increased as the magnitude of the violation of the proportional hazards assumption increased. While the use of pseudo-observations with a generalized linear model with a bounded logit link function did not have the best calibration across all scenarios or across all metrics, it often resulted in estimates of risk that displayed good calibration and were not meaningfully affected by the magnitude of the violation of the proportional hazards assumption. At shorter prediction time horizons (i.e., 1 and 2 years), the use of pseudo-observations with the bounded complementary log–log link function tended to result in estimates of risk that displayed the greatest lack of calibration when the magnitude of the violation of the proportional

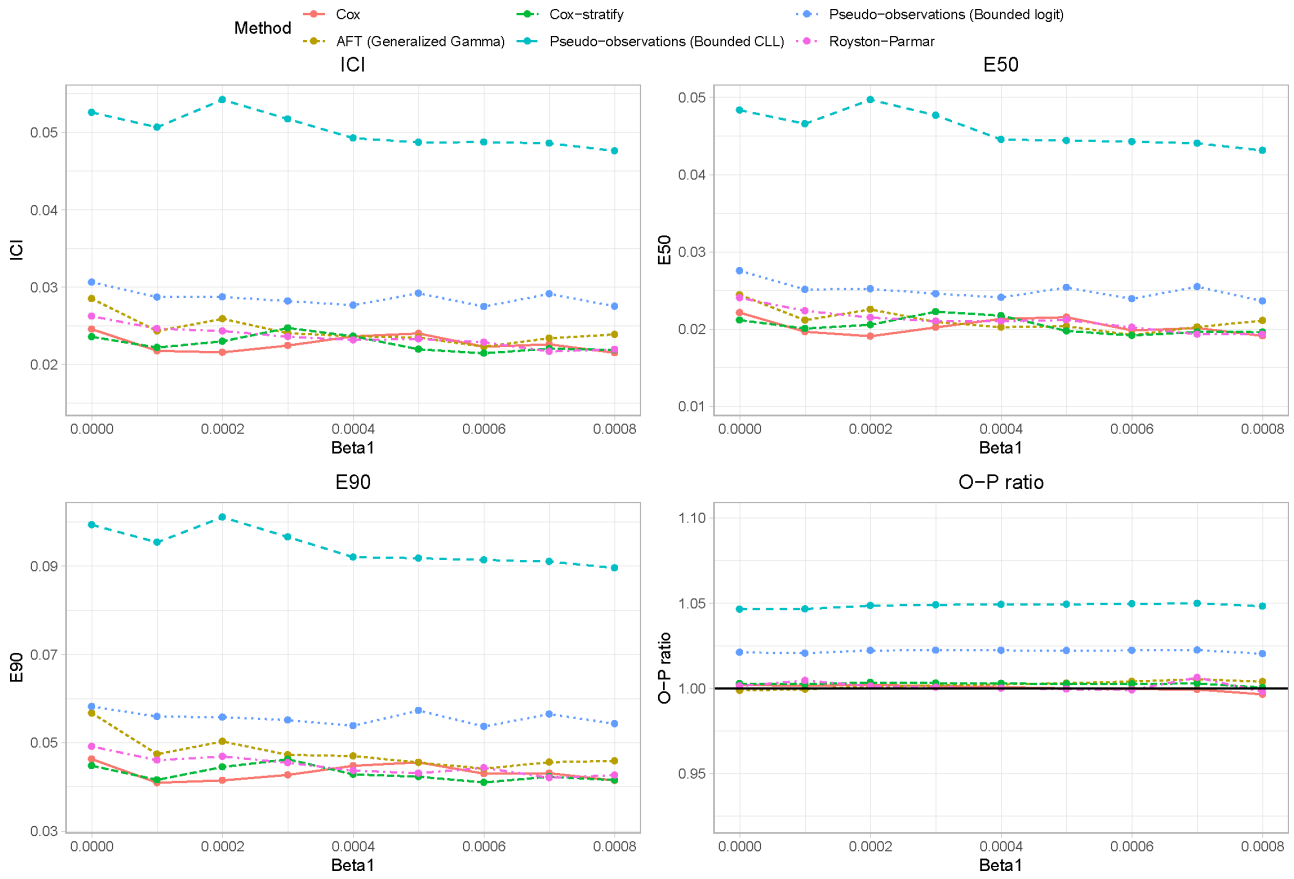

FIGURE 3 | Relationship between β_1 and calibration at 5 years (binary covariate).

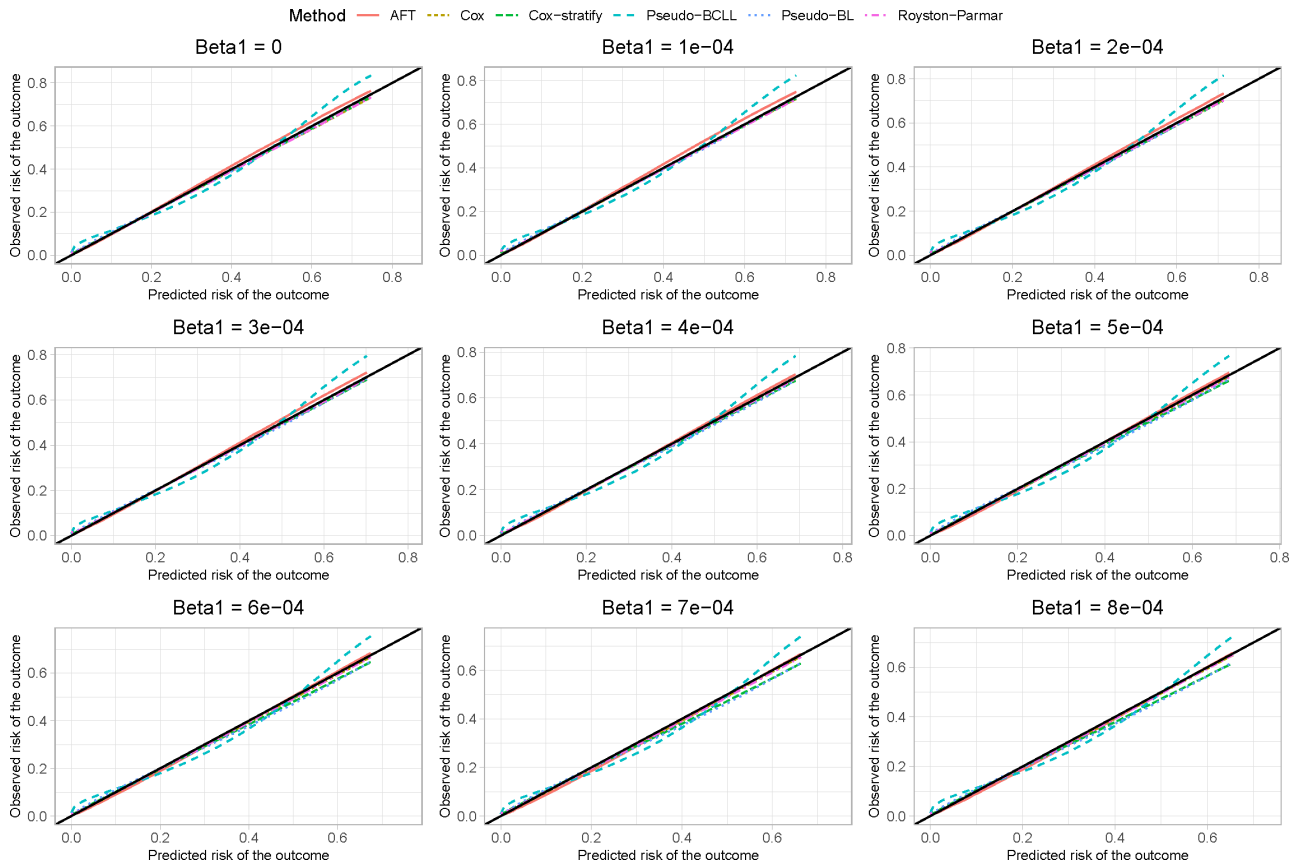

FIGURE 4 | Calibration curves: Predictions at 1 year (binary covariate).

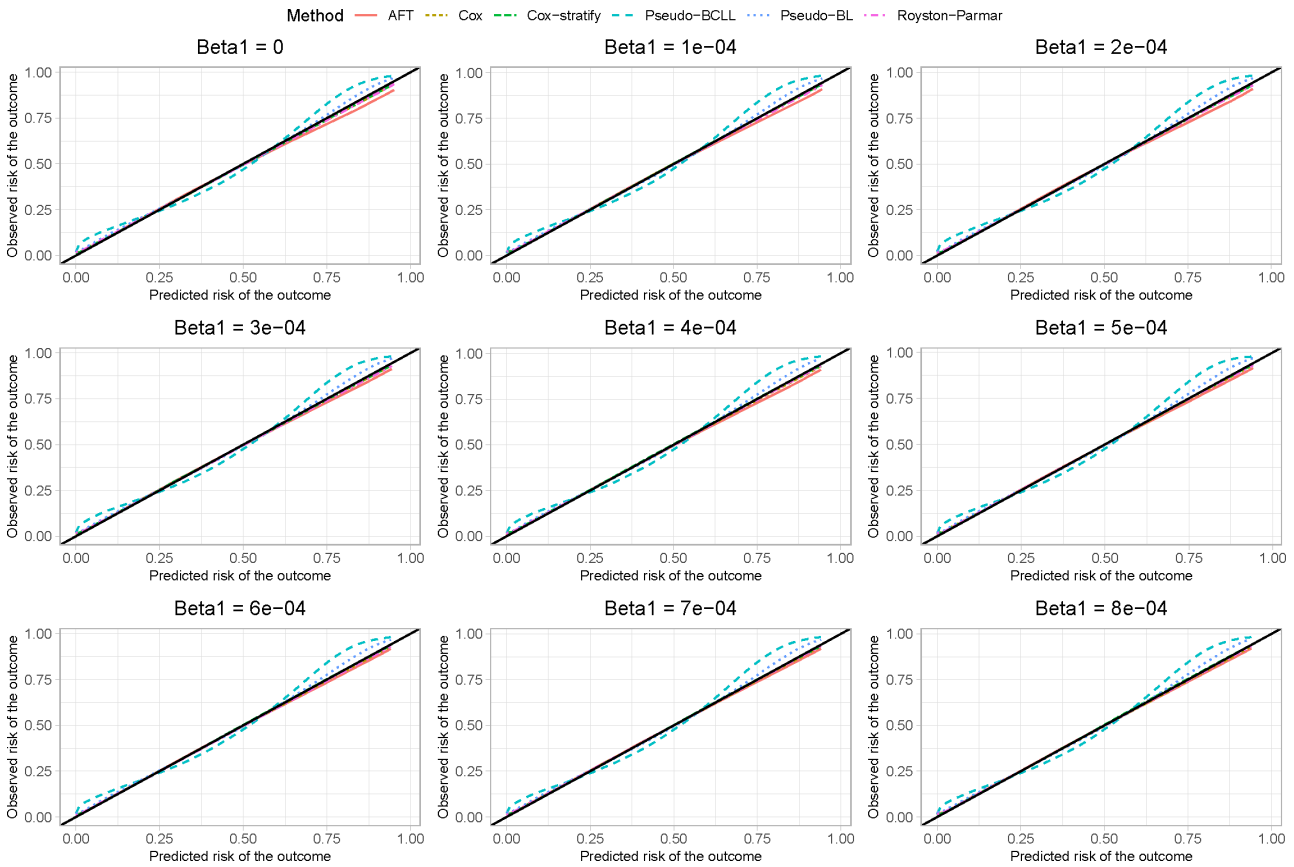

FIGURE 5 | Calibration curves: Predictions at 5 years (binary covariate).

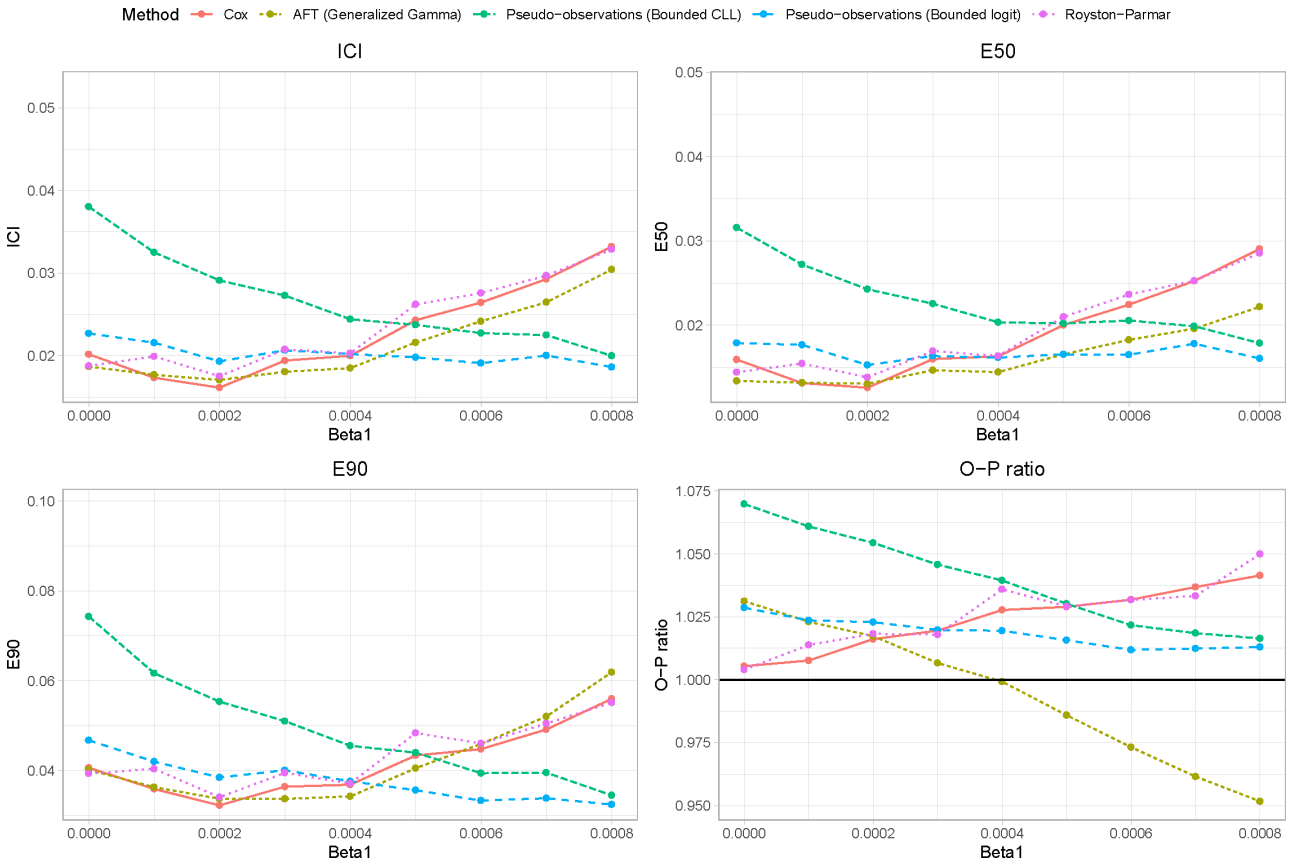

FIGURE 6 | Relationship between β_1 and calibration at 1 year (continuous covariate).

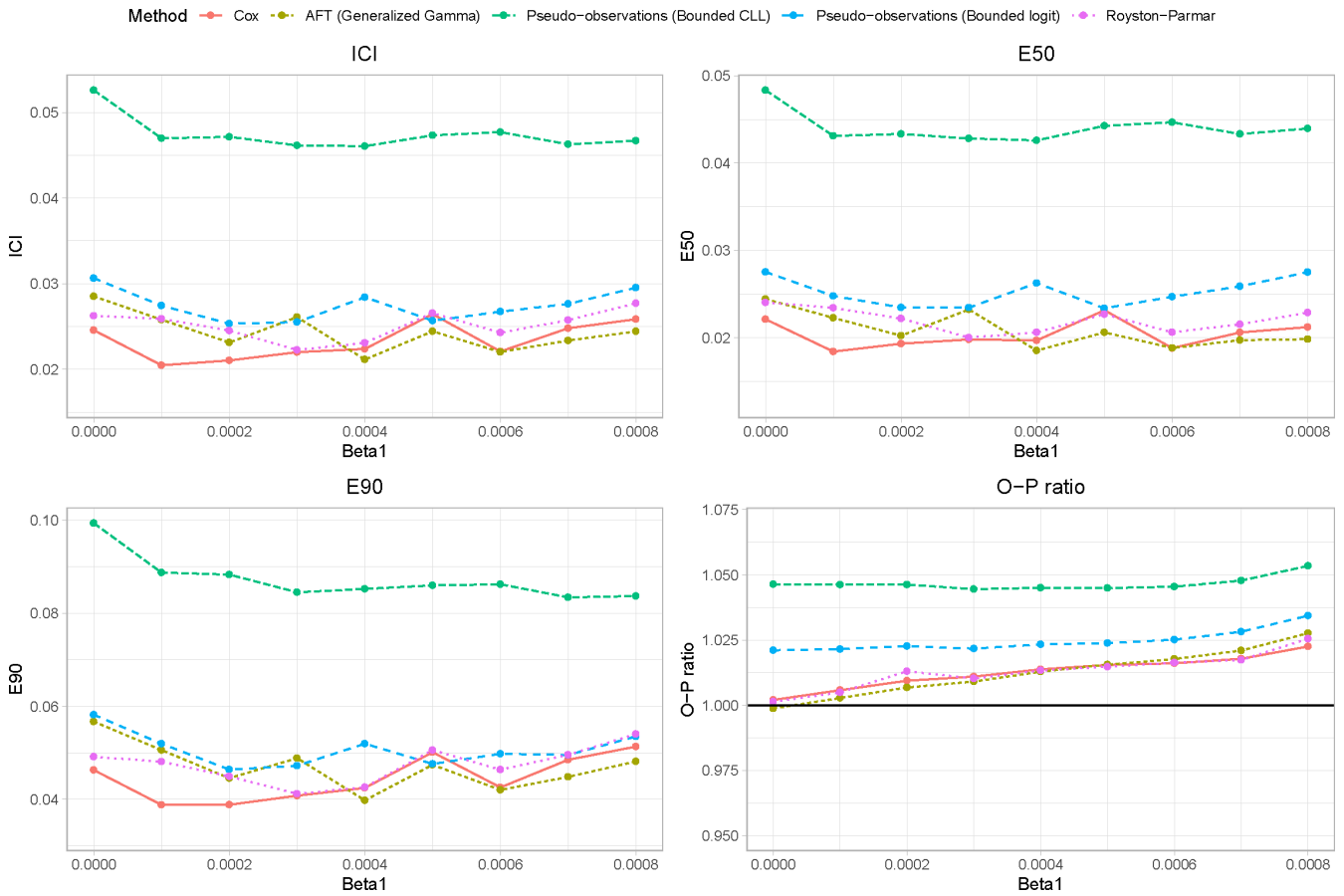

FIGURE 7 | Relationship between β_1 and calibration at 5 years (continuous covariate).

hazards assumption was weak to moderate. At longer prediction time horizons, this method tended to result in estimates with the greatest lack of calibration across all magnitudes of violation of the proportional hazards assumption. The Monte Carlo standard errors of the mean estimated ICI are described in Figure A11.

When assessing calibration using the observed-to-predicted ratio, the primary observation was that the use of the conventional Cox model that ignored the violation of the proportional hazards assumption displayed an increasing lack of calibration as the magnitude of the violation of the proportional hazards assumption increased. In particular, the magnitude of the underestimation of risk increased as the magnitude of the violation of the proportional hazards assumption increased. In contrast to this, the use of pseudo-observations with a generalized linear model with a bounded logit link function displayed a modest lack of calibration, but the magnitude of underestimation of risk was not affected by the magnitude of the violation of the proportional hazards assumption.

Graphical calibration curves are described in Figures 8 and 9 for the 1-year prediction horizon and 5-year prediction horizon, respectively, and in Figures A12–A14 for the 2- to 4-year prediction horizons, respectively. The primary observation was that at shorter prediction time horizons (i.e., 1 and 2 years), the use of the conventional Cox model, which ignored the violation of the proportional hazards assumption, tended to result in predictions that displayed very good calibration when the magnitude of the violation of the proportional hazards assumption was weak to modest. When the violation of the proportional hazards assumption was moderate to strong, this method resulted in a lack of calibration, particularly in those subjects for whom the predicted risk was high. When the prediction time horizon was longer (i.e., 3–5 years), this method tended to result in risk estimates that displayed very good calibration. Both the AFT (generalized gamma) model and the Royston–Parmar spline-based parametric survival model tended to result in risk estimates that displayed a calibration very similar to that of the conventional Cox regression model. Indeed, the calibration curve for the conventional Cox model that ignored the violation of the proportional hazards assumption and the calibration curve for Royston and Parmar's spline-based parametric survival model were essentially indistinguishable from one another. Consistent with the settings in which the proportional hazards assumption was violated for the binary variable, the use of pseudo-observations with the bounded complementary log–log link function tended to result in predictions that displayed a lack of calibration across all five prediction time horizons, particularly in those subjects for whom the predicted risk was high.

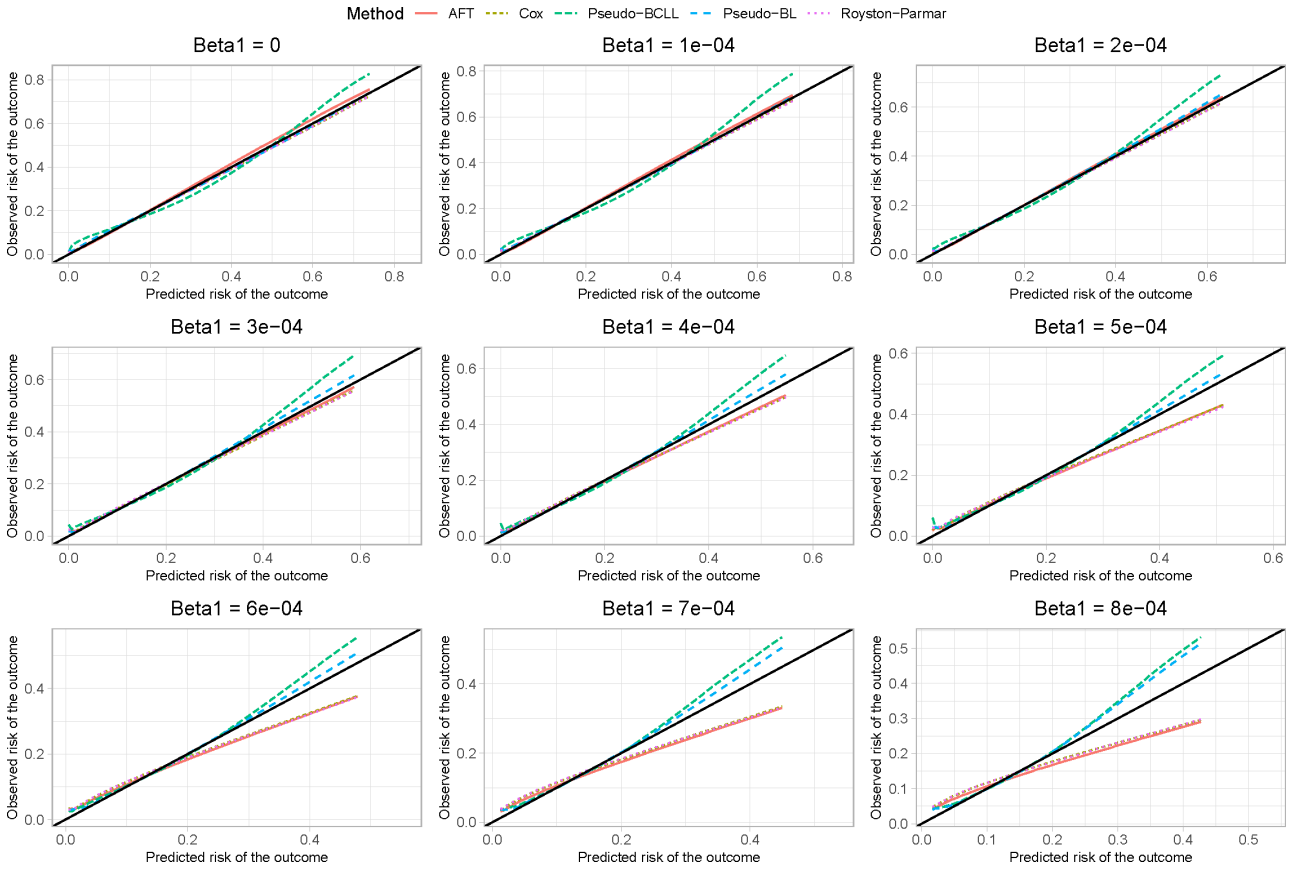

FIGURE 8 | Calibration curves: Predictions at 1 year (continuous covariate).

4 | Discussion

Our primary finding was that when the proportional hazards assumption was violated for a binary variable, the resultant misspecified Cox model still had excellent calibration. In particular, the magnitude of the violation of the proportional hazards assumption had no discernible impact on the magnitude of miscalibration. When the proportional hazards assumption was violated for a continuous variable, at most, minor to modest miscalibration was observed when the predicted risk of the outcome was high. Furthermore, in the settings with a continuous variable for which the proportional hazards assumption was violated, even when miscalibration was observed, it tended to be observed only for some prediction horizons and when the magnitude of the violation of the proportional hazards assumption was moderate to strong.

We also examined the performance of alternative prediction methods in settings in which the proportional hazards assumption was violated. The use of a parametric AFT (generalized gamma) survival model did not result in predictions with improved calibration compared to the Cox model. The likely reason for this is that, while the Cox model was misspecified by failing to account for nonproportional hazards, the AFT model was also misspecified. The generalized gamma AFT is not a proportional hazards model; however, it made an incorrect assumption about the distribution of event times. Event times were generated under a Weibull model, whereas the AFT model we used assumed a generalized gamma model for event times. While pseudo-observations appear to be rarely used in practice when developing clinical prediction models, we found that the use of pseudo-observations with a generalized linear model with a bounded logit link function tended to perform well across a wide variety of settings.

There is a paucity of research on the impact of the violation of the proportional hazards assumption on the accuracy of predictions obtained from a Cox regression model. van Houwelingen suggested that one can obtain survival probabilities from a misspecified Cox model in which one has ignored the violation of the proportional hazards assumption and that this will work well if "$\beta(t)$ does not vary too much over time, the effect of the covariate is not too big and the follow-up is not too long" [13]. van Houwelingen and Putter suggest that this is the case because it is the cumulative hazard function, and not the instantaneous hazard function, that is important for estimating the survival function (or its complement, the cumulative incidence function), and that the effect of a variable on the survival function is through a weighted average of the time-varying regression coefficient over the interval between zero and the time at which the survival function is being evaluated [14]. Royston and Altman, when discussing principles for the external validation of a clinical prediction model developed using Cox regression, suggested that investigators can assess the validity of the proportional hazards assumption in the external validation sample [15]. They proceeded to state that "even if the proportional hazards assumption is untenable, a model may still provide good discrimination . . . , but the calibration would need to be scrutinised. Such a 'partially validated' model may

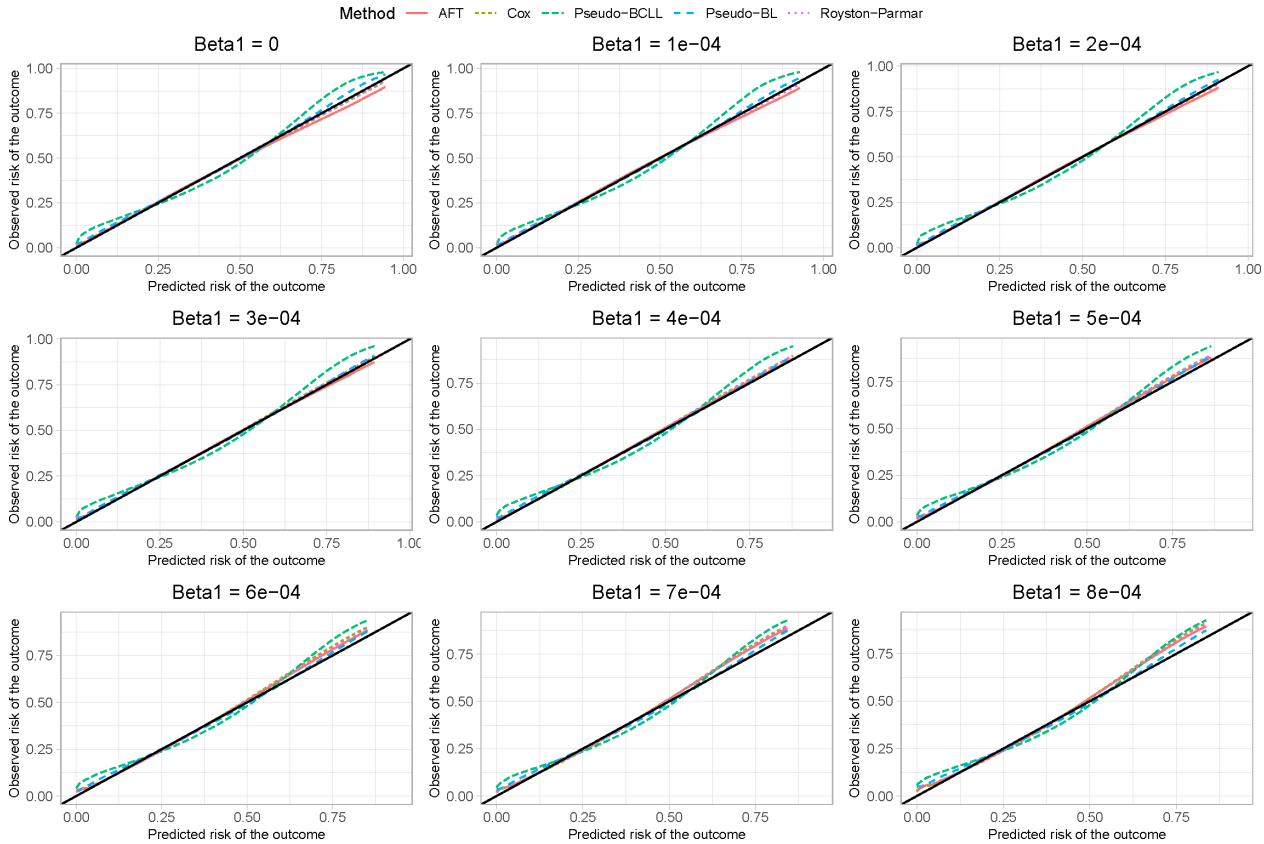

FIGURE 9 | Calibration curves: Predictions at 5 years (continuous covariate).

perform well enough to retain clinical utility." Our findings suggest that the proportional hazards assumption can be violated, and the calibration of the model can still be good. In particular, in many settings, the magnitude of the violation of the proportional hazards assumption can have, at most, a negligible impact on the calibration of the Cox model. van Houwelingen and Putter suggested that an alternative approach in the presence of the violation of the proportional hazards assumption is to use a stopped Cox model [16]. Using this approach, when making predictions at time t_{horizon}, one induces administrative censoring at time t_{horizon} and then uses a conventional Cox model that ignores the violation of the proportional hazards assumption.

A notable omission from the current study is Aalen's additive hazard model [17]. Aalen's additive hazard model is a model in which coefficients have an additive rather than a relative effect on the hazard function. Furthermore, Aalen's additive hazard model does not require the assumption of proportional hazards. For this reason, one might assume that Aalen's additive hazard model would result in risk estimates that display good calibration in settings in which the proportional hazards assumption was violated. In preliminary simulations, we found that Aalen's additive model tended to have worse calibration than all the other methods (data not shown). However, we did not include Aalen's additive model in the study because, strictly speaking, it is a misspecified regression model in this setting, as outcomes were generated under a relative model.

The primary limitation of the current study is its reliance on Monte Carlo simulations. The simulations were computationally intensive, not only because of the statistical models that were fit in the simulated datasets but also because of the complexity of the data-generating process. Using a Cox–Weibull model to simulate time-to-event outcomes required using numerical integration of the cumulative hazard function and root-finding techniques to invert the cumulative hazard function. This was computationally intensive, limiting the number of scenarios that we could examine. However, we examined 18 different scenarios characterized by the type of variable for which the proportional hazards assumption was violated and by the magnitude of the violation of the proportional hazards assumption. Furthermore, parameters used in the data-generating process were obtained from empirical analyses of individuals hospitalized with AMI. Thus, the simulated datasets reflected those observed in a specific clinical setting. A second limitation is that the use of pseudo-observations requires the assumption of non-informative censoring [11]. Further research is required on the impact of informative censoring on the performance of these methods. A third limitation is that the current study did not include a case study or an empirical analysis but relied solely on Monte Carlo simulations. While we could have used the different methods in an empirical analysis of data in which it was known that the proportional hazards assumption was violated, this would only have allowed us to compare the calibration of the different methods. It would not have allowed us to compare the observed calibration with the calibration that would have been observed had the proportional hazards assumption not been violated. This latter comparison was the focus of the study: the impact of the violation (and the magnitude of the violation) of the proportional hazards assumption on the calibration of predictions obtained from the Cox model.

Such a question cannot be addressed using empirical analyses. A final limitation was that in our simulations, we simulated outcomes using a Weibull parametric survival model. It is possible that different results would be observed when different models were used to generate outcomes.

In summary, the magnitude of the violation of the proportional hazards assumption had, at most, a minor impact on the calibration of the Cox regression model. In particular, when the proportional hazards assumption was violated for a binary variable, the calibration of the Cox model, even under a very strong violation of the proportional hazards assumption, was comparable to the calibration of the Cox model when the proportional hazards assumption was satisfied. The use of well-known alternative methods, such as AFT parametric survival models, did not result in improved calibration compared to the use of the Cox model.

Acknowledgments

ICES is an independent, nonprofit research institute funded by an annual grant from the Ontario Ministry of Health (MOH) and the Ministry of Long-Term Care (MLTC). As a prescribed entity under Ontario's privacy legislation, ICES is authorized to collect and use healthcare data for the purposes of health system analysis, evaluation, and decision support. Secure access to these data is governed by policies and procedures that are approved by the Information and Privacy Commissioner of Ontario. The use of the data in this project is authorized under section 45 of Ontario's Personal Health Information Protection Act (PHIPA) and does not require review by a Research Ethics Board. This study was supported by ICES, which is funded by an annual grant from the Ontario MOH and the Ministry of Long-Term Care (MLTC). This study also received funding from the Canadian Institutes of Health Research (CIHR) (PJT 166161). This document used data adapted from the Statistics Canada Postal CodeOM Conversion File, which is based on data licensed from Canada Post Corporation, and/or data adapted from the Ontario Ministry of Health Postal Code Conversion File, which contains data copied under license from Canada Post Corporation and Statistics Canada. Parts of this material are based on data and/or information compiled and provided by CIHI and the Ontario MOH. The analyses, conclusions, opinions, and statements expressed herein are solely those of the authors and do not reflect those of the funding or data sources; no endorsement is intended or should be inferred. The dataset from this study is held securely in coded form at ICES. While legal data sharing agreements between ICES and data providers (e.g., healthcare organizations and government) prohibit ICES from making the dataset publicly available, access may be granted to those who meet pre-specified criteria for confidential access, available at www.ices.on.ca/DAS (email: das@ices.on.ca). Daniele Giardiello is funded by the National Plan for NRRP Complementary Investments (PNC, established with the decree-law 6 May 2021, n. 59, converted by law n. 101 of 2021) in the call for the funding of research initiatives for technologies and innovative trajectories in the health and care sectors (Directorial Decree n. 931 of 06-06-2022) — project n. PNC0000003 — AdvaNced Technologies for Human-centrEd Medicine (project acronym: ANTHEM).

Conflicts of Interest

The authors declare no conflicts of interest.

Data Availability Statement

Research data are not shared.

References

1. D. R. Cox, "Regression Models and Life Tables (With Discussion)," *Journal of the Royal Statistical Society, Series B (Statistical Methodology)* 34 (1972): 187–202.

2. J. O'Quigley and R. Xu, "Robustness of Proportional Hazards Regression," in *Handbook of Survival Analysis*, ed. J. P. Klein, H. C. van Houwelingen, J. G. Ibrahim, and T. H. Scheike (CRC Press, 2014), 323–339.

3. M. J. Stensrud and M. A. Hernan, "Why Test for Proportional Hazards?," *Journal of the American Medical Association* 323, no. 14 (2020): 1401–1402.

4. M. A. Hernan, "The Hazards of Hazard Ratios," *Epidemiology* 21, no. 1 (2010): 13–15.

5. J. V. Tu, C. D. Naylor, and P. Austin, "Temporal Changes in the Outcomes of Acute Myocardial Infarction in Ontario, 1992-1996," *CMAJ* 161, no. 10 (1999): 1257–1261.

6. M. J. Crowther and P. C. Lambert, "Simulating Biologically Plausible Complex Survival Data," *Statistics in Medicine* 32, no. 23 (2013): 4118–4134.

7. P. C. Austin, F. E. Harrell, Jr., and D. van Klaveren, "Graphical Calibration Curves and the Integrated Calibration Index (ICI) for Survival Models," *Statistics in Medicine* 39, no. 21 (2020): 2714–2742.

8. C. Kooperberg, C. J. Stone, and Y. K. Truong, "Hazard Regression," *Journal of the American Statistical Association* 90, no. 429 (1995): 78–94.

9. P. Royston and M. K. Parmar, "Flexible Parametric Proportional-Hazards and Proportional-Odds Models for Censored Survival Data, With Application to Prognostic Modelling and Estimation of Treatment Effects," *Statistics in Medicine* 21, no. 15 (2002): 2175–2197.

10. P. K. Andersen and M. P. Perme, "Pseudo-Observations in Survival Analysis," *Statistical Methods in Medical Research* 19, no. 1 (2010): 71–99.

11. T. Therneau, "Pseudo-Values for Survival Data," accessed November 6, 2024, https://cran.r-project.org/web/packages/survivalVignettes/vignettes/pseudo.html.

12. R Core Team, *R: A Language and Environment for Statistical Computing* (R Foundation for Statistical Computing, 2005).

13. H. C. van Houwelingen, "Dynamic Prediction by Landmarking in Event History Analysis," *Scandinavian Journal of Statistics* 34 (2007): 70–85.

14. H. C. van Houwelingen and H. Putter, *Dynamic Prediction in Clinical Survival Analysis* (CRC Press, 2012).

15. P. Royston and D. G. Altman, "External Validation of a Cox Prognostic Model: Principles and Methods," *BMC Medical Research Methodology* 13 (2013): 33.

16. H. C. van Houwelingen and H. Putter, "Comparison of Stopped Cox Regression With Direct Methods Such as Pseudo-Values and Binomial Regression," *Lifetime Data Analysis* 21, no. 2 (2015): 180–196.

17. O. O. Aalen, "A Linear Regression Model for the Analysis of Life Times," *Statistics in Medicine* 8, no. 8 (1989): 907–925.

Supporting Information

Additional supporting information can be found online in the Supporting Information section. **Data S1.** Supporting figures.