Bias-corrected Cox regression with AI-extracted covariates via calibration summary statistics

Arjun Sondhi A2SONDHI@GMAIL.COM
Northwell Health
New York, NY, USA

Abstract

Large-scale observational studies increasingly rely on AI pipelines to extract structured variables from unstructured clinical records. A common workflow separates the data vendor, who validates extraction accuracy with a gold-standard sample, from the downstream researcher, who receives only the extracted dataset and summary accuracy statistics. We develop a bias-correction framework for the Cox proportional hazards model when covariates are subject to AI extraction error. Within a unified multivariate calibration framework, we show that the naive Cox estimator's bias decomposes into a leading-order calibration term and a second-order residual that vanishes as extraction accuracy improves. The leading-order term yields a corrected estimator that operates as a post-hoc matrix multiplication on the output of any standard Cox software. We further derive bias-adjusted confidence intervals that incorporate calibration uncertainty and a sensitivity diagnostic for assessing whether the neglected residual could materially affect inference. Synthetic data experiments with cross-dependent extraction errors and controlled nonlinear calibration violations confirm that the correction substantially reduces bias and achieves near-nominal coverage even under mild violations of the linear calibration assumption. The framework yields a concrete reporting specification: a short list of summary statistics that data vendors should provide alongside any AI-extracted covariate dataset used in survival analysis.

1. Introduction

The growing availability of electronic health records, clinical notes, and registry databases has enabled observational studies at unprecedented scale. Increasingly, AI-based extraction tools—including large language models (LLMs) and other natural language processing (NLP) pipelines—automate the extraction of structured analytic variables from unstructured source data, and a growing number of commercial data vendors offer AI-extracted clinical datasets as research products (Adamson et al., 2023; Agrawal et al., 2023).

A common workflow has emerged: a data vendor develops an AI extraction pipeline, validates it against manually abstracted gold-standard labels, and delivers the extracted dataset to end-user researchers. The researchers receive the extracted variables but *not* the paired validation data. Instead, what the vendors often provide alongside the extracted data is a set of summary statistics characterizing extraction accuracy. These often include sensitivity and specificity for categorical variables, while continuous variable error summaries may include mean absolute error or root mean squared error (Estevez et al., 2026). While these metrics are useful for broadly assessing the accuracy of the AI extraction task, they

do not directly address how reliable downstream statistical analyses are when applied to the AI-extracted dataset.

It is well established that using error-prone variables in statistical models (whether from manual abstraction mistakes, instrument noise, or algorithmic extraction) can produce systematic distortions in parameter estimates. AI-extracted covariates introduce a modern instance of this classical problem: the extraction algorithm acts as a noisy measurement device, and the resulting surrogates inherit error patterns that are shaped by the algorithm's architecture, training data, and the complexity of the source text. We focus on the Cox model because of its central role in clinical research: it is the default analysis model for time-to-event endpoints, and regulatory submissions routinely rely on Cox-based hazard ratio estimates for efficacy evaluation. The central question is:

> *Given accuracy summaries reported by the data vendor, can a downstream researcher correct for the bias in a Cox proportional hazards model fit to AI-extracted covariates and conduct valid inference on the corrected estimates?*

This paper focuses on the setting where covariate extraction is the primary source of measurement error: the event time T and event indicator Δ are assumed to be observed correctly (e.g., because they come from administrative or registry sources that require no language extraction).

Our main contributions are as follows. First, we derive a closed-form bias decomposition (Theorem 1) that expresses the difference $\boldsymbol{\beta} - \boldsymbol{\beta}^*$ between the true Cox coefficients and the probability limit of the naive estimator in terms of a *leading-order calibration correction* and a *second-order regression calibration (RC) residual*. The leading-order term depends only on the multivariate calibration slope matrix $\mathbf{B}$ and the naive estimate $\hat{\boldsymbol{\beta}}^*$, both of which are available to the researcher, yielding a fully computable corrected estimator $\hat{\boldsymbol{\beta}}_{\text{corr}} = (\hat{\mathbf{B}}^\top)^{-1}\hat{\boldsymbol{\beta}}^*$. Second, we develop a framework for bias-adjusted inference (Section 3.3) that propagates both sampling variability in $\hat{\boldsymbol{\beta}}^*$ and estimation uncertainty in $\hat{\mathbf{B}}$ through the correction. Crucially, the correction is *analysis-flexible*: it operates on the output of any standard Cox software, requiring no modifications to the partial likelihood fitting procedure itself, and accommodates arbitrary downstream contrasts (e.g., individual coefficients, linear combinations, or joint tests). Third, the bias decomposition cleanly separates quantities into those reported by the vendor and those computed by the researcher, yielding a concrete *reporting specification* of which accuracy summaries should accompany any AI-extracted covariate dataset used in survival analysis (Table 1).

The rest of this paper is organized as follows. Section 2 reviews the measurement error, post-prediction inference, and real-world data quality assurance literature, motivating the need for our work. Section 3 develops the methodology: Section 3.1 fixes notation and states the working assumptions, Section 3.2 presents the bias decomposition of Theorem 1 and the resulting corrected estimator, and Section 3.3 derives plug-in and propagated-variance confidence intervals along with a sensitivity diagnostic for the neglected residual. Section 4 reports simulation studies under cross-dependent extraction errors (Section 4.1) and under controlled violations of the linear calibration assumption at constant total error (Section 4.2). Section 5 concludes with a discussion of the vendor reporting standard and the framework's limitations. Appendix A contains the proof of Theorem 1, Appendix B

examines how confidence interval coverage converges with the validation sample size, and Appendix C presents a semi-synthetic analysis of MIMIC-IV data.

Generalizable Insights about Machine Learning in the Context of Healthcare

As AI extraction pipelines become standard infrastructure for large-scale clinical research, a recurring structural problem emerges: the team that builds and validates the extraction model is typically not the team that conducts the downstream analysis, and the validation data rarely travel with the extracted dataset.

1. We show that given appropriately chosen summary statistics from the validation stage, a downstream analyst can recover bias-corrected point estimates and valid confidence intervals from standard software output, without access to paired validation data or specialized estimation procedures.

2. The approach establishes a broader design principle: for any downstream analysis where ML-generated variables feed into a formal statistical procedure, one can ask what minimal summary statistics from the ML validation stage are sufficient for inferential correction, and design reporting standards around that answer.

3. We believe this shift from reporting extraction accuracy as an end in itself to reporting quantities that enable valid downstream inference is a principled design framework for the growing ecosystem of AI-extracted clinical research data.

2. Related Work

Measurement error in regression models. The statistical theory of measurement error is mature, with comprehensive treatments in the monographs of Fuller (2009) and Carroll et al. (2006). For linear models, additive covariate error with known variance produces the classical attenuation factor $\sigma_X^2/(\sigma_X^2 + \sigma_U^2)$, and regression calibration provides a consistent correction. For generalized linear models, the induced bias depends on the link function, and regression calibration is only approximately consistent, with the quality of the approximation depending on the curvature of the link and the magnitude of the measurement error (Carroll et al., 2006).

Measurement error in Cox models. Regression calibration for the Cox model was formalized by Wang et al. (1997), who showed that replacing the error-prone covariate with its conditional expectation given the surrogate is only approximately consistent because the induced hazard function does not satisfy proportional hazards; Xie et al. (2001) proposed a risk-set calibration variant that recalibrates at each event time. Simulation extrapolation (SIMEX) was adapted to survival analysis and extended to handle mixtures of Berkson and classical errors (Tapsoba et al., 2019; Oh et al., 2018). Bayesian approaches have also been proposed, using parametric models for the baseline hazard to enable full posterior inference over the measurement error structure (Bartlett and Keogh, 2018). Wang and Song (2021) provided a unified semiparametric regression calibration framework for general hazard models, characterizing the RC inconsistency in terms of the calibration residual variance and the baseline hazard. A common thread across all of these methods is that

the researcher is assumed to have direct access to either a validation dataset, replicate measurements, or a known error distribution.

Post-prediction inference. A separate and more recent line of work addresses the problem of conducting valid statistical inference when some variables (typically outcomes) are predicted by a machine learning model rather than directly observed. Wang et al. (2020) formalized the *post-prediction inference* problem, showing that naively substituting predicted outcomes into regression models produces biased estimates and anticonservative standard errors, and proposed corrections based on modeling the relationship between predicted and true outcomes in a labeled testing set. This was applied in the Cox regression setting by Sondhi et al. (2023). Angelopoulos et al. (2023) extended this to *prediction-powered inference* (PPI), a framework that provides valid inference for any estimand defined through an estimating equation. Miao et al. (2025) proposed the PoSt-Prediction Adaptive (PSPA) inference framework, which achieves element-wise variance reduction relative to using only the labeled data, and established connections to semiparametric efficiency theory. These methods assume the researcher has access to a labeled dataset in which both the true and predicted values are jointly observed; the correction is computed from this paired data. Moreover, the methods have been developed primarily for estimands defined through smooth M-estimation problems (means, quantiles, GLM coefficients), and their extension to the Cox partial likelihood (which involves risk-set-dependent weights and a nonparametric baseline hazard) is not immediate.

Quality-assurance frameworks for AI-extracted clinical data. A parallel applied literature specifies how vendors should certify RWD quality. Checklists such as PALISADE (Padula et al., 2022) and SUITABILITY (Fleurence et al., 2024) target transparency and fitness-for-purpose, but were not built specifically for extracted data quality. Closest to our setting is the VALID framework (Estevez et al., 2026), which validates LLM-extracted oncology datasets through three pillars: variable-level accuracy metrics; verification checks for conformance, plausibility, and internal consistency; and replication analyses comparing an analysis run on extracted data against a human-abstracted reference. VALID is the most complete such specification we are aware of, but fails to yield a statement about the reliability of a downstream effect estimate. Its accuracy metrics are marginal, and marginal metrics are not sufficient statistics for Cox bias: two pipelines with identical recall, precision, and F_1 can induce different calibration slope matrices and hence different asymptotic bias in $\hat{\boldsymbol{\beta}}^*$ (Theorem 1). Agreement on a marginal survival curve, as in VALID's case study, does not certify adjusted coefficients in a multivariable Cox model, where extraction errors across several covariates interact. Verification checks are diagnostic rather than corrective. The replication pillar is the most inferentially ambitious, yet it requires the vendor to hold the paired data and to run the analysis, so its guarantee does not transport to the analyses a downstream researcher actually specifies.

3. Estimating Cox Model Bias from Summary Statistics

3.1. Setup and Notation

Let T denote the event time, Δ the event indicator ($\Delta = 1$ if the event is observed, $\Delta = 0$ if censored), and $\mathbf{X} = (X_1, \ldots, X_p)^\top$ a p-vector of covariates. T and Δ are observed exactly;

the covariates are observed only through AI-extracted surrogates $\mathbf{X}^* = (X_1^*, \ldots, X_p^*)^\top$. The true Cox model is

$$\lambda(t \mid \mathbf{X}) = \lambda_0(t) \exp(\boldsymbol{\beta}^\top \mathbf{X}), \tag{1}$$

and $\boldsymbol{\beta}$ is estimated by maximizing the partial likelihood

$$\mathrm{PL}(\boldsymbol{\beta}) = \prod_{i:\Delta_i=1} \frac{\exp(\boldsymbol{\beta}^\top \mathbf{X}_i)}{\sum_{j \in \mathcal{R}(T_i)} \exp(\boldsymbol{\beta}^\top \mathbf{X}_j)},$$

where $\mathcal{R}(t) = \{j : T_j \geq t\}$ is the risk set at time t. We write $\hat{\boldsymbol{\beta}}^*$ for the naive estimator obtained by substituting $\mathbf{X}^*$ for $\mathbf{X}$ in the partial likelihood while keeping T and Δ at their true values, and $\boldsymbol{\beta}^*$ for its asymptotic target. The bias is $\mathbf{b} := \boldsymbol{\beta} - \boldsymbol{\beta}^*$.

Vendor/researcher separation. The data vendor has access to a validation dataset of n_v paired observations $(\mathbf{X}_i, \mathbf{X}_i^*)$ (with outcomes, when available, serving only as stratification variables for diagnostics) and reports accuracy summaries derived from it. The researcher receives only the extracted data $(T_i, \Delta_i, \mathbf{X}_i^*)$ and these summaries. We use "vendor-reported" and "researcher-computed" throughout to indicate provenance.

Working assumptions. Unless otherwise stated, we assume:

(A1) *Cox data-generating model and regularity*: The Cox model (1) is correctly specified for the true data, and censoring is independent of the event time conditional on the covariates. Additionally, we assume the following regularity conditions: (i) the true coefficient $\boldsymbol{\beta}$ lies in a compact set $\mathcal{B} \subset \mathbb{R}^p$; (ii) $(\mathbf{X}, \mathbf{X}^*)$ is jointly sub-Gaussian, the moment generating function $\mathbb{E}[\exp(\boldsymbol{\beta}^\top \mathbf{X})]$ is uniformly bounded on the compact set $\mathcal{B}$, and all polynomial moments of $\mathbf{X}$ and $\mathbf{X}^*$ are finite; (iii) the study horizon $\tau < \infty$ is finite, the baseline hazard λ_0 and censoring survival $G(\cdot \mid \mathbf{X})$ are bounded on $[0, \tau]$, and $s_X^{(0)}(\boldsymbol{\beta}, t) = \mathbb{E}[\exp(\boldsymbol{\beta}^\top \mathbf{X}) \mathbf{1}(T \geq t)]$ is bounded below uniformly in $t \in [0, \tau]$; (iv) the Fisher information is finite and the usual second-moment conditions hold.

(A2) *Non-differential covariate extraction error*: $\mathbf{X}^* \perp (T, \Delta) \mid \mathbf{X}$. The AI extracts covariates based only on the source record's covariate content, not on downstream outcome information.

(A3) *Conditional linear calibration*: the multivariate calibration function is linear: $\mathbb{E}[\mathbf{X} \mid \mathbf{X}^* = \mathbf{x}^*] = \mathbf{a} + \mathbf{B}\mathbf{x}^*$, where $\mathbf{B} = \mathrm{Cov}(\mathbf{X}, \mathbf{X}^*)[\mathrm{Var}(\mathbf{X}^*)]^{-1}$ is the $p \times p$ calibration slope matrix and $\mathbf{a} = \mathbb{E}[\mathbf{X}] - \mathbf{B}\mathbb{E}[\mathbf{X}^*]$. Moreover, $\mathbf{B}$ is nonsingular, and its condition number $\kappa(\mathbf{B})$ is bounded.

(A2) is a standard assumption for characterizing bias due to measurement error. Note that no assumption of independent extraction errors *across covariates* is needed: the unified calibration matrix $\mathbf{B}$ in Section 3.2 captures all cross-covariate correlations automatically. (A3) holds exactly when $(\mathbf{X}, \mathbf{X}^*)$ is jointly Gaussian; it is most defensible for continuous covariates and only approximate for binary ones. Crucially, (A3) is testable on the validation data via the goodness-of-fit of the calibration regression, so vendors can flag the magnitude of violations.

Figure 1 summarizes the methodology. The vendor provides the multivariate calibration slope $\hat{\mathbf{B}}$ from a validation dataset comparing AI-extracted covariates to gold-standard labels. The researcher uses $\hat{\mathbf{B}}$ together with the naive Cox fit on the extracted data to estimate and correct the measurement-error bias. Additional summary statistics and computations yield bias-corrected statistical inference (Section 3.3).

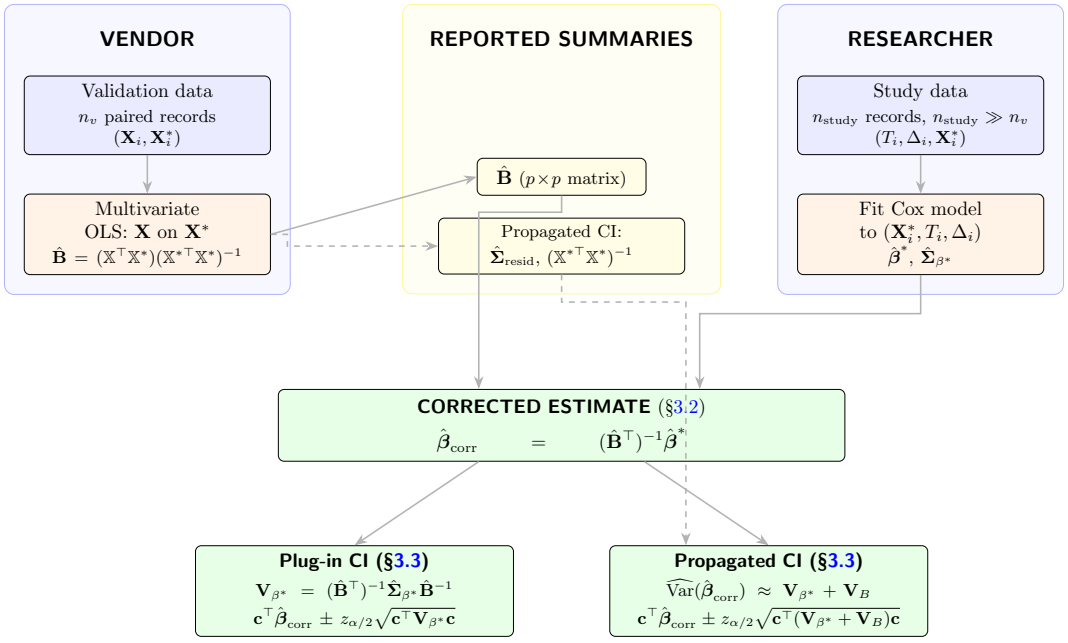

Figure 1: Workflow for bias-corrected inference from AI-extracted covariates in a Cox proportional hazards analysis.

3.2. Bias from Covariate Extraction Error

We treat continuous and binary covariates within a single multivariate regression calibration framework. This unified approach handles all cross-covariate correlations automatically and requires no assumption of independent errors across covariates.

The *multivariate calibration function*

$$\mathbf{m}(\mathbf{x}^*) = \mathbb{E}[\mathbf{X} \mid \mathbf{X}^* = \mathbf{x}^*]$$

maps the full vector of extracted values to the conditional expectation of the truth. Under (A3), $\mathbf{m}(\mathbf{x}^*) = \mathbf{a} + \mathbf{B}\mathbf{x}^*$ with $\mathbf{B}$ and $\mathbf{a}$ as defined in the assumption.

Vendor computation of $\hat{\mathbf{B}}$. The vendor has n_v paired observations $(\mathbf{X}_i, \mathbf{X}_i^*)$ from the validation dataset. The vendor computes $\hat{\mathbf{B}}$ by multivariate OLS: regressing each true covariate column X_j on the full matrix of extracted covariates $\mathbf{X}^*$, for all j. In matrix form, let $\mathbb{X} \in \mathbb{R}^{n_v \times p}$ be the matrix of true covariates and $\mathbb{X}^* \in \mathbb{R}^{n_v \times p}$ the matrix of extracted covariates (both centered). Then

$$\hat{\mathbf{B}} = \left(\mathbb{X}^\top\mathbb{X}^*\right)\left(\mathbb{X}^{*\top}\mathbb{X}^*\right)^{-1}.$$

This is equivalent to running p linear regressions with each column of $\mathbb{X}$ as an outcome on all columns of $\mathbb{X}^*$ as covariates, and forming a $p \times p$ coefficient matrix. Off-diagonal entries capture how extraction error in one variable predicts the truth of another (e.g., if the AI systematically overestimates age for treated patients, $\hat{\mathbf{B}}_{\text{age,trt}} \neq 0$).

Theorem 1 provides a bias derivation that allows us to use $\hat{\mathbf{B}}$ in a corrected estimator:

Theorem 1 (Covariate extraction bias via calibration) *Under (A1)–(A3), the naive Cox estimator's bias satisfies*

$$\mathbf{b} \;=\; \boldsymbol{\beta} - \boldsymbol{\beta}^* \;=\; \left[(\mathbf{B}^\top)^{-1} - \mathbf{I}\right] \boldsymbol{\beta}^* \;-\; \boldsymbol{\Omega}_{\text{RC}}^{-1}\, \mathbf{u}_{\text{RC}}(\boldsymbol{\beta}), \tag{2}$$

where $\boldsymbol{\Omega}_{\text{RC}}^{-1}\, \mathbf{u}_{\text{RC}}(\boldsymbol{\beta})$ *is the* RC residual*:* $\mathbf{u}_{\text{RC}}(\boldsymbol{\beta})$ *is the population regression-calibration score evaluated at the true* $\boldsymbol{\beta}$*, and* $\boldsymbol{\Omega}_{\text{RC}}$ *is the RC information matrix. This residual is of order*

$$\|\boldsymbol{\Omega}_{\text{RC}}^{-1}\, \mathbf{u}_{\text{RC}}(\boldsymbol{\beta})\| \;=\; O(\|\boldsymbol{\beta}\| \cdot \|\mathbb{E}[\boldsymbol{\Sigma}_{X|X^*}]\|^{1/2}),$$

where

$$\boldsymbol{\Sigma}_{X|X^*} = \text{Var}(\mathbf{X} \mid \mathbf{X}^*) = \mathbb{E}[(\mathbf{X} - \mathbf{m}(\mathbf{X}^*))(\mathbf{X} - \mathbf{m}(\mathbf{X}^*))^\top \mid \mathbf{X}^*]$$

is the calibration residual variance*—the conditional variance of the true covariates given the extracted covariates.*

Theorem 1 is a leading-order asymptotic correction valid in the accurate-extraction, approximately-linear regime; it is not an exact finite-sample result. It has been shown that regression calibration is not consistent for the Cox model (Carroll et al., 2006; Xie et al., 2001; Wang and Song, 2021). Bias arises because the induced hazard given calibrated covariates includes a factor depending on $\text{Var}(\mathbf{X} \mid \mathbf{X}^*, T \geq t)$, which varies with t as the risk set composition changes. $\boldsymbol{\Sigma}_{X|X^*}$ measures how much uncertainty about $\mathbf{X}$ remains after observing $\mathbf{X}^*$. When extraction is perfect up to a linear transformation ($\mathbf{X} = \mathbf{a} + \mathbf{B}\mathbf{X}^*$), $\boldsymbol{\Sigma}_{X|X^*} = \mathbf{0}$ and the RC residual vanishes.

The RC residual is not directly estimable from the observed data, as it depends on the unknown baseline hazard and true $\boldsymbol{\beta}$. When the residual is small, which holds when $\|\mathbb{E}[\boldsymbol{\Sigma}_{X|X^*}]\|$ is small (accurate extraction), the bias is dominated by the calibration term:

$$\mathbf{b} \;\approx\; \left[(\mathbf{B}^\top)^{-1} - \mathbf{I}\right] \boldsymbol{\beta}^*, \qquad \hat{\boldsymbol{\beta}}_{\text{corr}} \;=\; (\hat{\mathbf{B}}^\top)^{-1} \hat{\boldsymbol{\beta}}^*. \tag{3}$$

The correction is fully computable from the vendor-reported $\hat{\mathbf{B}}$ and the researcher's naive estimate $\hat{\boldsymbol{\beta}}^*$. When the RC residual is suspected to be non-negligible, its potential magnitude should be assessed with a sensitivity diagnostic (Section 3.3.3).

Table 1 consolidates the validation summaries the vendor must report to make this correction and the inference that follows available to a downstream researcher: the calibration slope matrix $\hat{\mathbf{B}}$ alone suffices for the point correction, while the remaining entries support the propagated intervals of Section 3.3 and the sensitivity diagnostic of Section 3.3.3.

Table 1: Vendor-reported validation summary statistics.

Purpose	**Summary statistic**	**Symbol**
Point correction	Full calibration slope matrix	$\hat{\mathbf{B}}$, $p \times p$
Propagated inference	Residual covariance of calibration regressions	$\hat{\boldsymbol{\Sigma}}_{\text{resid}}$, $p \times p$
	Design Gram inverse	$(\mathbb{X}^{*\top}\mathbb{X}^{*})^{-1}$, $p \times p$
Sensitivity diagnostic	Calibration residual variance norm	$\|\hat{\boldsymbol{\Sigma}}_{X\|X^*}\|_{\text{op}}^{1/2}$

The binary covariate setting. While Theorem 1 covers continuous, binary, and mixed covariates, the discrete case is where the leading-order correction is most strained. The linear calibration assumption (A3) is only an approximation for a binary covariate: the true calibration function $\mathbb{E}[X_j \mid \mathbf{X}^*] = \mathbb{P}(X_j = 1 \mid \mathbf{X}^*)$ is a conditional probability confined to $[0, 1]$, whereas the linear form $\mathbf{a} + \mathbf{B}\mathbf{x}^*$ is unbounded. The vendor's OLS fit therefore recovers the best *linear* predictor of X_j rather than the conditional mean itself, leaving a curvature error $\text{Var}\big(\mathbb{E}[X_j \mid \mathbf{X}^*] - (\mathbf{a} + \mathbf{B}\mathbf{X}^*)_j\big)$ that is zero under Gaussian (A3) but generally nonzero here. A fully nonparametric calibration of $\mathbb{E}[\mathbf{X} \mid \mathbf{X}^*]$ would eliminate it, at the cost of requiring the vendor to report richer, nonlinear calibration summaries.

3.3. Bias-Adjusted Confidence Intervals

The bias formula in Theorem 1 expresses the difference $\boldsymbol{\beta} - \boldsymbol{\beta}^*$ in terms of estimable quantities and provides a corrected estimator $\hat{\boldsymbol{\beta}}_{\text{corr}}$. While this is valid as a point estimate, the extraction error also needs to be incorporated into statistical inference.

Because the calibration matrix $\mathbf{B}$ is generally non-diagonal with off-diagonal entries encoding cross-covariate extraction error correlations, inference should be conducted at the vector level rather than coordinate-wise. To leading order in the RC residual, Theorem 1 gives

$$\hat{\boldsymbol{\beta}}^* \;=\; \mathbf{B}^\top \boldsymbol{\beta} \;+\; \boldsymbol{\varepsilon}, \qquad \boldsymbol{\varepsilon} \sim N(\mathbf{0}, \boldsymbol{\Sigma}_{\beta^*}),$$

where $\boldsymbol{\Sigma}_{\beta^*}$ is the sampling covariance of $\hat{\boldsymbol{\beta}}^*$. The vector-corrected estimator is

$$\hat{\boldsymbol{\beta}}_{\text{corr}} \;=\; (\hat{\mathbf{B}}^\top)^{-1} \hat{\boldsymbol{\beta}}^*.$$

Wald inference on any linear contrast $\mathbf{c}^\top \boldsymbol{\beta}$ (e.g., $\mathbf{c} = \mathbf{e}_k$ for the kth coefficient) proceeds from the multivariate delta method applied to this correction.

3.3.1. Plug-in CI (calibration treated as known)

Assuming $\hat{\mathbf{B}} \approx \mathbf{B}$ with negligible estimation error, the Jacobian is $\partial \hat{\boldsymbol{\beta}}_{\text{corr}} / \partial \hat{\boldsymbol{\beta}}^* = (\hat{\mathbf{B}}^\top)^{-1}$, and the covariance of the corrected estimator is

$$\mathbf{V}_{\beta^*} \;=\; (\hat{\mathbf{B}}^\top)^{-1}\, \hat{\boldsymbol{\Sigma}}_{\beta^*}\, \hat{\mathbf{B}}^{-1}. \tag{4}$$

For a contrast $\mathbf{c}^\top \boldsymbol{\beta}$, the plug-in CI is

$$\text{CI}_{\text{plug}}(\mathbf{c}) \;=\; \mathbf{c}^\top \hat{\boldsymbol{\beta}}_{\text{corr}} \;\pm\; z_{\alpha/2} \sqrt{\mathbf{c}^\top \mathbf{V}_{\beta^*} \mathbf{c}}.$$

The marginal CI for β_k is obtained with $\mathbf{c} = \mathbf{e}_k$ and uses the kth diagonal entry of $\mathbf{V}_{\beta^*}$. This approach is appropriate when the validation sample n_v is large relative to the study, so sampling variability in $\hat{\mathbf{B}}$ is negligible compared to that in $\hat{\boldsymbol{\beta}}^*$.

3.3.2. Propagated CI (calibration estimated with uncertainty)

When $\hat{\mathbf{B}}$ carries non-negligible estimation error, the variance must also reflect uncertainty in the vendor's calibration fit. Treating $\hat{\boldsymbol{\beta}}^*$ and $\hat{\mathbf{B}}$ as independent (the study fit is independent of the validation fit), the multivariate delta method applied to $\hat{\boldsymbol{\beta}}_{\text{corr}} = (\hat{\mathbf{B}}^\top)^{-1}\hat{\boldsymbol{\beta}}^*$ yields two additive variance contributions $\widehat{\text{Var}}(\hat{\boldsymbol{\beta}}_{\text{corr}}) \approx \mathbf{V}_{\beta^*} + \mathbf{V}_B$.

The study-data component $\mathbf{V}_{\beta^*}$ is (4). Assuming all p calibration regressions share the same extracted-covariate design $\mathbb{X}^*$, the covariance of $\hat{\mathbf{B}}$ factors as $\widehat{\text{Cov}}(\hat{B}_{ij}, \hat{B}_{i'j'}) = [\hat{\boldsymbol{\Sigma}}_{\text{resid}}]_{ii'}\,[(\mathbb{X}^{*\top}\mathbb{X}^*)^{-1}]_{jj'}$, where $\hat{\boldsymbol{\Sigma}}_{\text{resid}}$ is the $p \times p$ residual covariance across the p calibration regressions, computed as follows. Consider the $n_v \times p$ residual matrix

$$\mathbf{Z} \;=\; \begin{pmatrix} \mathbf{z}_1^\top \\ \vdots \\ \mathbf{z}_{n_v}^\top \end{pmatrix} \;=\; \mathbb{X} \;-\; \mathbb{X}^*\hat{\mathbf{B}}^\top,$$

whose (i, j) entry is the residual of subject i in the j-th calibration regression. The residual covariance matrix is the sample cross-product, scaled by the residual degrees of freedom:

$$\hat{\boldsymbol{\Sigma}}_{\text{resid}} \;=\; \frac{1}{n_v - p - 1}\,\mathbf{Z}^\top\mathbf{Z} \;\in\; \mathbb{R}^{p\times p}.$$

Propagating this through $\partial\hat{\boldsymbol{\beta}}_{\text{corr}}/\partial\hat{B}_{ij} = -[\hat{\boldsymbol{\beta}}_{\text{corr}}]_i\,(\hat{\mathbf{B}}^\top)^{-1}\mathbf{e}_j$ (from $\partial M^{-1}/\partial M_{ab} = -M^{-1}\mathbf{e}_a\mathbf{e}_b^\top M^{-1}$) yields

$$\mathbf{V}_B \;=\; \big(\hat{\boldsymbol{\beta}}_{\text{corr}}^\top\,\hat{\boldsymbol{\Sigma}}_{\text{resid}}\,\hat{\boldsymbol{\beta}}_{\text{corr}}\big)\,(\hat{\mathbf{B}}^\top)^{-1}\,(\mathbb{X}^{*\top}\mathbb{X}^*)^{-1}\,\hat{\mathbf{B}}^{-1}.$$

This yields the Wald confidence interval

$$\text{CI}_{\text{prop}}(\mathbf{c}) \;=\; \mathbf{c}^\top\hat{\boldsymbol{\beta}}_{\text{corr}} \;\pm\; z_{\alpha/2}\sqrt{\mathbf{c}^\top\big(\mathbf{V}_{\beta^*} + \mathbf{V}_B\big)\mathbf{c}}.$$

The vendor reports two $p \times p$ matrices for propagated inference: the residual covariance $\hat{\boldsymbol{\Sigma}}_{\text{resid}}$ and the design Gram inverse $(\mathbb{X}^{*\top}\mathbb{X}^*)^{-1}$ from the calibration regressions. When $\hat{\mathbf{B}}$ is approximately diagonal, the formulas collapse to coordinate-wise scalar expressions:

$$\text{Var}(\hat{\beta}_{k,\text{corr}}) \;\approx\; \frac{\hat{\sigma}^2_{\beta^*_k}}{\hat{B}^2_{kk}} \;+\; \frac{\hat{\beta}^2_{k,\text{corr}}\,\widehat{\text{Var}}(\hat{B}_{kk})}{\hat{B}^2_{kk}};$$

the general vector form should be used whenever off-diagonal entries of $\hat{\mathbf{B}}$ are non-negligible, representing cross-variable correlation of extraction errors.

If the vendor fits the p calibration regressions with column-specific designs (due to e.g., variable selection or regularization applied separately per outcome), the Kronecker factorization above does not hold and $\mathbf{V}_B$ takes the general form $(\hat{\mathbf{B}}^\top)^{-1}\mathbf{Q}(\hat{\boldsymbol{\beta}}_{\text{corr}})\hat{\mathbf{B}}^{-1}$ with $[\mathbf{Q}(\mathbf{v})]_{jj'} = \sum_{i,i'} v_i v_{i'}\,\widehat{\text{Cov}}(\hat{B}_{ij}, \hat{B}_{i'j'})$, requiring the vendor to report the full covariance of $\hat{\mathbf{B}}$. Note also that regularization biases $\hat{\mathbf{B}}$ toward zero; this bias propagates into $\hat{\boldsymbol{\beta}}_{\text{corr}}$ and should be reported alongside the covariance so the researcher can account for it.

3.3.3. Sensitivity analysis for the RC residual

The plug-in and propagated CIs account for sampling variability in $\hat{\boldsymbol{\beta}}^*$ and (in the propagated case) estimation uncertainty in $\hat{\mathbf{B}}$, but both treat the leading-order approximation $\boldsymbol{\beta} \approx (\mathbf{B}^\top)^{-1}\boldsymbol{\beta}^*$ as exact. Theorem 1 shows that the neglected RC residual contributes an additional systematic error of order $O(\|\boldsymbol{\beta}\| \cdot \|\mathbb{E}[\boldsymbol{\Sigma}_{X|X^*}]\|^{1/2})$. This term cannot be estimated from the observed data, but it can be bounded, and the bound provides a sensitivity diagnostic for judging whether the leading-order correction is adequate.

Suppose the vendor reports the estimated calibration residual variance $\hat{\boldsymbol{\Sigma}}_{\text{resid}}$, which is equal to $\boldsymbol{\Sigma}_{X|X^*}$ in expectation under (A3). Let $\bar{\sigma} = \|\hat{\boldsymbol{\Sigma}}_{\text{resid}}\|_{\text{op}}^{1/2}$ denote the square root of its largest eigenvalue. The RC residual bound implies that the true $\boldsymbol{\beta}$ lies within

$$\|\boldsymbol{\beta} - \hat{\boldsymbol{\beta}}_{\text{corr}}\| \;\le\; \underbrace{(\text{sampling error})}_{\text{covered by Plug-in/Propagated CI}} \;+\; \underbrace{C\,\|\hat{\boldsymbol{\beta}}_{\text{corr}}\|\,\bar{\sigma}}_{\text{RC residual bound}},$$

where C is an unknown constant depending on the covariate, hazard, and censoring distributions.

To judge whether the RC residual would materially change the inference, define the ratio of the RC residual bound to the propagated CI half-width:

$$\rho(\mathbf{c}) \;=\; \frac{|\mathbf{c}^\top \hat{\boldsymbol{\beta}}_{\text{corr}}|\,\bar{\sigma}}{z_{\alpha/2}\sqrt{\mathbf{c}^\top(\mathbf{V}_{\beta^*} + \mathbf{V}_B)\mathbf{c}}}. \tag{5}$$

When $\rho \ll 1$, the RC residual is small relative to the statistical uncertainty already reflected in the CI, and the leading-order correction is adequate. When ρ is appreciable, the propagated CI may understate the true uncertainty, and should be interpreted with caution. Because ρ omits the unknown constant C, it screens for the *potential* of excess bias rather than certifying its absence; a small ρ indicates the residual is dominated by statistical uncertainty for any plausible C, not that the residual is zero.

4. Simulation studies

In this section, we implement simulation studies to examine the performance of our bias-corrected estimator and confidence intervals. We simulate $n = n_{\text{study}} + n_v$ subjects with $p = 4$ covariates subject to AI-extraction error. The true covariate vector $X = (X_1, X_2, X_3, X_4)^\top$ is generated with $X_1 \sim N(0,1)$ and $X_2 \sim N(0,1)$ as continuous covariates, and $X_3 \sim$ Bernoulli(0.5) and $X_4 \sim$ Bernoulli(0.5) as binary covariates. All four are mutually independent. The event time follows a Cox model with exponential baseline hazard,

$$\lambda(t \mid X) = \lambda_0 \exp(\boldsymbol{\beta}^\top X),$$

Censoring times are independently generated as $C \sim$ Exponential(ν). The observed data are $(Y, \Delta) = (\min(T, C),\ \mathbf{1}\{T \le C\})$. In both studies, the true coefficient vector is $\boldsymbol{\beta} = (0.60,\ -0.40,\ 0.50,\ -0.35)^\top$. The baseline hazard rate is $\lambda_0 = 0.1$ and the censoring rate is $\nu = 0.05$. This results in an observed event rate of 0.66. $n_v = 300$ subjects form the validation set, in which the vendor has access to paired observations $(\mathbf{X}_i, \mathbf{X}_i^*)$. The

remaining $n_{\text{study}} = 1500$ subjects form the study set, in which the researcher observes only $(\mathbf{X}_i^*, Y_i, \Delta_i)$ and the vendor-reported calibration summaries.

In each simulation replicate, the vendor computes $\hat{\mathbf{B}}$, $\hat{\boldsymbol{\Sigma}}_{\text{resid}}$, and $(\mathbb{X}^{*\top}\mathbb{X}^*)^{-1}$ from the validation data. The researcher then fits the naive Cox model $\hat{\boldsymbol{\beta}}^*$ on the study data using $\mathbf{X}^*$ in place of $\mathbf{X}$, and computes the corrected estimate $\hat{\boldsymbol{\beta}}_{\text{corr}} = (\hat{\mathbf{B}}^\top)^{-1}\hat{\boldsymbol{\beta}}^*$ together with plug-in and propagated confidence intervals. Each scenario is replicated $R = 500$ times. For each covariate k and estimator (naive or corrected), we report the signed bias $R^{-1}\sum_{r=1}^{R}(\hat{\beta}_k^{(r)} - \beta_k)$ and the root mean squared error $\left[R^{-1}\sum_{r=1}^{R}(\hat{\beta}_k^{(r)} - \beta_k)^2\right]^{1/2}$. We also report the CI coverage (the proportion of replicates in which the 95% confidence interval contains the true β_k) for each CI method. All simulations are run in R version 4.4.1 (R Core Team, 2024), using the `survival` package version 3.8-3 (Therneau, 2024) for fitting Cox models with the default Efron approximation for ties.

We additionally conduct a study of real data with synthetic covariate measurement error applied. This is reported in Appendix C.

4.1. Study 1: Cross-Dependent Extraction Errors

In this study, a shared latent factor $Z_i \sim N(0,1)$, drawn independently per subject and independently of X, drives correlated extraction errors across all four covariates. The extracted continuous covariates ($j = 1, 2$) are generated as

$$X_j^* = X_j + \lambda_j Z + \varepsilon_j, \qquad \varepsilon_j \overset{\text{iid}}{\sim} N(0, \sigma_{\varepsilon,j}^2),$$

where ε_j, Z, X are mutually independent. The factor loading λ_j induces cross-covariate correlation through the shared Z, while ε_j adds covariate-specific error. For X_1, the factor loading is $\lambda_1 = 0.6$ with noise $\sigma_{\varepsilon,1} = 0.25$; for X_2, $\lambda_2 = 0.5$ and $\sigma_{\varepsilon,2} = 0.20$. For binary covariates ($j = 3, 4$), the misclassification probability depends on the same latent factor:

$$P(X_j^* \neq X_j \mid Z) = \text{logit}^{-1}(\alpha_j + \gamma_j|Z|).$$

For X_3, the intercept is $\alpha_3 = -2.5$ and the sensitivity to the latent factor is $\gamma_3 = 1.5$; for X_4, $\alpha_4 = -2.8$ and $\gamma_4 = 1.2$. Because both the continuous additive error ($\lambda_j Z$) and the binary flip probability ($\gamma_j|Z|$) depend on the same Z, extraction errors are correlated across all four covariates. The multivariate calibration matrix $\mathbf{B}$ captures this structure through its off-diagonal entries.

Table 2 summarizes the simulation results. Extraction quality is moderate to high across all covariates. The naive estimator exhibits substantial bias: X_1 is attenuated by -0.114 (19% of $\beta_1 = 0.60$) and X_3 by -0.256 (50% of $\beta_3 = 0.50$), while X_4 is biased by $+0.121$ (35% of $|\beta_4| = 0.35$). The correction eliminates most of this bias, reducing signed bias to below 0.03 in absolute value for all covariates. Notably, X_2 has near-zero naive bias; the correction slightly increases its absolute bias to 0.02, reflecting the variance cost of the matrix inversion in $(\hat{\mathbf{B}}^\top)^{-1}$. RMSE follows the same pattern: the correction reduces RMSE for most variables where bias dominates the naive error, but slightly inflates RMSE for X_2, where the naive estimator was already nearly unbiased and the correction adds estimation variance.

Naive confidence intervals have poor coverage in general. Plug-in CIs, which treat $\hat{\mathbf{B}}$ as known, improve coverage substantially (82–92%) but remain below the nominal 95% level

Table 2: Simulation results for cross-dependent extraction errors (Study 1). Each cell reports the point estimate with a 95% Monte Carlo confidence interval in brackets. Extraction quality is R^2 for continuous covariates (X_1, X_2) and classification accuracy for binary covariates (X_3, X_4).

		Signed Bias		RMSE		Coverage (%)		
Covariate	Ext. Quality	Naive	Corrected	Naive	Corrected	Naive	Plug-in	Propagated
X_1	0.70	−0.114	−0.029	0.118	0.057	3.6	82.4	98.4
	[0.70, 0.70]	[−0.117, −0.111]	[−0.033, −0.024]	[0.115, 0.121]	[0.053, 0.060]	[2.0, 5.2]	[79.1, 85.7]	[97.3, 99.5]
X_2	0.77	−0.008	0.020	0.031	0.048	94.4	88.0	99.2
	[0.77, 0.78]	[−0.010, −0.005]	[0.017, 0.024]	[0.029, 0.033]	[0.045, 0.051]	[92.4, 96.4]	[85.2, 90.8]	[98.4, 100.0]
X_3	0.76	−0.256	−0.015	0.264	0.148	3.8	90.0	94.4
	[0.75, 0.76]	[−0.262, −0.251]	[−0.028, −0.002]	[0.259, 0.270]	[0.139, 0.158]	[2.1, 5.5]	[87.4, 92.6]	[92.4, 96.4]
X_4	0.84	0.121	0.012	0.136	0.108	50.0	92.2	95.2
	[0.84, 0.84]	[0.115, 0.126]	[0.003, 0.022]	[0.131, 0.141]	[0.102, 0.114]	[45.6, 54.4]	[89.8, 94.6]	[93.3, 97.1]

for most covariates, indicating that calibration uncertainty is non-negligible with $n_v = 300$. Propagated CIs, which account for uncertainty in $\hat{\mathbf{B}}$, achieve near or above nominal coverage for all covariates. Appendix B demonstrates the convergence of CI coverage with increasing n_v under this study's data-generating process.

4.2. Study 2: Nonlinear Calibration (Constant Total Error)

This study tests robustness to deliberate, controlled violations of assumption (A3). The extracted continuous covariates ($j = 1, 2$) are generated as

$$X_j^* = X_j + \underbrace{\delta_j X_j^2}_{\text{quadratic distortion}} + \underbrace{\sigma_{0,j}(1 + \kappa_j X_j^2)\,\varepsilon_j}_{\text{heteroscedastic noise}}, \qquad \varepsilon_j \sim N(0,1).$$

The error has two nonlinear components. The quadratic distortion term $\delta_j X_j^2$ introduces a systematic, mean-shifting nonlinearity. The heteroscedastic noise term, governed by κ_j, makes the extraction noisier for extreme covariate values. Together, $\text{MSE}_j = 3\delta_j^2 + \sigma_{0,j}^2(1 + 2\kappa_j + 3\kappa_j^2)$. For binary covariates ($j = 3, 4$), the misclassification probability is modulated by the true continuous covariates through

$$P(X_j^* \neq X_j \mid X_1, X_2) = p_j\big(1 + \eta_j \tanh(X_1 + X_2)\big),$$

where p_j is the target marginal flip rate and $\eta_j \in [0, 1)$ controls the strength of the covariate dependence. Since tanh is an odd function and $X_1 + X_2$ is symmetric around zero, $E[\tanh(X_1 + X_2)] = 0$, so the marginal flip rate is exactly p_j regardless of η_j. When $\eta_j = 0$, flips are uniform and the calibration $E[X_j \mid \mathbf{X}^*]$ is linear. When $\eta_j > 0$, the flip probability varies smoothly from $p_j(1 - \eta_j)$ to $p_j(1 + \eta_j)$ across the full range of $X_1 + X_2$, making $E[X_j \mid \mathbf{X}^*]$ a nonlinear function of (X_1^*, X_2^*) and violating (A3).

The continuous errors ε_1 and ε_2 are independent, and the binary flips for X_3 and X_4 are conditionally independent given (X_1, X_2). The only source of dependence between extracted covariates is through the true covariates themselves: X_3^* and X_4^* both depend on

$X_1 + X_2$ (which governs their flip probabilities), so they are marginally correlated with X_1^* and X_2^*. Off-diagonal entries in the estimated $\hat{\mathbf{B}}$ will be nonzero, but the extraction *errors* are conditionally uncorrelated given the true covariates. This contrasts with Study 1, where the shared Z induces genuine cross-covariate error correlation. We investigate three levels of nonlinearity:

1. Under **mild nonlinearity**, $\delta = 0.02$, $\kappa = 0.05$, and $\eta = 0.10$.
2. Under **moderate nonlinearity**, $\delta = 0.15$, $\kappa = 0.30$, and $\eta = 0.50$.
3. Under **severe nonlinearity**, $\delta = 0.25$, $\kappa = 0.55$, and $\eta = 0.90$.

To isolate the effect of nonlinearity shape, $\sigma_{0,j}$ is calibrated so that the marginal MSE is identical across severity levels, with targets $\text{MSE}_1 = \text{MSE}_2 = 0.70$. The binary flip rates are fixed at $p_3 = p_4 = 0.20$ and are automatically preserved by the tanh modulation. Thus, any change in bias, RMSE, or CI coverage between severity levels is attributable to the nonlinearity of the calibration function, not to the amount of error.

Table 3 displays RMSE and CI coverage as nonlinearity increases. Naive RMSE is large and mostly flat across severity levels for all four covariates, consistent with the constant-error design: the naive estimator is always biased, and the shape of the calibration function does not affect its magnitude. Corrected RMSE is substantially smaller at all severity levels, while increasing with nonlinearity for covariates X_1, X_4.

Naive coverage is near zero for X_1 and X_2 at all severity levels, reflecting severe attenuation bias in those covariates. Plug-in coverage improves substantially over naive but falls short of the nominal level, as it does not account for uncertainty in $\hat{\mathbf{B}}$. Propagated CI coverage degrades with severity, maintaining nominal coverage only for X_2. The binary covariates (X_3, X_4) exhibit below-nominal propagated coverage even at mild severity (88.0% and 87.6%), reflecting residual bias from the calibration function that the leading-order correction does not fully capture.

5. Discussion

This paper develops a framework for quantifying and correcting the bias introduced by AI-extracted covariates in the Cox proportional hazards model, under the assumption that event times and indicators are accurately observed. The central theoretical result (Theorem 1) yields a fully computable corrected estimator $\hat{\boldsymbol{\beta}}_{\text{corr}} = (\hat{\mathbf{B}}^\top)^{-1}\hat{\boldsymbol{\beta}}^*$ that requires only the vendor-reported $\hat{\mathbf{B}}$ and the researcher's naive Cox fit. Bias-adjusted confidence intervals provide inference that honestly reflects both sampling variability in the study data and estimation uncertainty in the vendor's validation sample. Our framework cleanly separates vendor and researcher responsibilities: the vendor should report the calibration summaries in Table 1 and the researcher combines these with the output of any standard Cox software. We encourage vendors to adopt Table 1 as a reporting standard for any AI-extracted covariate dataset intended for use in survival analysis, particularly as extensions to emerging and existing guidelines for real-world-evidence reporting.

R code for implementing our bias-correction methodology and for replicating the empirical studies described in Section 4 and Appendix C is available at https://github.com/asondhi/cox-bias-cov.

Table 3: Simulation results under nonlinear calibration (Study 2). Each cell gives the point estimate over its 95% Monte Carlo CI.

		RMSE		95% CI Coverage (%)		
Cov.	Severity	Naive	Corr.	Naive	Plug-in	Prop.
X_1	Mild	0.290 [0.288, 0.292]	0.090 [0.086, 0.094]	0.0 [0.0, 0.0]	57.8 [53.5, 62.1]	95.6 [93.8, 97.4]
	Moderate	0.288 [0.286, 0.290]	0.094 [0.090, 0.098]	0.0 [0.0, 0.0]	53.4 [49.0, 57.8]	91.6 [89.2, 94.0]
	Severe	0.293 [0.290, 0.296]	0.114 [0.108, 0.119]	0.0 [0.0, 0.0]	35.8 [31.6, 40.0]	81.2 [77.8, 84.6]
X_2	Mild	0.194 [0.192, 0.196]	0.074 [0.070, 0.078]	0.0 [0.0, 0.0]	71.4 [67.4, 75.4]	96.8 [95.3, 98.3]
	Moderate	0.186 [0.184, 0.188]	0.066 [0.062, 0.070]	0.0 [0.0, 0.0]	80.4 [76.9, 83.9]	98.2 [97.0, 99.4]
	Severe	0.176 [0.173, 0.178]	0.069 [0.065, 0.073]	0.0 [0.0, 0.0]	77.6 [73.9, 81.3]	97.2 [95.8, 98.6]
X_3	Mild	0.239 [0.233, 0.245]	0.151 [0.142, 0.160]	5.8 [3.8, 7.8]	84.2 [81.0, 87.4]	88.0 [85.2, 90.8]
	Moderate	0.246 [0.240, 0.252]	0.148 [0.139, 0.157]	5.0 [3.1, 6.9]	85.4 [82.3, 88.5]	88.0 [85.2, 90.8]
	Severe	0.239 [0.234, 0.245]	0.150 [0.140, 0.160]	5.0 [3.1, 6.9]	83.4 [80.1, 86.7]	86.6 [83.6, 89.6]
X_4	Mild	0.178 [0.173, 0.184]	0.144 [0.135, 0.152]	25.8 [22.0, 29.6]	85.8 [82.7, 88.9]	87.6 [84.7, 90.5]
	Moderate	0.176 [0.170, 0.182]	0.151 [0.141, 0.160]	30.6 [26.6, 34.6]	84.6 [81.4, 87.8]	88.0 [85.2, 90.8]
	Severe	0.176 [0.171, 0.182]	0.158 [0.147, 0.168]	28.4 [24.4, 32.4]	83.4 [80.1, 86.7]	85.6 [82.5, 88.7]

Limitations. The framework rests on several assumptions whose relaxation defines natural directions for future work. First, we assume that the event time T and event indicator Δ are observed exactly. This is reasonable when outcomes come from administrative or registry sources that do not require language extraction, but it excludes settings where outcomes are themselves AI-extracted; for example, when the event of interest (disease recurrence, treatment discontinuation) must be identified from unstructured clinical notes.

Each error source enters the partial likelihood differently. Covariate error leaves every risk set correct and produces the linear attenuation of Theorem 1, which vanishes at $\boldsymbol{\beta} = \mathbf{0}$. Misclassified Δ changes only which records count as events. Events missed at a rate that does not depend on $\mathbf{X}$ or t merely thin the event set and are absorbed into λ_0, so imperfect sensitivity costs precision but not consistency. False events do bias $\boldsymbol{\beta}$: a censored record is labeled an event at the time its follow-up ends, not at a time generated by the failure hazard, so its covariate is effectively a draw from the *unweighted* risk set at that time, whereas the true score subtracts the risk-set mean $\bar{\mathbf{X}}(\boldsymbol{\beta}, t)$ weighted by $\exp(\boldsymbol{\beta}^\top \mathbf{X})$; the discrepancy is $-\mathrm{Var}(\mathbf{X} \mid T \geq t)\boldsymbol{\beta}$ to first order and therefore pulls the estimate toward zero. Error in T perturbs event ordering and risk-set membership together, so the partial likelihood is invariant to a common monotone distortion of the time scale but is biased by heterogeneous error.

When T or Δ is subject to extraction error, additional bias terms arise from the interaction between outcome misclassification and the partial likelihood structure, and the non-differential error assumption requires a different form. Extending the framework to handle errors in T, Δ, or both is an important open problem; the joint treatment of covariate and outcome extraction error in the Cox model has not, to our knowledge, been addressed.

Second, the present framework addresses estimation of the full-population coefficient vector $\boldsymbol{\beta}$ and does not cover subpopulation analyses in which the analytic sample is selected on the basis of an AI-extracted variable. Such analyses are common in practice: for example, a researcher may restrict attention to patients with a particular AI-extracted diagnosis or biomarker status, effectively conditioning on X_j^* falling in a specific category. Here, even if the extraction error is non-differential in the full population, the subpopulation defined by $X_j^* = 1$ will systematically differ from the subpopulation defined by $X_j = 1$ due to misclassification, and the induced bias depends on both the misclassification rates and the joint distribution of X_j with the other covariates and with the outcome.

Finally, the bias formulas are leading-order approximations that rely on the linear calibration assumption (A3). As the simulations in Study 2 show, the correction remains effective under mild departures from linearity but degrades under severe nonlinearity, particularly for binary covariates. Nonlinear extensions—for example, replacing the linear calibration model with a flexible nonparametric or semiparametric specification of $\mathbb{E}[\mathbf{X} \mid \mathbf{X}^*]$—would broaden the applicability of the framework, though at the cost of requiring the vendor to report richer summaries.

References

Blythe Adamson, Michael Waskom, Auriane Blarre, Jonathan Kelly, Konstantin Krismer, Sheila Nemeth, James Gippetti, John Ritten, Katherine Harrison, George Ho, et al. Approach to machine learning for extraction of real-world data variables from electronic health records. *Frontiers in Pharmacology*, 14:1180962, 2023.

Smita Agrawal, Rohini George, Vivek Prabhakar Vaidya, Sangavai Chakkrapani, Rambaksh Prajapati, Srikanth Tankala, Dhaval Parmar, Vinay Phani Santosh Lakkimsetty, Tapasya Bhardwaj, Ashwani Ashwani, et al. Development of natural language processing (nlp) models for extracting key features from unstructured notes to create real-world data (rwd) assets for clinical research at scale., 2023.

Anastasios N Angelopoulos, Stephen Bates, Clara Fannjiang, Michael I Jordan, and Tijana Zrnic. Prediction-powered inference. *Science*, 382(6671):669–674, 2023.

Jonathan W Bartlett and Ruth H Keogh. Bayesian correction for covariate measurement error: A frequentist evaluation and comparison with regression calibration. *Statistical Methods in Medical Research*, 27(6):1695–1708, 2018.

Raymond J Carroll, David Ruppert, Leonard A Stefanski, and Ciprian M Crainiceanu. *Measurement error in nonlinear models: a modern perspective*. Chapman and Hall/CRC, 2006.

Melissa Estevez, Nisha Singh, Lauren Dyson, Blythe Adamson, Konstantin Krismer, Kelly Magee, Qianyu Yuan, Megan W Hildner, Erin Fidyk, Tori Williams, et al. Ensuring reliability of curated electronic health record–derived data: The validation of accuracy for large language model–/machine learning–extracted information and data (valid) framework. *JCO Clinical Cancer Informatics*, 10:e2500215, 2026.

Rachael L Fleurence, Seamus Kent, Blythe Adamson, James Tcheng, Ran Balicer, Joseph S Ross, Kevin Haynes, Patrick Muller, Jon Campbell, Elsa Bouée-Benhamiche, et al. Assessing real-world data from electronic health records for health technology assessment: the suitability checklist: a good practices report of an ispor task force. *Value in Health*, 27(6):692–701, 2024.

Wayne A Fuller. *Measurement error models*. John Wiley & Sons, 2009.

Ary L Goldberger, Luis AN Amaral, Leon Glass, Jeffrey M Hausdorff, Plamen Ch Ivanov, Roger G Mark, Joseph E Mietus, George B Moody, Chung-Kang Peng, and H Eugene Stanley. Physiobank, physiotoolkit, and physionet: components of a new research resource for complex physiologic signals. *circulation*, 101(23):e215–e220, 2000.

Alistair Johnson, Lucas Bulgarelli, Tom Pollard, Brian Gow, Benjamin Moody, Steven Horng, Leo Anthony Celi, and Roger Mark. MIMIC-IV. *PhysioNet*, October 2024. doi: 10.13026/kpb9-mt58. URL https://doi.org/10.13026/kpb9-mt58. Version 3.1.

Alistair EW Johnson, Lucas Bulgarelli, Lu Shen, Alvin Gayles, Ayad Shammout, Steven Horng, Tom J Pollard, Sicheng Hao, Benjamin Moody, Brian Gow, et al. Mimic-iv, a freely accessible electronic health record dataset. *Scientific data*, 10(1):1, 2023.

Jiacheng Miao, Xinran Miao, Yixuan Wu, Jiwei Zhao, and Qiongshi Lu. Assumption-lean and data-adaptive post-prediction inference. *Journal of Machine Learning Research*, 26 (179):1–31, 2025.

Eric J Oh, Bryan E Shepherd, Thomas Lumley, and Pamela A Shaw. Considerations for analysis of time-to-event outcomes measured with error: bias and correction with simex. *Statistics in medicine*, 37(8):1276–1289, 2018.

William V Padula, Noemi Kreif, David J Vanness, Blythe Adamson, Juan-David Rueda, Federico Felizzi, Pall Jonsson, Maarten J IJzerman, Atul Butte, and William Crown. Machine learning methods in health economics and outcomes research—the palisade checklist: a good practices report of an ispor task force. *Value in health*, 25(7):1063–1080, 2022.

R Core Team. *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria, 2024. URL https://www.R-project.org/.

Arjun Sondhi, Alexander S Rich, Siruo Wang, and Jeffery T Leek. Postprediction inference for clinical characteristics extracted with machine learning on electronic health records. *JCO Clinical Cancer Informatics*, 7:e2200174, 2023.

Jean de Dieu Tapsoba, Edward C Chao, and Ching-Yun Wang. Simulation extrapolation method for cox regression model with a mixture of berkson and classical errors in the covariates using calibration data. *The international journal of biostatistics*, 15(2):20180028, 2019.

Terry M Therneau. *A Package for Survival Analysis in R*, 2024. URL https://CRAN.R-project.org/package=survival. R package version 3.8-3.

Ching-Yun Wang and Xiao Song. Semiparametric regression calibration for general hazard models in survival analysis with covariate measurement error; surprising performance under linear hazard. *Biometrics*, 77(2):561–572, 2021.

CY Wang, Li Hsu, ZD Feng, and Ross L Prentice. Regression calibration in failure time regression. *Biometrics*, pages 131–145, 1997.

Siruo Wang, Tyler H McCormick, and Jeffrey T Leek. Methods for correcting inference based on outcomes predicted by machine learning. *Proceedings of the National Academy of Sciences*, 117(48):30266–30275, 2020.

Sharon X Xie, CY Wang, and Ross L Prentice. A risk set calibration method for failure time regression by using a covariate reliability sample. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 63(4):855–870, 2001.

Appendix A. Proof of Theorem 1

Asymptotic regime and conventions. All asymptotic bounds in this appendix are stated in the limit $\|\mathbb{E}[\boldsymbol{\Sigma}_{X|X^*}]\| \to 0$ (the accurate-extraction regime), with $\boldsymbol{\beta} \in \mathcal{B}$ held fixed per (A1). Under (A1), hidden constants in big-O bounds depend on $\mathcal{B}$ and the covariate, hazard, and censoring distributions. Specifically, under the joint sub-Gaussianity in (A1)(ii), the moment generating function $M_X(\boldsymbol{\beta}) = \mathbb{E}[\exp(\boldsymbol{\beta}^\top \mathbf{X})]$ is finite and continuous, hence uniformly bounded on the compact set $\mathcal{B}$, and the same holds for the tilted moments $\mathbb{E}[\exp(\boldsymbol{\beta}^\top \mathbf{X})\mathbf{1}(T \geq t)]$ and $\mathbb{E}[\|\mathbf{X}\|^k \exp(\boldsymbol{\beta}^\top \mathbf{X})\mathbf{1}(T \geq t)]$ uniformly on $\mathcal{B} \times [0, \tau]$; $s_X^{(0)}(\boldsymbol{\beta}, t)$ and $\mathbb{P}(T \geq t)$ are bounded below uniformly on $[0, \tau]$ by (A1)(iii). Expectations of $\|\mathbf{X}\|^k$ and $\|\boldsymbol{\eta}(\mathbf{X})\|^k$ are finite; and, since under (A3) the calibration residual $\boldsymbol{\epsilon}_m = \mathbf{X} - \mathbf{m}(\mathbf{X}^*)$ is a linear combination of the jointly sub-Gaussian $(\mathbf{X}, \mathbf{X}^*)$, it is itself sub-Gaussian.

Proof We proceed in three steps.

Step 1: Equivalence of naive and regression calibration estimators. The *regression calibration (RC) estimator* replaces $\mathbf{X}^*$ with $\mathbf{m}(\mathbf{X}^*) = \mathbf{a} + \mathbf{B}\mathbf{X}^*$ in the partial likelihood and maximizes over $\boldsymbol{\beta}$. The partial likelihood contribution at event time T_i under RC is

$$\frac{\exp(\boldsymbol{\beta}^\top \mathbf{m}(\mathbf{X}_i^*))}{\sum_{j \in \mathcal{R}(T_i)} \exp(\boldsymbol{\beta}^\top \mathbf{m}(\mathbf{X}_j^*))}.$$

Expanding the exponent: $\boldsymbol{\beta}^\top \mathbf{m}(\mathbf{X}_i^*) = \boldsymbol{\beta}^\top(\mathbf{a} + \mathbf{B}\mathbf{X}_i^*) = \boldsymbol{\beta}^\top \mathbf{a} + (\mathbf{B}^\top \boldsymbol{\beta})^\top \mathbf{X}_i^*$. Since $\exp(\boldsymbol{\beta}^\top \mathbf{a})$ appears in every term of both numerator and denominator, it cancels:

$$\frac{\exp(\boldsymbol{\beta}^\top \mathbf{m}(\mathbf{X}_i^*))}{\sum_{j \in \mathcal{R}(T_i)} \exp(\boldsymbol{\beta}^\top \mathbf{m}(\mathbf{X}_j^*))} = \frac{\exp((\mathbf{B}^\top \boldsymbol{\beta})^\top \mathbf{X}_i^*)}{\sum_{j \in \mathcal{R}(T_i)} \exp((\mathbf{B}^\top \boldsymbol{\beta})^\top \mathbf{X}_j^*)}.$$

The right-hand side is the naive partial likelihood with covariates $\mathbf{X}^*$ at coefficient $\mathbf{B}^\top \boldsymbol{\beta}$. Since the map $\boldsymbol{\beta} \mapsto \mathbf{B}^\top \boldsymbol{\beta}$ is a bijection (assuming $\mathbf{B}$ is invertible), the maximizers are related by

$$\hat{\boldsymbol{\beta}}^* = \mathbf{B}^\top \hat{\boldsymbol{\beta}}_{\mathrm{RC}}.$$

The bias problem therefore reduces to showing that $\hat{\boldsymbol{\beta}}_{\mathrm{RC}}$ approximately targets $\boldsymbol{\beta}$.

Step 2: Approximate consistency of the RC estimator. The population RC score at $\boldsymbol{\beta}$ is

$$\mathbf{u}_{\mathrm{RC}}(\boldsymbol{\beta}) = \mathbb{E}\Big[\Delta\big\{\mathbf{m}(\mathbf{X}^*) - \bar{\mathbf{m}}(\boldsymbol{\beta}, T)\big\}\Big],$$

where

$$\bar{\mathbf{m}}(\boldsymbol{\beta}, t) = \frac{\mathbb{E}[\mathbf{m}(\mathbf{X}^*) \exp(\boldsymbol{\beta}^\top \mathbf{m}(\mathbf{X}^*))\mathbf{1}(T \geq t)]}{\mathbb{E}[\exp(\boldsymbol{\beta}^\top \mathbf{m}(\mathbf{X}^*))\mathbf{1}(T \geq t)]}$$

is the risk-set weighted average of the calibrated covariates. We need to show that $\mathbf{u}_{\mathrm{RC}}(\boldsymbol{\beta}) \approx \mathbf{0}$.

Add and subtract $\mathbf{X}$ and $\bar{\mathbf{X}}(\boldsymbol{\beta}, T)$ (the true risk-set mean) to decompose:

$$\begin{aligned}\mathbf{u}_{\mathrm{RC}}(\boldsymbol{\beta}) = & \underbrace{\mathbb{E}\big[\Delta\big\{\mathbf{m}(\mathbf{X}^*) - \mathbf{X}\big\}\big]}_{\text{Term A: double-projection}} \\ & + \underbrace{\mathbb{E}\big[\Delta\big\{\mathbf{X} - \bar{\mathbf{X}}(\boldsymbol{\beta}, T)\big\}\big]}_{\text{Term B: true Cox score}} \\ & + \underbrace{\mathbb{E}\big[\Delta\big\{\bar{\mathbf{X}}(\boldsymbol{\beta}, T) - \bar{\mathbf{m}}(\boldsymbol{\beta}, T)\big\}\big]}_{\text{Term C: risk-set weight discrepancy}}.\end{aligned}$$

Term B is the true Cox score $\mathbf{u}(\boldsymbol{\beta}) = \mathbf{0}$ by definition. We bound Terms A and C.

Step 2a: The double-projection residual. Under non-differential error (A2), $\mathbf{X}^* \perp (\Delta, T) \mid \mathbf{X}$, so for any function $h(\mathbf{X}^*)$:

$$\mathbb{E}[\Delta \cdot h(\mathbf{X}^*)] = \mathbb{E}\big[\mathbb{E}[\Delta \mid \mathbf{X}] \cdot \mathbb{E}[h(\mathbf{X}^*) \mid \mathbf{X}]\big].$$

Applying with $h(\mathbf{X}^*) = \mathbf{m}(\mathbf{X}^*) - \mathbf{X}$:

$$\mathbb{E}\big[\Delta\big\{\mathbf{m}(\mathbf{X}^*) - \mathbf{X}\big\}\big] = \mathbb{E}\big[\mathbb{E}[\Delta \mid \mathbf{X}] \cdot \big(\mathbb{E}[\mathbf{m}(\mathbf{X}^*) \mid \mathbf{X}] - \mathbf{X}\big)\big].$$

Define the *double-projection residual*

$$\boldsymbol{\eta}(\mathbf{X}) := \mathbb{E}[\mathbf{m}(\mathbf{X}^*) \mid \mathbf{X}] - \mathbf{X}. \tag{6}$$

This is the discrepancy between $\mathbf{X}$ and its double projection $\mathbb{E}[\mathbb{E}[\mathbf{X} \mid \mathbf{X}^*] \mid \mathbf{X}]$. By the law of total expectation, $\mathbb{E}[\boldsymbol{\eta}(\mathbf{X})] = \mathbb{E}[\mathbb{E}[\mathbf{X} \mid \mathbf{X}^*]] - \mathbb{E}[\mathbf{X}] = \mathbf{0}$.

Under (A3), $\mathbf{m}(\mathbf{x}^*) = \mathbf{a} + \mathbf{B}\mathbf{x}^*$ with $\mathbf{a} = \boldsymbol{\mu}_X - \mathbf{B}\boldsymbol{\mu}_{X^*}$:

$$\begin{aligned}\mathbb{E}[\mathbf{m}(\mathbf{X}^*) \mid \mathbf{X}] &= \mathbf{a} + \mathbf{B}\,\mathbb{E}[\mathbf{X}^* \mid \mathbf{X}] \\ &= \boldsymbol{\mu}_X - \mathbf{B}\boldsymbol{\mu}_{X^*} + \mathbf{B}\,\mathbb{E}[\mathbf{X}^* \mid \mathbf{X}],\end{aligned}$$

so

$$\boldsymbol{\eta}(\mathbf{X}) = \mathbf{B}\big(\mathbb{E}[\mathbf{X}^* \mid \mathbf{X}] - \boldsymbol{\mu}_{X^*}\big) - (\mathbf{X} - \boldsymbol{\mu}_X).$$

We now show that $\mathrm{Var}(\boldsymbol{\eta}(\mathbf{X}))$ is controlled by $\mathbb{E}[\boldsymbol{\Sigma}_{X|X^*}]$. Define the calibration residual $\boldsymbol{\epsilon}_m := \mathbf{X} - \mathbf{m}(\mathbf{X}^*)$ (the difference between the truth and the calibration prediction). Note that

$$\mathbb{E}[\boldsymbol{\epsilon}_m] = \mathbb{E}[\mathbf{X}] - \mathbb{E}[\mathbf{m}(\mathbf{X}^*)] = \mathbb{E}[\mathbf{X}] - \mathbb{E}[\mathbb{E}[\mathbf{X} \mid \mathbf{X}^*]] = \mathbf{0}$$

and

$$\mathbb{E}[\boldsymbol{\epsilon}_m \mid \mathbf{X}] = \mathbf{X} - \mathbb{E}[\mathbf{m}(\mathbf{X}^*) \mid \mathbf{X}] = -\boldsymbol{\eta}(\mathbf{X}).$$

Then, the law of total variance gives

$$\begin{aligned}\mathrm{Var}(\boldsymbol{\epsilon}_m) &= \mathrm{Var}(\mathbb{E}[\boldsymbol{\epsilon}_m \mid \mathbf{X}]) + \mathbb{E}[\mathrm{Var}(\boldsymbol{\epsilon}_m \mid \mathbf{X})] \\ &= \mathrm{Var}(-\boldsymbol{\eta}(\mathbf{X})) + \mathbb{E}[\mathrm{Var}(\boldsymbol{\epsilon}_m \mid \mathbf{X})]\end{aligned}$$

Taking traces and using $\mathbb{E}[\boldsymbol{\epsilon}_m] = \mathbf{0}$:

$$\underbrace{\mathbb{E}[\|\boldsymbol{\epsilon}_m\|^2]}_{\mathrm{tr}(\mathrm{Var}(\boldsymbol{\epsilon}_m))} = \underbrace{\mathbb{E}[\|\boldsymbol{\eta}(\mathbf{X})\|^2]}_{\mathrm{tr}(\mathrm{Var}(\boldsymbol{\eta}))} + \underbrace{\mathbb{E}[\mathbb{E}[\|\boldsymbol{\epsilon}_m - \mathbb{E}[\boldsymbol{\epsilon}_m \mid \mathbf{X}]\|^2 \mid \mathbf{X}]]}_{\mathrm{tr}(\mathbb{E}[\mathrm{Var}(\boldsymbol{\epsilon}_m|\mathbf{X})]) \geq 0},$$

where the first equality on the left uses $\mathrm{tr}(\mathrm{Var}(Y)) = \mathbb{E}[\|Y\|^2]$ when $\mathbb{E}[Y] = \mathbf{0}$, and similarly for $\boldsymbol{\eta}$.

The left hand side equals $\mathrm{tr}(\mathbb{E}[\boldsymbol{\Sigma}_{X|X^*}])$ since $\mathbb{E}[\|\mathbf{X}-\mathbf{m}(\mathbf{X}^*)\|^2] = \mathrm{tr}(\mathbb{E}[(\mathbf{X}-\mathbf{m}(\mathbf{X}^*))(\mathbf{X}-\mathbf{m}(\mathbf{X}^*))^\top]) = \mathrm{tr}(\mathbb{E}[\boldsymbol{\Sigma}_{X|X^*}])$. Dropping the non-negative third term:

$$\mathrm{tr}(\mathrm{Var}(\boldsymbol{\eta}(\mathbf{X}))) \le \mathrm{tr}(\mathbb{E}[\boldsymbol{\Sigma}_{X|X^*}]). \tag{7}$$

Thus, the variance of the double-projection residual is bounded by the calibration residual variance.

Since $\mathbb{E}[\boldsymbol{\eta}(\mathbf{X})] = \mathbf{0}$, Term A becomes a covariance:

$$\text{Term A} = \mathrm{Cov}\big(\mathbb{E}[\Delta \mid \mathbf{X}],\ \boldsymbol{\eta}(\mathbf{X})\big). \tag{8}$$

We bound Terms A and C by decomposing each into a $\boldsymbol{\beta}$-independent censoring contribution and a $\boldsymbol{\beta}$-dependent hazard contribution.

Step 2b: Decomposing Term A. Under the Cox model with censoring survival $G(t \mid \mathbf{X}) = \mathbb{P}(C \ge t \mid \mathbf{X})$,

$$\mathbb{E}[\Delta \mid \mathbf{X}] = \int_0^\tau \lambda_0(t) \exp(\boldsymbol{\beta}^\top \mathbf{X})\, S_0(t)^{\exp(\boldsymbol{\beta}^\top \mathbf{X})}\, G(t \mid \mathbf{X})\, dt.$$

When censoring depends on covariates, $\mathbb{E}[\Delta \mid \mathbf{X}]$ depends on $\mathbf{X}$ through both the event hazard ($\boldsymbol{\beta}$-dependent) and the censoring mechanism ($\boldsymbol{\beta}$-independent). Write

$$\mathbb{E}[\Delta \mid \mathbf{X}] = \underbrace{h(\mathbf{X})}_{\text{censoring contribution}} + \underbrace{g(\mathbf{X}, \boldsymbol{\beta})}_{\text{hazard contribution}},$$

where $h(\mathbf{X}) := \mathbb{E}[\Delta \mid \mathbf{X}]\big|_{\boldsymbol{\beta}=\mathbf{0}} = \int_0^\tau \lambda_0(t)\, S_0(t)\, G(t \mid \mathbf{X})\, dt$ is the event probability under the null model (which depends on $\mathbf{X}$ only through censoring), and $g(\mathbf{X}, \boldsymbol{\beta}) = \mathbb{E}[\Delta \mid \mathbf{X}] - h(\mathbf{X})$ captures the effect of the covariates on the event hazard. By construction, $g(\mathbf{X}, \mathbf{0}) = 0$. Moreover $\mathbb{E}[\Delta \mid \mathbf{X}]$ is smooth in $\boldsymbol{\beta}$ with gradient $\nabla_{\boldsymbol{\beta}} \mathbb{E}[\Delta \mid \mathbf{X}]\big|_{\boldsymbol{\beta}=\mathbf{0}} = \mathbf{X}\, \phi(\mathbf{X})$, where $\phi(\mathbf{X}) = \int_0^\tau \lambda_0(t)\, S_0(t) \big(1 + \log S_0(t)\big) G(t \mid \mathbf{X})\, dt$ is bounded on $[0, \tau]$ because the survival-damped integrand $u\, S_0(t)^u$ is bounded over $u \ge 0$ and $\log S_0(t)$ is bounded by (A1)(iii). Hence, by a first-order expansion $g(\mathbf{X}, \boldsymbol{\beta}) = O(\|\boldsymbol{\beta}\|)$ in L^2 and $\mathrm{Var}(g(\mathbf{X}, \boldsymbol{\beta})) = O(\|\boldsymbol{\beta}\|^2)$, with the constant governed by $\mathbb{E}\|\mathbf{X}\|^2 < \infty$.

Substituting into (8):

$$\text{Term A} = \underbrace{\mathrm{Cov}(h(\mathbf{X}),\, \boldsymbol{\eta}(\mathbf{X}))}_{\mathrm{A}_0\text{: censoring–calibration}} + \underbrace{\mathrm{Cov}(g(\mathbf{X}, \boldsymbol{\beta}),\, \boldsymbol{\eta}(\mathbf{X}))}_{\mathrm{A}_\beta\text{: hazard–calibration}}.$$

The hazard–calibration term A_β is $O(\|\boldsymbol{\beta}\| \cdot \|\mathbb{E}[\boldsymbol{\Sigma}_{X|X^*}]\|^{1/2})$ by Cauchy–Schwarz and (7), since $\sqrt{\mathrm{Var}(g)} = O(\|\boldsymbol{\beta}\|)$ and $\sqrt{\mathrm{tr}(\mathrm{Var}(\boldsymbol{\eta}))} \le \sqrt{\mathrm{tr}(\mathbb{E}[\boldsymbol{\Sigma}_{X|X^*}])}$. The censoring–calibration term $\mathrm{A}_0 = \mathrm{Cov}(h(\mathbf{X}), \boldsymbol{\eta}(\mathbf{X}))$ does not involve $\boldsymbol{\beta}$ and is generally nonzero when censoring depends on covariates.

Step 2c: Decomposing Term C. A parallel decomposition applies to Term C. The risk-set weighted averages are:

$$\bar{\mathbf{X}}(\boldsymbol{\beta},t)=\frac{s_X^{(1)}(\boldsymbol{\beta},t)}{s_X^{(0)}(\boldsymbol{\beta},t)}, \qquad s_X^{(k)}(\boldsymbol{\beta},t)=\mathbb{E}[\mathbf{X}^{\otimes k}\exp(\boldsymbol{\beta}^\top\mathbf{X})\mathbf{1}(T\geq t)],$$
$$\bar{\mathbf{m}}(\boldsymbol{\beta},t)=\frac{s_m^{(1)}(\boldsymbol{\beta},t)}{s_m^{(0)}(\boldsymbol{\beta},t)}, \qquad s_m^{(k)}(\boldsymbol{\beta},t)=\mathbb{E}[\mathbf{m}(\mathbf{X}^*)^{\otimes k}\exp(\boldsymbol{\beta}^\top\mathbf{m}(\mathbf{X}^*))\mathbf{1}(T\geq t)],$$

and we define the risk-set discrepancy $\mathbf{D}(\boldsymbol{\beta},t):=\bar{\mathbf{X}}(\boldsymbol{\beta},t)-\bar{\mathbf{m}}(\boldsymbol{\beta},t)$, so

$$\text{Term C}=\mathbb{E}\big[\Delta\,\mathbf{D}(\boldsymbol{\beta},T)\big]. \tag{9}$$

Recall the calibration residual $\boldsymbol{\epsilon}_m=\mathbf{X}-\mathbf{m}(\mathbf{X}^*)$ from Step 2a, so that $\mathbf{m}(\mathbf{X}^*)=\mathbf{X}-\boldsymbol{\epsilon}_m$ and by a Taylor expansion:

$$\exp(\boldsymbol{\beta}^\top\mathbf{m}(\mathbf{X}^*))=\exp(\boldsymbol{\beta}^\top\mathbf{X})\cdot\exp(-\boldsymbol{\beta}^\top\boldsymbol{\epsilon}_m)=\exp(\boldsymbol{\beta}^\top\mathbf{X})\big[1-\boldsymbol{\beta}^\top\boldsymbol{\epsilon}_m\big]+O(\|\boldsymbol{\beta}\|^2\|\boldsymbol{\epsilon}_m\|^2).$$

Substituting into $s_m^{(0)}$ and $s_m^{(1)}$, and using $\boldsymbol{\epsilon}_m\perp T\mid\mathbf{X}$ from (A2) together with $\mathbb{E}[\boldsymbol{\epsilon}_m\mid\mathbf{X}]=-\boldsymbol{\eta}(\mathbf{X})$:

$$s_m^{(0)}(\boldsymbol{\beta},t)=s_X^{(0)}(\boldsymbol{\beta},t)+\mathbb{E}\big[\exp(\boldsymbol{\beta}^\top\mathbf{X})\,\boldsymbol{\beta}^\top\boldsymbol{\eta}(\mathbf{X})\,\mathbf{1}(T\geq t)\big]+R_0(\boldsymbol{\beta},t), \tag{10}$$
$$\begin{aligned}s_m^{(1)}(\boldsymbol{\beta},t)&=\mathbb{E}\big[(\mathbf{X}-\boldsymbol{\epsilon}_m)\exp(\boldsymbol{\beta}^\top\mathbf{X})(1-\boldsymbol{\beta}^\top\boldsymbol{\epsilon}_m)\mathbf{1}(T\geq t)\big]+R_1(\boldsymbol{\beta},t)\\&=s_X^{(1)}(\boldsymbol{\beta},t)+\mathbb{E}\big[\mathbf{X}\exp(\boldsymbol{\beta}^\top\mathbf{X})\,\boldsymbol{\beta}^\top\boldsymbol{\eta}(\mathbf{X})\,\mathbf{1}(T\geq t)\big]\\&\quad+\mathbb{E}\big[\boldsymbol{\eta}(\mathbf{X})\exp(\boldsymbol{\beta}^\top\mathbf{X})\,\mathbf{1}(T\geq t)\big]+R_1(\boldsymbol{\beta},t),\end{aligned} \tag{11}$$

where $\|R_0\|,\|R_1\|=O(\|\boldsymbol{\beta}\|^2\,\|\mathbb{E}[\boldsymbol{\Sigma}_{X|X^*}]\|)$.

Write $s_m^{(k)}=s_X^{(k)}+\Delta_k$, so that from (10)–(11):

$$\Delta_0(\boldsymbol{\beta},t)=\mathbb{E}\big[\exp(\boldsymbol{\beta}^\top\mathbf{X})\,\boldsymbol{\beta}^\top\boldsymbol{\eta}(\mathbf{X})\,\mathbf{1}(T\geq t)\big]+R_0(\boldsymbol{\beta},t), \tag{12}$$
$$\Delta_1(\boldsymbol{\beta},t)=\underbrace{\mathbb{E}\big[\boldsymbol{\eta}(\mathbf{X})\exp(\boldsymbol{\beta}^\top\mathbf{X})\,\mathbf{1}(T\geq t)\big]}_{=:\,\Delta_1^{(\mathrm{a})}(\boldsymbol{\beta},t)}+\underbrace{\mathbb{E}\big[\mathbf{X}\exp(\boldsymbol{\beta}^\top\mathbf{X})\,\boldsymbol{\beta}^\top\boldsymbol{\eta}(\mathbf{X})\,\mathbf{1}(T\geq t)\big]}_{=:\,\Delta_1^{(\mathrm{b})}(\boldsymbol{\beta},t)}+R_1(\boldsymbol{\beta},t). \tag{13}$$

Note that $\|\mathbb{E}[\boldsymbol{\eta}(\mathbf{X})\,\omega(\mathbf{X},t)]\|\leq\mathbb{E}[\|\boldsymbol{\eta}(\mathbf{X})\|\cdot|\omega(\mathbf{X},t)|]$ is bounded via Cauchy–Schwarz by $\sqrt{\mathbb{E}[\|\boldsymbol{\eta}\|^2]}\cdot\sqrt{\mathbb{E}[\omega^2]}$, and $\sqrt{\mathbb{E}[\|\boldsymbol{\eta}(\mathbf{X})\|^2]}=O(\|\mathbb{E}[\boldsymbol{\Sigma}_{X|X^*}]\|^{1/2})$ by (7). Applying this:

(i) Δ_0 *carries an explicit* $\boldsymbol{\beta}^\top\boldsymbol{\eta}$ *factor.* From (12), $\Delta_0(\boldsymbol{\beta},t)=\mathbb{E}[\exp(\boldsymbol{\beta}^\top\mathbf{X})\,\boldsymbol{\beta}^\top\boldsymbol{\eta}(\mathbf{X})\,\mathbf{1}(T\geq t)]+R_0$. The leading term is bounded by $\|\boldsymbol{\beta}\|\cdot\mathbb{E}[\exp(\boldsymbol{\beta}^\top\mathbf{X})\,\|\boldsymbol{\eta}(\mathbf{X})\|\,\mathbf{1}(T\geq t)]\leq\|\boldsymbol{\beta}\|\cdot\sqrt{\mathbb{E}[\exp(2\boldsymbol{\beta}^\top\mathbf{X})\mathbf{1}(T\geq t)]}\cdot\sqrt{\mathbb{E}[\|\boldsymbol{\eta}\|^2]}$ by Cauchy–Schwarz, and $R_0=O(\|\boldsymbol{\beta}\|^2\,\|\mathbb{E}[\boldsymbol{\Sigma}_{X|X^*}]\|)$ is strictly subdominant. Therefore

$$\|\Delta_0(\boldsymbol{\beta},t)\|=O\big(\|\boldsymbol{\beta}\|\cdot\|\mathbb{E}[\boldsymbol{\Sigma}_{X|X^*}]\|^{1/2}\big), \tag{14}$$

and in particular $\Delta_0(\mathbf{0},t)=0$.

(ii) Δ_1 *has a nonzero* $\boldsymbol{\beta}$*-independent piece.* From (13), $\Delta_1^{(\mathrm{a})}(\mathbf{0},t) = \mathbb{E}[\boldsymbol{\eta}(\mathbf{X})\mathbf{1}(T \geq t)] \neq \mathbf{0}$ in general, with $\|\Delta_1^{(\mathrm{a})}(\mathbf{0},t)\| = O(\|\mathbb{E}[\boldsymbol{\Sigma}_{X|X^*}]\|^{1/2})$ by Cauchy–Schwarz and (7). The same bound holds for $\Delta_1^{(\mathrm{a})}(\boldsymbol{\beta},t)$ at any $\boldsymbol{\beta}$ (using boundedness of $\exp(\boldsymbol{\beta}^\top \mathbf{X})$ on the support). The piece $\Delta_1^{(\mathrm{b})}$ carries an explicit $\boldsymbol{\beta}^\top$ factor, so $\|\Delta_1^{(\mathrm{b})}(\boldsymbol{\beta},t)\| = O(\|\boldsymbol{\beta}\| \cdot \|\mathbb{E}[\boldsymbol{\Sigma}_{X|X^*}]\|^{1/2})$. Combining:

$$\|\Delta_1(\boldsymbol{\beta},t)\| = O\big(\|\mathbb{E}[\boldsymbol{\Sigma}_{X|X^*}]\|^{1/2}\big), \qquad \|\Delta_1(\boldsymbol{\beta},t) - \Delta_1(\mathbf{0},t)\| = O\big(\|\boldsymbol{\beta}\| \cdot \|\mathbb{E}[\boldsymbol{\Sigma}_{X|X^*}]\|^{1/2}\big). \tag{15}$$

At $\boldsymbol{\beta} = \mathbf{0}$, $\Delta_0(\mathbf{0},t) = 0$ but $\Delta_1(\mathbf{0},t) \neq \mathbf{0}$. So the risk-set discrepancy at $\boldsymbol{\beta} = \mathbf{0}$ is

$$\mathbf{D}(\mathbf{0},t) = \bar{\mathbf{X}}(\mathbf{0},t) - \bar{\mathbf{m}}(\mathbf{0},t) = -\frac{\Delta_1(\mathbf{0},t)}{s_X^{(0)}(\mathbf{0},t)} = -\frac{\mathbb{E}[\boldsymbol{\eta}(\mathbf{X})\mathbf{1}(T \geq t)]}{\mathbb{P}(T \geq t)}, \tag{16}$$

which is $\boldsymbol{\beta}$-independent, nonzero in general, and bounded by $\|\mathbf{D}(\mathbf{0},t)\| = O(\|\mathbb{E}[\boldsymbol{\Sigma}_{X|X^*}]\|^{1/2})$ via Cauchy–Schwarz and (7).

Take $\bar{\mathbf{m}}(\boldsymbol{\beta},t) = \dfrac{s_X^{(1)} + \Delta_1}{s_X^{(0)} + \Delta_0}$. The ratio expansion uses $1/(a+x) = 1/a - x/a^2 + x^2/a^3 + O(x^3)$ applied to $1/(s_X^{(0)}(\boldsymbol{\beta},t) + \Delta_0(\boldsymbol{\beta},t))$:

$$\begin{aligned}
\bar{\mathbf{m}}(\boldsymbol{\beta},t) &= \big(s_X^{(1)} + \Delta_1\big) \cdot \left[\frac{1}{s_X^{(0)}} - \frac{\Delta_0}{(s_X^{(0)})^2} + O\big(\Delta_0^2/(s_X^{(0)})^3\big)\right] \\
&= \bar{\mathbf{X}}(\boldsymbol{\beta},t) + \frac{\Delta_1(\boldsymbol{\beta},t) - \bar{\mathbf{X}}(\boldsymbol{\beta},t)\,\Delta_0(\boldsymbol{\beta},t)}{s_X^{(0)}(\boldsymbol{\beta},t)} \underbrace{- \frac{\Delta_1(\boldsymbol{\beta},t)\,\Delta_0(\boldsymbol{\beta},t)}{(s_X^{(0)}(\boldsymbol{\beta},t))^2} + O\big(\|\Delta_0\|^2\big)}_{\text{second-order remainder}},
\end{aligned}$$

so

$$\mathbf{D}(\boldsymbol{\beta},t) = -\frac{\Delta_1(\boldsymbol{\beta},t) - \bar{\mathbf{X}}(\boldsymbol{\beta},t)\,\Delta_0(\boldsymbol{\beta},t)}{s_X^{(0)}(\boldsymbol{\beta},t)} + \frac{\Delta_1(\boldsymbol{\beta},t)\,\Delta_0(\boldsymbol{\beta},t)}{(s_X^{(0)}(\boldsymbol{\beta},t))^2} + O(\|\Delta_0\|^2).$$

The second-order remainder contains a $\Delta_1\Delta_0$ cross term and a Δ_0^2 piece. Using (14) and (15):

$$\begin{aligned}
\|\Delta_1\Delta_0/(s_X^{(0)})^2\| &= O\big(\|\boldsymbol{\beta}\| \cdot \|\mathbb{E}[\boldsymbol{\Sigma}_{X|X^*}]\|\big), \\
\|\Delta_0^2/(s_X^{(0)})^3\| &= O\big(\|\boldsymbol{\beta}\|^2 \cdot \|\mathbb{E}[\boldsymbol{\Sigma}_{X|X^*}]\|\big).
\end{aligned}$$

Both are subdominant to the leading-order bound $O(\|\boldsymbol{\beta}\| \cdot \|\mathbb{E}[\boldsymbol{\Sigma}_{X|X^*}]\|^{1/2})$ established below.

Subtracting $\mathbf{D}(\mathbf{0},t) = -\Delta_1(\mathbf{0},t)/\mathbb{P}(T \geq t)$ from $\mathbf{D}(\boldsymbol{\beta},t)$:

$$\begin{aligned}
\mathbf{D}(\boldsymbol{\beta},t) - \mathbf{D}(\mathbf{0},t) = \underbrace{\Delta_1(\mathbf{0},t)\left[\frac{1}{\mathbb{P}(T \geq t)} - \frac{1}{s_X^{(0)}(\boldsymbol{\beta},t)}\right]}_{\text{(I)}} \\
\underbrace{- \frac{\Delta_1(\boldsymbol{\beta},t) - \Delta_1(\mathbf{0},t)}{s_X^{(0)}(\boldsymbol{\beta},t)}}_{\text{(II)}} + \underbrace{\frac{\bar{\mathbf{X}}(\boldsymbol{\beta},t)\,\Delta_0(\boldsymbol{\beta},t)}{s_X^{(0)}(\boldsymbol{\beta},t)}}_{\text{(III)}} + (\text{subdominant}).
\end{aligned} \tag{17}$$

We bound each piece. For (I): $s_X^{(0)}(\boldsymbol{\beta}, t) - \mathbb{P}(T \geq t) = \mathbb{E}[(\exp(\boldsymbol{\beta}^\top \mathbf{X}) - 1)\mathbf{1}(T \geq t)] = O(\|\boldsymbol{\beta}\|)$ by Taylor expansion, so the bracket is $O(\|\boldsymbol{\beta}\|)$; combined with $\|\Delta_1(\mathbf{0}, t)\| = O(\|\mathbb{E}[\boldsymbol{\Sigma}_{X|X^*}]\|^{1/2})$ from (15), this gives $\|(\mathrm{I})\| = O(\|\boldsymbol{\beta}\| \cdot \|\mathbb{E}[\boldsymbol{\Sigma}_{X|X^*}]\|^{1/2})$.

For (II): the second bound in (15) gives directly $\|\Delta_1(\boldsymbol{\beta}, t) - \Delta_1(\mathbf{0}, t)\| = O(\|\boldsymbol{\beta}\| \cdot \|\mathbb{E}[\boldsymbol{\Sigma}_{X|X^*}]\|^{1/2})$, and $s_X^{(0)}(\boldsymbol{\beta}, t)$ is bounded below by regularity, so $\|(\mathrm{II})\| = O(\|\boldsymbol{\beta}\| \cdot \|\mathbb{E}[\boldsymbol{\Sigma}_{X|X^*}]\|^{1/2})$.

For (III): (14) gives $\|\Delta_0(\boldsymbol{\beta}, t)\| = O(\|\boldsymbol{\beta}\| \cdot \|\mathbb{E}[\boldsymbol{\Sigma}_{X|X^*}]\|^{1/2})$, and $\|\bar{\mathbf{X}}(\boldsymbol{\beta}, t)\|$ is bounded on the support, so $\|(\mathrm{III})\| = O(\|\boldsymbol{\beta}\| \cdot \|\mathbb{E}[\boldsymbol{\Sigma}_{X|X^*}]\|^{1/2})$.

Combining:

$$\mathbf{D}(\boldsymbol{\beta}, t) = \mathbf{D}(\mathbf{0}, t) + \boldsymbol{\rho}(\boldsymbol{\beta}, t), \qquad \|\boldsymbol{\rho}(\boldsymbol{\beta}, t)\| = O\big(\|\boldsymbol{\beta}\| \cdot \|\mathbb{E}[\boldsymbol{\Sigma}_{X|X^*}]\|^{1/2}\big), \tag{18}$$

so $\mathbf{D}(\boldsymbol{\beta}, t)$ splits cleanly into a $\boldsymbol{\beta}$-independent piece $\mathbf{D}(\mathbf{0}, t)$ of size $O(\|\mathbb{E}[\boldsymbol{\Sigma}_{X|X^*}]\|^{1/2})$ and a $\boldsymbol{\beta}$-dependent remainder $\boldsymbol{\rho}(\boldsymbol{\beta}, t)$ of size $O(\|\boldsymbol{\beta}\| \cdot \|\mathbb{E}[\boldsymbol{\Sigma}_{X|X^*}]\|^{1/2})$.

Step 2c(ii): Decomposing Term C via the censoring/hazard split. Substituting (18) into (9) and using iterated expectations with $\mathbb{E}[\Delta \mid \mathbf{X}] = h(\mathbf{X}) + g(\mathbf{X}, \boldsymbol{\beta})$ from Step 2b:

$$\begin{aligned} \text{Term C} &= \mathbb{E}\big[\Delta\, \mathbf{D}(\mathbf{0}, T)\big] + \mathbb{E}\big[\Delta\, \boldsymbol{\rho}(\boldsymbol{\beta}, T)\big] \\ &= \underbrace{\mathbb{E}\big[h(\mathbf{X})\, \mathbf{D}(\mathbf{0}, T)\big]}_{=:\,\mathrm{C}_0} + \underbrace{\mathbb{E}\big[g(\mathbf{X}, \boldsymbol{\beta})\, \mathbf{D}(\mathbf{0}, T)\big] + \mathbb{E}\big[\Delta\, \boldsymbol{\rho}(\boldsymbol{\beta}, T)\big]}_{=:\,\mathrm{C}_\beta}. \end{aligned}$$

By construction, C_0 is $\boldsymbol{\beta}$-independent (since $\mathbf{D}(\mathbf{0}, t)$ depends only on baseline risk-set composition, not on $\boldsymbol{\beta}$). The two terms defining C_β are each $O(\|\boldsymbol{\beta}\| \cdot \|\mathbb{E}[\boldsymbol{\Sigma}_{X|X^*}]\|^{1/2})$: the first by Cauchy–Schwarz with $\sqrt{\mathrm{Var}(g(\mathbf{X}, \boldsymbol{\beta}))} = O(\|\boldsymbol{\beta}\|)$ and $\sup_t \|\mathbf{D}(\mathbf{0}, t)\|$ controlled by $\|\mathbb{E}[\boldsymbol{\Sigma}_{X|X^*}]\|^{1/2}$ via (7) applied at $\boldsymbol{\beta} = \mathbf{0}$; the second directly from (18) since $|\Delta| \leq 1$. Therefore

$$\|\mathrm{C}_\beta\| = O\big(\|\boldsymbol{\beta}\| \cdot \|\mathbb{E}[\boldsymbol{\Sigma}_{X|X^*}]\|^{1/2}\big),$$

matching the rate of A_β from Step 2b.

Step 2d: Cancellation of censoring cross-terms. At $\boldsymbol{\beta} = \mathbf{0}$, $\mathbf{u}_{\mathrm{RC}}(\mathbf{0}) = \mathrm{A}_0 + \mathrm{C}_0$ (since $\mathbf{u}(\boldsymbol{\beta}) = \mathbf{0}$ and $\mathrm{A}_\beta, \mathrm{C}_\beta$ vanish). A martingale argument shows $\mathbf{u}_{\mathrm{RC}}(\mathbf{0}) = \mathbf{0}$: at $\boldsymbol{\beta} = \mathbf{0}$, the RC score is $\mathbb{E}[\int_0^\tau \{\mathbf{m}(\mathbf{X}^*) - \bar{\mathbf{m}}(\mathbf{0}, t)\}\, dM(t)]$, where $M(t) = N(t) - \int_0^t Y(s)\lambda_0(s)\, ds$ is a martingale under the true null model. Since $\mathbf{m}(\mathbf{X}^*) - \bar{\mathbf{m}}(\mathbf{0}, t)$ is predictable, this integral has zero expectation. Therefore $\mathrm{A}_0 + \mathrm{C}_0 = \mathbf{0}$: the censoring cross-terms from Terms A and C cancel exactly.

The surviving contributions are the $\boldsymbol{\beta}$-dependent parts A_β and C_β, both $O(\|\boldsymbol{\beta}\| \cdot \|\mathbb{E}[\boldsymbol{\Sigma}_{X|X^*}]\|^{1/2})$. Therefore

$$\|\mathbf{u}_{\mathrm{RC}}(\boldsymbol{\beta})\| = \|\mathrm{A}_\beta + \mathrm{C}_\beta\| = O\big(\|\boldsymbol{\beta}\| \cdot \|\mathbb{E}[\boldsymbol{\Sigma}_{X|X^*}]\|^{1/2}\big). \tag{19}$$

Step 2e: Taylor inversion of the score. From (19):

$$\mathbf{u}_{\mathrm{RC}}(\boldsymbol{\beta}) = O(\|\boldsymbol{\beta}\| \cdot \|\mathbb{E}[\boldsymbol{\Sigma}_{X|X^*}]\|^{1/2}).$$

Let $\boldsymbol{\beta}^*_{\mathrm{RC}}$ denote the population target of $\hat{\boldsymbol{\beta}}_{\mathrm{RC}}$ (where $\mathbf{u}_{\mathrm{RC}}(\boldsymbol{\beta}^*_{\mathrm{RC}}) = \mathbf{0}$). Taylor-expanding around $\boldsymbol{\beta}$:

$$\mathbf{0} = \mathbf{u}_{\mathrm{RC}}(\boldsymbol{\beta}^*_{\mathrm{RC}}) \approx \mathbf{u}_{\mathrm{RC}}(\boldsymbol{\beta}) + \left.\frac{\partial \mathbf{u}_{\mathrm{RC}}}{\partial \boldsymbol{\beta}^\top}\right|_{\boldsymbol{\beta}} (\boldsymbol{\beta}^*_{\mathrm{RC}} - \boldsymbol{\beta}).$$

The derivative $\partial \mathbf{u}_{\mathrm{RC}}/\partial \boldsymbol{\beta}^\top|_{\boldsymbol{\beta}}$ is $-\boldsymbol{\Omega}_{\mathrm{RC}}$, the negative RC information matrix. Solving:

$$\boldsymbol{\beta}^*_{\mathrm{RC}} - \boldsymbol{\beta} = \boldsymbol{\Omega}_{\mathrm{RC}}^{-1}\, \mathbf{u}_{\mathrm{RC}}(\boldsymbol{\beta}).$$

This is the RC residual: $\|\boldsymbol{\Omega}_{\mathrm{RC}}^{-1}\mathbf{u}_{\mathrm{RC}}(\boldsymbol{\beta})\| = O(\|\boldsymbol{\beta}\| \cdot \|\mathbb{E}[\boldsymbol{\Sigma}_{X|X^*}]\|^{1/2})$.

Step 3: Combining via the equivalence. From Step 2, $\boldsymbol{\beta}^*_{\mathrm{RC}} = \boldsymbol{\beta} + \boldsymbol{\Omega}_{\mathrm{RC}}^{-1}\mathbf{u}_{\mathrm{RC}}(\boldsymbol{\beta})$. From Step 1, $\boldsymbol{\beta}^* = \mathbf{B}^\top \boldsymbol{\beta}^*_{\mathrm{RC}}$, so

$$\boldsymbol{\beta}^* = \mathbf{B}^\top \boldsymbol{\beta} + \mathbf{B}^\top \boldsymbol{\Omega}_{\mathrm{RC}}^{-1}\mathbf{u}_{\mathrm{RC}}(\boldsymbol{\beta}).$$

Solving for $\boldsymbol{\beta}$:

$$\boldsymbol{\beta} = \boldsymbol{\beta}^* + \big[(\mathbf{B}^\top)^{-1} - \mathbf{I}\big]\boldsymbol{\beta}^* - \boldsymbol{\Omega}_{\mathrm{RC}}^{-1}\mathbf{u}_{\mathrm{RC}}(\boldsymbol{\beta}),$$

which gives (2). Dropping the RC residual gives the leading-order approximation (3). ■

Appendix B. Coverage convergence of CIs with n_v

To assess how the two tiers of confidence interval respond to the amount of validation data, we reran the cross-dependent error design of Study 1 while varying the validation-sample size n_v over the grid $\{50, 100, 200, 300, 400, 800, 1600, 3200\}$ and holding the study-sample size fixed at $n_{\text{study}} = 1500$. The data-generating process is otherwise identical. We report empirical coverage across methods and n_v size in Figure 2.

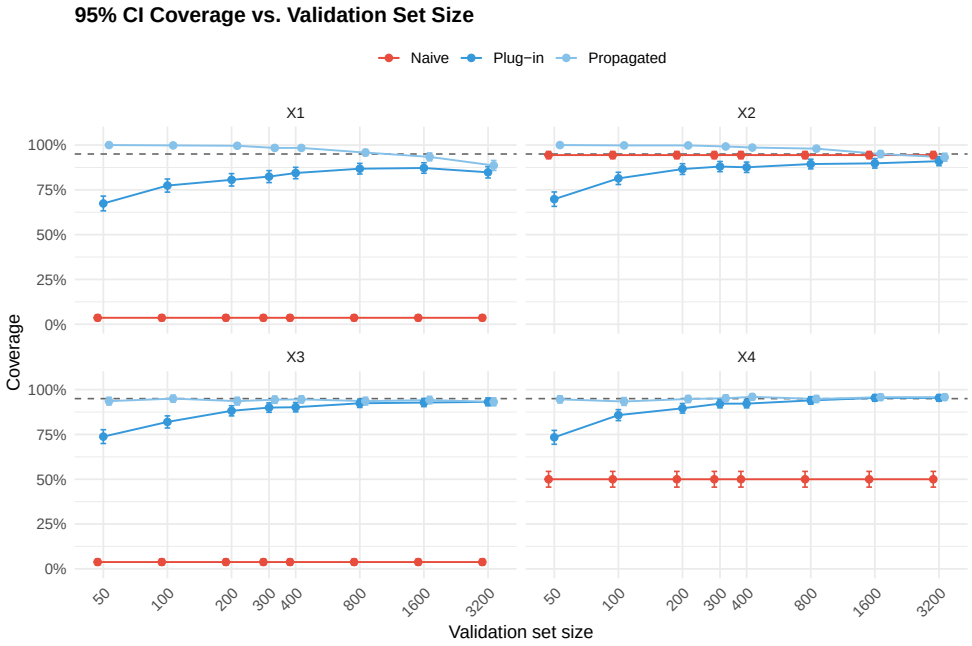

Figure 2: 95% CI coverage vs n_v for Study 1 data-generating process; error bars are 95% Monte Carlo CIs.

As expected, naive coverage is invariant to n_v, since it never uses the validation data: it remains badly miscalibrated for the covariates with appreciable attenuation and is near nominal for X_2 only because its naive bias is small. The plug-in interval undercovers when n_v is small because it ignores the sampling variability of $\hat{\mathbf{B}}$, which is largest for small validation sets. Its coverage rises monotonically toward the nominal level as n_v grows.

The propagated intervals attain near-nominal coverage across the entire grid. At small n_v it is conservative: for continuous covariates it covers essentially 100% of the time at $n_v = 50$ because the variance inflation that accounts for $\hat{\mathbf{B}}$ is largest exactly when the validation set is small; as n_v increases this inflation shrinks and propagated coverage descends toward the nominal 95%. Consequently the gap between the plug-in and propagated intervals is widest at small n_v and narrows as the two converge, quantifying the practical value of propagating calibration uncertainty precisely in the small n_v regime. For the binary covariates the propagated interval tracks the nominal level closely throughout, remaining between roughly 93% and 96% for both X_3 and X_4.

Appendix C. Semi-synthetic example using MIMIC-IV dataset

Here, we analyze data derived from the Medical Information Mart for Intensive Care (MIMIC)-IV version 3.1 (Johnson et al., 2024, 2023; Goldberger et al., 2000) on patients admitted to an intensive care unit. We obtain baseline clinical factors and time to inpatient mortality, censoring surviving patients at the end of their stay. We aim to fit a Cox proportional hazards model with covariates contaminated by measurement error, and demonstrate the realistic use and performance of our bias correction. The covariates analyzed are age at admission, sex, Sequential Organ Failure Assessment (SOFA) score, history of diabetes, and history of chronic obstructive pulmonary disease (COPD). We leave age and sex as error-free measures, and simulate moderate error among the remaining covariates. The SOFA score receives independent zero-mean additive Gaussian noise with standard deviation $\sigma_{\text{noise}} = 0.7\,\sigma_{\text{SOFA}}$; the corrupted values are rounded and clipped to the non-negative integers to respect SOFA's support. Diabetes and COPD status (binary variables) are corrupted by correlated bit-flips sharing a common latent factor $Z \sim \mathcal{N}(0,1)$: each has flip probability $p_k = \text{logit}^{-1}(\alpha_k + \gamma_k|Z|)$, calibrated to a marginal flip rate of $\sim 15\%$.

We randomly sample a study dataset of $n_{\text{study}} = 5000$ and a disjoint validation dataset with $n_v = 500$ used to compute error summaries. The validation set yields marginal error metrics visualized in Figure 3, and the calibration slope matrix used for bias correction:

$$\hat{\mathbf{B}} = \begin{array}{c} \\ \text{age} \\ \text{female} \\ \text{sofa} \\ \text{diab} \\ \text{copd} \end{array} \begin{array}{c} \begin{array}{ccccc} \text{age} & \text{female} & \text{sofa} & \text{diab} & \text{copd} \end{array} \\ \begin{pmatrix} 1.0000 & 0.0000 & 0.0000 & 0.0000 & 0.0000 \\ 0.0000 & 1.0000 & 0.0000 & 0.0000 & 0.0000 \\ -0.0007 & -0.1204 & 0.7430 & 0.1215 & 0.4018 \\ 0.0033 & 0.0140 & 0.0026 & 0.6351 & 0.0360 \\ 0.0036 & 0.0793 & -0.0009 & 0.0001 & 0.5470 \end{pmatrix} \end{array}$$

While age and sex are error-free, their correlations with other covariates impact the resulting bias correction. The condition number of $\hat{\mathbf{B}}$ is 2.28, indicating that the matrix inversion and bias correction should be stable.

The results of our analysis are visualized in Figure 4. We compute the ground truth model parameters by fitting a Cox model on error-free covariates in study data. The naive model fit results in moderate bias for HR estimates, most critically for SOFA score, where the naive confidence interval does not cover the ground truth parameter. The corrected parameters uniformly reduce bias relative to the ground truth. With the exception of SOFA score, the widths of plug-in confidence intervals are similar to those of the propagated CIs, indicating that the variance elements for these parameters are reasonably estimated from the validation set.

We also report the sensitivity diagnostic ρ for each covariate. We observe that the sex and COPD coefficients are least susceptible to bias due to the RC residual term, while those for age, SOFA score, and diabetes may be more sensitive. While our ground truth comparisons in this example show reasonably low bias among all corrected estimates, in practice, a large ρ value could indicate unreliability for larger extraction error settings.

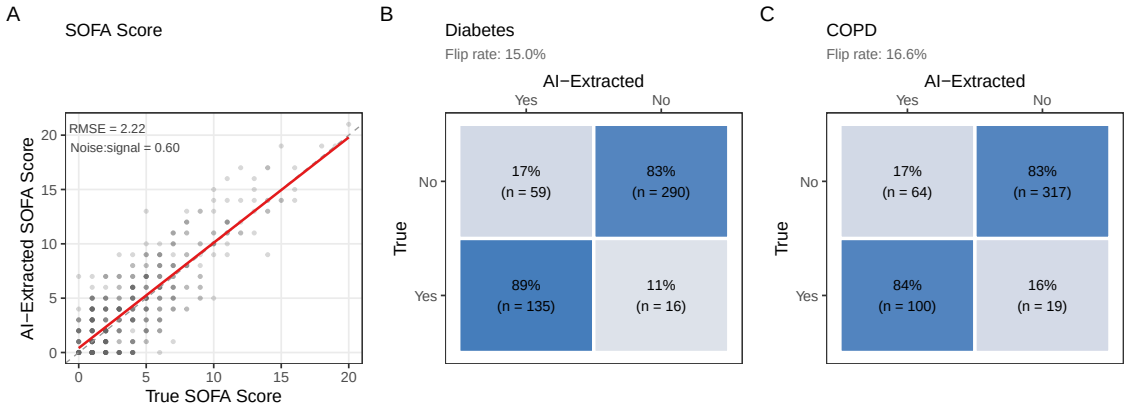

Figure 3: Marginal measurement error distributions for covariates in MIMIC-IV data experiment.

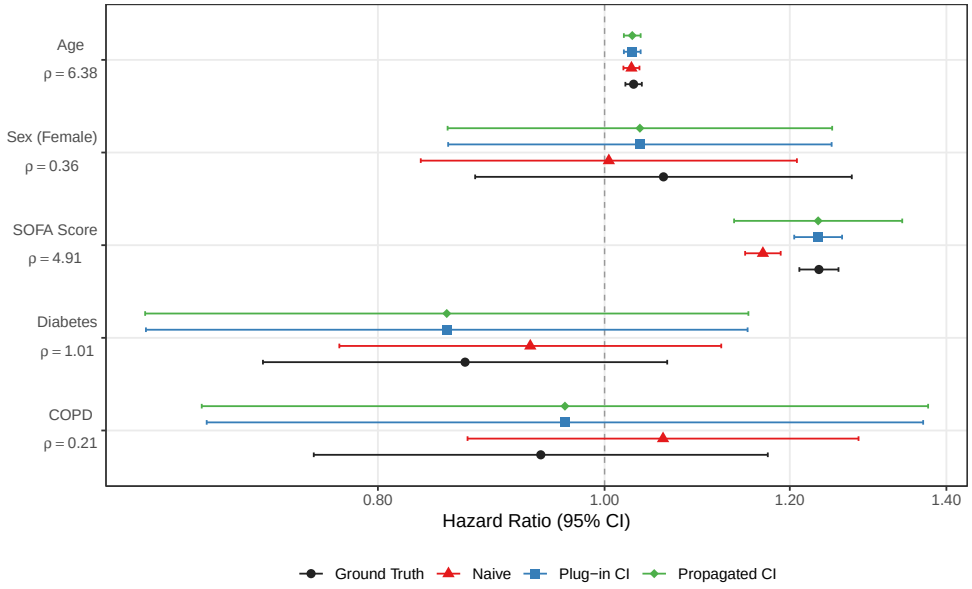

Figure 4: Estimated naive and bias-corrected hazard ratios with 95% confidence intervals in MIMIC-IV data experiment.