

Urinary metabolomics for improved prediction of overall survival beyond tumor stage in colorectal cancer: results from the ColoCare Study

Tengda Lin

Huntsman Cancer Institute

Boyi Guo

University of Utah

Victoria M Bandera

Huntsman Cancer Institute

David B Liesenfeld

German Cancer Research Center

Jincheng Shen

University of Utah

Benjamin Haaland

University of Utah

Paul A Stewart

Huntsman Cancer Institute

Kenneth M Boucher

Huntsman Cancer Institute

Patricia A Erickson

Huntsman Cancer Institute

Sheetal Hardikar

Huntsman Cancer Institute

Victoria Damerell

National Center for Tumor Diseases (NCT), German Cancer Research Center (DKFZ)

Doratha A Byrd

H. Lee Moffitt Cancer Center and Research Institute

Jane C Figueiredo

Cedars-Sinai Medical Center, Samuel Oschin Comprehensive Cancer Institute

Adetunji T Toriola

Washington University in St. Louis

David Shibata

University of Tennessee Health Science Center

Erin M Siegel

H. Lee Moffitt Cancer Center and Research Institute

Christopher I Li

Fred Hutchinson Cancer Center

Alexis B Ulrich

Lukaskrankenhaus

Christoph Kahlert

University Hospital Heidelberg

Daniel O Scharfstein

University of Utah

Biljana Gigic

National Center for Tumor Diseases (NCT), German Cancer Research Center (DKFZ)

Cornelia M Ulrich

Huntsman Cancer Institute

Jennifer Ose

`jennifer.ose@hs-hannover.de`

Hochschule Hannover

Research Article

Keywords:

Posted Date: June 1st, 2026

DOI: <https://doi.org/10.21203/rs.3.rs-9737689/v1>

License:   This work is licensed under a Creative Commons Attribution 4.0 International License.

[Read Full License](#)

Additional Declarations: Competing interest reported. D.L. is currently an employee of Roche (Roche Pharma AG, Emil-Barell-Str. 1, 79639 Grenzach-Wyhlen, Germany). This work was conducted while D.L. was affiliated with NCT/DKFZ and is unrelated to his current employment. C.M.U., as cancer center director, oversees research funded by several pharmaceutical companies but has not received direct personal funding.

Version of Record: A version of this preprint was published at Metabolomics on September 11th, 2026. See the published version at <https://doi.org/10.1007/s11306-026-02516-3>.

Urinary metabolomics for improved prediction of overall survival beyond tumor stage in colorectal cancer: results from the ColoCare Study

Tengda Lin^{1,2}, Boyi Guo², Victoria M Bandera^{1,2}, David B Liesenfeld³, Jincheng Shen⁴, Benjamin Haaland², Paul A Stewart^{1,5}, Kenneth M Boucher¹, Patricia A Erickson^{1,2}, Sheetal Hardikar^{1,2}, Victoria Damerell⁶, Doratha A Byrd⁷, Jane C Figueiredo⁸, Adetunji T Toriola⁹, David Shibata¹⁰, Erin M Siegel⁷, Christopher I Li¹¹, Alexis B Ulrich¹², Christoph Kahlert⁶, Daniel O Scharfstein^{2,*}, Biljana Gigic^{6,*}, Cornelia M Ulrich^{1,2,*}, Jennifer Ose^{13,*}

¹Huntsman Cancer Institute, University of Utah, Salt Lake City, UT, United States.

²Department of Population Health Sciences, University of Utah, Salt Lake City, UT, United States.

³Division of Preventive Oncology, National Center for Tumor Diseases (NCT), German Cancer Research Center (DKFZ), Heidelberg, Germany.

⁴Department of Internal Medicine, University of Utah, Salt Lake City, UT, United States.

⁵Department of Nutrition and Integrative Physiology, University of Utah, Salt Lake City, UT, United States.

⁶Department of General, Visceral, and Transplantation Surgery, Heidelberg University Hospital, Heidelberg, Germany.

⁷H. Lee Moffitt Cancer Center and Research Institute, Tampa, FL, United States.

⁸Department of Medicine, Cedars-Sinai Medical Center, Samuel Oschin Comprehensive Cancer Institute, Los Angeles, CA, United States.

⁹Washington University School of Medicine in St. Louis, St. Louis, MO, United States.

¹⁰Department of Surgery, University of Tennessee Health Science Center, Memphis, TN, United States.

¹¹Fred Hutchinson Cancer Center, Seattle, WA, United States.

¹²Lukaskrankenhaus Neuss, Neuss, Germany.

¹³Department of Information and Communication, Faculty for Media, Information and Design, University of Applied Sciences and Arts, Hannover, Germany.

* Shared last authors

1 Abstract (word count: 246)

Background: Colorectal cancer (CRC) is a leading cause of cancer-related mortality. Prognosis is primarily guided by tumor stage despite substantial molecular heterogeneity. Urinary metabolomics may capture systemic and tumor-related biology beyond staging and could improve prognostic assessment. We hypothesized that incorporating urinary metabolomic profiles would improve overall survival (OS) prediction performance compared with a stage- and age-based reference model. **Method:** A total of $n = 76$ stage I-IV CRC patients recruited as part of the ColoCare Study in Heidelberg Germany with pre-surgery urinary metabolomics were included (23 deaths; median follow-up 3.03 years). Four metabolomics-based penalized Cox models adjusted for tumor stage and age at diagnosis were developed using LASSO, adaptive LASSO, spike-and-slab LASSO, and iterative sure independence screening (iSIS)-LASSO. Model discrimination was assessed using Harrell's C-index and time-dependent AUC based on the nested cross-validation. **Results:** Compared with the reference model (Cox model including only tumor stage and age at diagnosis), all metabolomics-based models improved discrimination. The spike-and-slab LASSO Cox model demonstrated the best performance, achieving a C-index of 0.75 (vs. 0.68) and consistently higher time-dependent AUCs at 1–5 years of follow-up, with a peak AUC of 0.76 at year 3 (vs. 0.68). Three urinary metabolites were consistently selected across all metabolomics-based models: indolelactate, 2-hydroxyisobutyrate and a uridine-like metabolite. **Conclusions:** Urinary metabolomic profiles improved CRC OS prediction beyond tumor stage and age at diagnosis, especially with the spike-and-slab LASSO Cox model. These results support urinary metabolomics as a promising noninvasive prognostic tool that merits external validation.

2 Introduction

Colorectal cancer (CRC) is the third most common type of cancer and the second leading cause of cancer-related mortality worldwide (World Health Organization, 2024), with over 1.9 million annual cases and 0.9 million deaths per year reported in 2022 (International Agency for Research on Cancer, 2024). In the US, CRC is also the second leading cause of cancer death, with a 5-year relative survival rate of 65% (Siegel, Giaquinto, & Jemal, 2026; Siegel, Miller, et al., 2026; Society, 2024). Prognosis is primarily based on tumor stage at diagnosis, despite the heterogeneity molecular characteristics and treatment response (Koncina et al., 2020; Lawler et al., 2018; Rosati et al., 2022). Consequently, stage-based decision-making may not adequately support individualized risk stratification and can lead to over-treatment, with associated patient and economic burdens (Kneuert et al., 2015; Scheele et al., 2018). More accurate prognostic tools are needed to enable personalized interventions.

Metabolomics offers a promising avenue for addressing these gaps, which provides a unique window into the functional state of biological systems by capturing the end products of cellular regulatory processes (Fiehn, 2002; Griffin & Shockcor, 2004). This makes metabolomics particularly valuable in cancer research, where metabolic dysregulation is a disease hallmark (Schmidt et al., 2021). However, prognostic application remains limited (Erben et al., 2018; Gold et al., 2022; Zhang et al., 2017). Urine-based metabolomics is attractive because its noninvasive collection and minimal processing offer accessible patient prognostic stratification (Bouatra et al., 2013; Brezmes et al., 2022). Urinary CRC biomarkers have been identified, including volatile organic compounds (Widlak et al., 2018), modified cytosine nucleosides (Guo et al., 2018), and polyamines (Nakajima et al., 2018; Venäläinen et al., 2018). Although primarily investigated for

diagnosis, these biomarkers reflect metabolic alterations and suggest prognostic potential.

In CRC metabolomics, prognostic studies commonly use one of two strategies: (i) univariate screening, where metabolites are tested individually with multiple-testing correction and the significant metabolites are used to build survival models; or (ii) multivariate dimension reduction, typically partial least squares (PLS)–based, where latent components are derived, their corresponding component scores are used as predictors in Cox models to generate risk scores, and patients are stratified for Kaplan–Meier analysis (Bertini et al., 2012; Farshidfar et al., 2016; Jiménez et al., 2013; Liesenfeld et al., 2015; Ose et al., 2023; Phua et al., 2014; Qiu et al., 2014; Zaimenko et al., 2019). One major drawback of the univariate approach for feature selection is that it considers only marginal associations of individual features. Consequently, features with strong joint associations but weak marginal associations may be excluded (Fan & Lv, 2008). The PLS and orthogonal PLS (OPLS) methods are favored in chemometrics and adopted by metabolomics for dimension reduction (Trygg & Wold, 2002; Trygg et al., 2007; Wold, 1975). However, standard PLS uses all predictors to form latent components and does not yield sparse solutions, so all metabolites typically have nonzero loadings, complicating the interpretation of individual metabolite contributions. Although OPLS improves interpretability by separating predictive and orthogonal variation and has been extended into several variants (Magnanensi et al., 2021; Mehmood et al., 2012, 2020; Pinto, 2017), they still lack explicit variable selection. These limitations have led to increasing use of penalized Cox models for identifying prognostic metabolite signatures (Seum et al., 2025; Y. Xu et al., 2022). However, penalized methods fit as a single model can be prone to model-selection instability. Because metabolites are highly correlated and true signals are likely sparse, the selected metabolite set can vary substantially with tuning parameters, random resampling/splits, or across datasets, limiting the selection stability and

potentially hindering reproducibility and external validation in independent cohorts. (Jardillier et al., 2022; Waldmann et al., 2013).

To address these limitations, we applied a multi-model framework using four complementary high-dimensional variable-selection methods: least absolute shrinkage and selection operator (LASSO) (Simon et al., 2011; Tibshirani, 1996), adaptive LASSO (Zou, 2006), spike-and-slab LASSO (Rocková & George, 2014; Tang et al., 2017; Yi et al., 2019), and iterative sure independence screening (iSIS)-LASSO (Fan & Lv, 2008; Fan et al., 2009, 2010), to identify robust prognostic metabolite signatures from a prospective CRC cohort. By adjusting all models for key clinical covariates (tumor stage and age at diagnosis), we aim to identify prognostic information contributed by metabolomics beyond these established risk factors.

3 Materials and Methods

3.1 *Study population*

We included patients recruited as part of the ColoCare Study in Heidelberg, Germany with available pre-surgery urinary metabolomics data. ColoCare is an international prospective cohort study designed to discover and validate novel prognostic biomarkers and risk prediction models (Ulrich et al., 2019). The study has been approved by the ethics committee of the University of Heidelberg and all study participants provided written informed consent. ColoCare participants were followed longitudinally through in-person follow-up, systematic medical record review, and linkage to cancer registry and vital status records. Standardized medical record abstraction obtained detailed information on surgical procedures, oncologic treatments, and clinical outcomes. Vital status was ascertained through multiple sources, including local and external medical records, routine follow-up mailings, and linkage to state and national cancer and death registries.

3.2 *Biospecimen collection and assays*

The details regarding the sample collection/preparation and data acquisition/pre-processing were described by Liesenfeld et al. (2015). Briefly, first-morning urine samples (predominantly fasting) were collected prior to surgery, aliquoted immediately, and stored at -80°C . Patients reporting alcohol or tobacco use 24 hours before urine collection were excluded. Sample processing followed Cheng et al. (2012), without urease treatment. Pooled quality control (QC) samples and Kovat's alkane mixtures were injected with each analytical batch. Gas chromatography–mass spectrometry (GC-MS) analysis was performed on an Agilent 6890 GC/5973 MS single quadrupole system. Raw data were converted to netCDF format and processed with MZMine 2.0 (Pluskal et al., 2010). Chromatographic peaks were annotated at four levels (Sumner et al., 2007):

Level 1 (reference standard available), Level 2 (electron impact (EI) spectral match ($> 75\%$) and Kovat’s retention indices (RI) match (± 20) to the National Institute of Standards and Technology (NIST) database), Level 3 (class-level EI spectral match $> 75\%$), Level 4 (unknown; excluded). We excluded metabolites with $> 30\%$ coefficient of variance (CV) in QC samples (Dunn et al., 2011), and subsequently removed those with $> 40\%$ missing values. Remaining missing values were imputed using the minimum observed value for each metabolite, consistent with our prior work showing no substantial differences between imputation methods (Liesenfeld et al., 2015). A total of 121 Level 1–3 annotated metabolites were included and normalized by sum (Sun & Xia, 2024), $\log_{10}$ -transformed, and scaled (mean-centered and divided by the standard deviation of each variable) for subsequent statistical analysis.

3.3 Statistical method

3.3.1 Methods of modeling

The primary outcome was overall survival (OS), defined as the time from the date of CRC primary surgery to death from any cause or censoring at last follow-up. OS was modeled following the Cox proportional hazards framework, with the baseline hazard function estimated non-parametrically via the Breslow estimator (Breslow, 1975). We fit five Cox proportional hazard models: a clinical reference model including tumor stage (I–III vs. IV) and age at diagnosis (continuous), and four penalized models to handle the high-dimensional metabolite data, including LASSO, adaptive LASSO, spike-and-slab LASSO, iSIS-LASSO. LASSO imposes an ℓ_1 penalty to produce sparse solutions and allows simultaneous estimation in settings with many potentially correlated predictors (Tibshirani, 1996). Adaptive LASSO uses data-driven penalty weights, and under suitable conditions enjoys “oracle” properties (Fan & Li, 2001), achieving consistent variable

selection as if the true underlying model were known (Zou, 2006). Spike-and-slab LASSO adopts a Bayesian mixture (spike-and-slab) prior to better separate signal from noise and mitigates the shrinkage bias of standard LASSO (Tang et al., 2017; Yi et al., 2019). SIS was designed for ultrahigh-dimensional problems and can reduce dimensionality from very large p to a manageable subset while retaining the true signals with high probability (“sure screening” property) (Fan & Lv, 2008; Fan et al., 2009, 2010). In the iSIS–LASSO model, SIS is iteratively combined with LASSO, repeatedly screening and penalizing predictors until a stable set of signals is identified, thereby improving recovery of relevant metabolites. In all penalized models, tumor stage and age at diagnosis were included as unpenalized covariates. Detailed model specifications and tuning are provided in **Supplement Table S1.1**.

3.3.2 Prediction performance assessment

Model discrimination was assessed using the concordance index (C-index) (Harrell et al., 1996). The C-index measures the probability that, for a randomly selected pair of subjects whose survival times can be definitively ordered, the one who experiences the event earlier is assigned a higher predicted risk (shorter predicted survival); To assess how discrimination may change at different time points (1st, 2nd, 3rd, 4th, and 5th year after surgery), we evaluated the time-dependent area under the receiver operating characteristic curve (tAUC) (Heagerty et al., 2000), using the cumulative/dynamic (C/D) definition (Heagerty & Zheng, 2005; Kamarudin et al., 2017). We also calculated the integrated AUC (iAUC), defined as the weighted average of the tAUC over the follow-up period, to summarize overall discrimination across time. Both tAUC and iAUC were estimated using the method of Uno et al. (2007), which applies inverse probability of censoring weighting and yields consistent estimates under right censoring, assuming that censoring

is independent of event times.

The above performance metrics were estimated using nested cross-validation, consisting of an outer loop (cross-validation layer) for performance estimation and an inner loop for hyperparameter tuning. In the outer loop, the data were randomly partitioned into five approximately equal-sized folds (outer folds) with similar event rates. Each outer fold served once as the test set, while the remaining four outer folds formed the training set. Within each outer training set, an inner 10-fold cross-validation was performed to select hyperparameters. The model was then refit on the full outer training set using the selected hyperparameters and evaluated on the corresponding outer test set. Predictions from each outer test set were used to compute the C-index, tAUC, and iAUC; the performance was summarized by averaging each metric across the five outer test folds. This procedure prevents optimistic bias by ensuring that test data are never used for model tuning (Cawley & Talbot, 2010; Varma & Simon, 2006) and provides an approximately unbiased estimate of the generalization performance of the entire modeling procedure. The reported discrimination metrics reflect expected model performance on new data from the same underlying population (Hastie et al., 2009; Stone, 1974).

3.3.3 Final model fit and calibration assessment

Final models were trained on all study participants, with hyperparameters selected *via* 10-fold cross-validation using the same criterion as in the nested procedure. As no independent test set was available, the performance of these final models cannot be directly evaluated; instead, nested cross-validation provides an approximately unbiased estimate of generalization performance of the modeling procedure. Selected predictors, corresponding effect estimates, and model calibration analyses are based on the final models.

Calibration was assessed by comparing observed and model-predicted event rates across groups defined by the model's risk score (e.g., quartiles of the linear predictor) (Harrell et al., 1996; Royston & Altman, 2013). Within each group, observed event rates were calculated as event proportions and predicted event rates as the mean model-based event probability. Overall calibration was evaluated using a Poisson regression model, with a likelihood ratio test assessing whether observed event counts were consistent with model-expected counts across groups, thereby reflecting agreement between observed and predicted event rates (Atkinson et al., 2008; Berry, 1983; Crowson et al., 2016). Additional details are provided in **Supplement Table S1.2**.

3.3.4 Sensitivity analysis

Given accumulating evidence that tumor site may be associated with OS in CRC (Liesenfeld et al., 2015; Ose et al., 2022; Petrelli et al., 2017; Shida et al., 2020), we conducted sensitivity analyses with additional adjustment for tumor site (rectum vs. colon). Tumor site was added to the clinical reference model and treated as an unpenalized covariate as other clinical variables in all penalized Cox models. All other model specification, tuning, and validation were unchanged. Results from the sensitivity analyses are presented in **Supplement 2**.

3.3.5 Simulation Study

To evaluate the robustness and finite-sample properties of the penalized Cox models under the empirical correlation structure of the metabolomics data, we conducted a real data-based simulation study leveraging the empirical metabolite design matrix. We evaluated eighteen scenarios by systematically varying signal strength (three levels), multicollinearity (two levels), and censoring rates (three levels). To ensure stable performance estimates, we performed 200

iterations per scenario. Model performance was characterized by discrimination, true signal recovery, and coefficient estimation accuracy (see **Supplement 3**).

All statistical analyses were conducted in R (v4.4.2). Penalized Cox proportional hazards models were implemented using the `glmnet` package for LASSO and adaptive LASSO. Spike-and-slab LASSO models were developed using custom wrappers interfacing with the `BhGLM` package, and the iSIS-LASSO procedure was implemented using custom functions with `glmnet` utilities. Predictive performance evaluation regarding C-index was implemented *via* the `survival` package, while iAUC and tAUC were calculated using the `survAUC` package.

3.4 Metabolite Set Enrichment Analysis

For the metabolites identified from each model, we conducted metabolite set enrichment analysis using the MetaboAnalyst 6.0 (Pang et al., [2024](#)). For the input reference set, we chose the ‘Disease-associated metabolite sets’ (reported in urine) which contain 384 entries (accessed December 2025). The hypergeometric test was applied to evaluate whether a particular metabolite set is represented more than expected by chance within the given reference set. Benjamini-Hochberg procedure was used to adjusted p-values for multiple comparisons (Benjamini & Hochberg, [1995](#)). The results are presented in **Supplement Figure S1.1** and **Figure S1.2**.

4 Results

4.1 Patient characteristics and survival

Seventy-six patients were included in the current study, of whom 23 (30%) died during follow-up. Median follow-up time was 3.03 years. The estimated survival probabilities were 0.93 (95% CI: 0.88–0.99) at 6 months, 0.84 (95% CI: 0.76–0.93) at 1 year, 0.80 (95% CI: 0.71–0.89) at 2 years, and 0.59 (95% CI: 0.45–0.79) at 5 years. Baseline patient characteristics are summarized in **Table 1**.

4.2 Model discrimination

All metabolomics-based penalized models improved discrimination over the reference model (**Figure 1**). C-indices were 0.74 for LASSO, 0.75 for adaptive LASSO, 0.75 for spike-and-slab LASSO, 0.73 for iSIS-LASSO, compared with 0.68 for the reference model. Consistent pattern was observed for iAUC: spike-and-slab LASSO and adaptive LASSO had the highest iAUC (0.68), followed closely by LASSO and iSIS-LASSO (0.67), representing an improvement of about 8 ~ 9.7% compared to the reference model (iAUC 0.62).

Time-dependent AUCs across the 5-year horizon showed how discrimination changed over follow-up time. At 1st year of follow-up, tAUCs for the penalized models ranged from 0.72 to 0.75 (LASSO: 0.73 , adaptive LASSO: 0.74 , spike-and-slab LASSO: 0.75, iSIS-LASSO: 0.72), compared with 0.65 for the reference model, indicating better short-term discrimination. The performance peaked at year 3 for all penalized models (LASSO: 0.77, adaptive LASSO: 0.78, spike-and-slab LASSO: 0.79, iSIS-LASSO: 0.76), compared with 0.68 for the reference model. Although tAUCs declined after 3 years of follow-up, the penalized models, particularly spike-

and-slab and adaptive LASSO, maintained higher discrimination than the reference at every time point.

4.3 Final models and model calibration

For all models, tumor stage (stage I–III vs. stage IV as the reference) and age at diagnosis were included as clinical covariates. Tumor stage showed consistently large effects, with lower hazard for patients with stage I–III vs. stage IV ($\hat{\beta}_{\text{median}} \approx -2.09$; range -2.12 to -2.08 ; median and range denote the median and minimum-to-maximum coefficient estimates across models), whereas age at diagnosis showed a small positive effect overall ($\hat{\beta}_{\text{median}} \approx 0.005$; range -0.003 to 0.010), suggesting slightly higher risk with increasing age (**Figure 3, Table 2**). Several metabolites showed reproducible and directionally consistent associations with OS. In particular, 2-hydroxyisobutyrate was selected by all penalized models with negative coefficients ($\hat{\beta}_{\text{median}} \approx -0.23$; range -0.43 to -0.02), indicating a potentially inverse association with mortality. In contrast, indolelactate ($\hat{\beta}_{\text{median}} \approx 0.31$; range 0.13 to 0.58) and a uridine-like metabolite ($\hat{\beta}_{\text{median}} \approx 0.31$; range 0.24 to 1.00) were consistently selected with positive coefficients, suggesting that higher levels of these metabolites are associated with increased hazard.

For model calibration, we compared observed and predicted event rates across risk quartiles defined by the linear predictor (**Figure 2**). Calibration was summarized using the observed-to-expected (O/E) ratio of observed to model-predicted event counts. Across models, O/E values were generally close to 1 across quartiles, indicating good agreement between observed and predicted event rates (**Table 3**). Most models showed mild overprediction in the lowest-risk quartile ($\text{O/E} < 1$), and near agreement or slight underprediction in the highest-risk quartile (e.g., $\text{O/E} \approx 1.12$). In the Poisson calibration models, likelihood ratio tests for group effects yielded p -values of 0.66

(LASSO), 0.35 (adaptive LASSO), 0.18 (spike-and-slab LASSO), 0.54 (iSIS-LASSO), and 1.00 (reference Cox model), indicating no evidence that observed event counts differed from model-expected counts across quartiles.

4.4 Sensitivity Analysis

In sensitivity analyses additionally adjusting for tumor site, model discrimination and calibration were comparable to those observed in the primary analyses (see **Supplement 2**). The composition of the final models remained stable, with key metabolites, particularly the uridine-like metabolite and indolelactate, consistently selected with similar effect estimates across models.

4.5 Simulation Study

Simulation analyses showed that, across all scenarios (varying signal strength, multicollinearity and censoring rates), the four modeling procedures demonstrated comparable discrimination performance. However, clear trade-offs emerged in variable selection and parameter estimation. Notably, iSIS-LASSO model achieved the highest rate of true signal recovery (full-set recovery), but also exhibited the largest ℓ_2 -Norm of model coefficient estimation error among the penalized models (see **Supplement 3**).

4.6 Metabolite enrichment set analysis

The core metabolite lists (2-hydroxyisobutyrate(HMDB0000729), indolelactate(HMDB0000671), uridine(HMDB0000296)) yielded enrichment of two metabolite sets, “Argininemia” and “Colorectal cancer”, however, the enrichment sets did not meet the FDR threshold of 0.05 after adjusted for multiple comparisons. (**Figure S1.1**). Using the spike-and-slab LASSO model-derived metabolite list (2-hydroxyisobutyrate, valine (HMDB0000883), xylitol(HMDB0242149),

indolelactate, uridine), we observed nominal enrichment across several disease-metabolite sets, most prominently for “Colorectal cancer”. After adjusting for multiple comparisons, these enrichments did not meet the FDR threshold of 0.05. (**Figure S1.2**).

5 Discussion

We evaluated whether preoperative urinary metabolites could improve CRC OS prediction beyond tumor stage and age at diagnosis. As expected, tumor stage remained the dominant prognostic factor, whereas age showed a small effect. In addition to these clinical covariates, several urinary metabolites contributed to prognostic prediction. Across four penalized Cox models (LASSO, adaptive LASSO, spike-and-slab LASSO, and iSIS-LASSO), variable selection showed highly concordance, with comparable coefficient estimates, and 2-hydroxyisobutyrate, indolelactate, and a uridine-like metabolite were consistently identified as predictors of CRC OS.

In nested cross-validation, metabolite-augmented penalized models achieved C-indices of 0.72–0.75, outperforming the stage- and age-only reference model (C-index 0.68), with higher iAUC and consistently improved tAUCs at all time points. Although these gains are small, they are notable given the limited samples (76 patients, 23 deaths) and suggest that urinary metabolomics provides incremental prognostic information beyond established clinical factors. Calibration analyses showed reasonable agreement between predicted and observed event rate, with no evidence of systematic miscalibration.

The simulation study assessed the predictive discrimination achievable across four penalized models, under the empirical correlation structure of the metabolomic data. By preserving this structure and varying censoring levels, moderate discrimination (C-index > 0.70) was attainable under stronger signal scenarios (e.g., $\beta = 0.7$), even with high censoring. These results demonstrate that, under the empirical correlation structure, metabolite signals of plausible strength can achieve moderate discrimination, supporting the performance observed when metabolites are incorporated into the clinically adjusted model. While the discrimination performance was similar across

penalized models, the simulations highlighted differences in variable selection and estimation. In particular, iSIS-LASSO achieved higher recovery of true signals but at the cost of increased model coefficient estimation error. This trade-off is consistent with known challenges of penalized regression in high-dimensional settings, where shrinkage can introduce bias in coefficient estimates (Fan & Lv, 2008; Tibshirani, 1996). In the primary analysis, the other penalized methods identified metabolites overlapping with those selected by iSIS-LASSO, suggesting these overlapping features reflect robust underlying signal rather than method-specific artifacts. These findings support the robustness of the identified metabolite signatures at the feature-selection level and the observed predictive performance. However, individual coefficient estimates may be biased and should be interpreted cautiously.

Several biologically relevant metabolites were consistently associated with OS across models. Our findings point to branched-chain amino acid and energy metabolism (2-hydroxyisobutyrate and valine) (Brown et al., 2026; Huang et al., 2018; E. Xu et al., 2023; Y. Zhou et al., 2025), microbial tryptophan–indole metabolism (indolelactate) (Z. Q. Liu et al., 2025), carbohydrate and polyol metabolism (xylitol and a sugar-acid feature) (Cannon et al., 2025), and nucleoside turnover (a uridine-like metabolite) (Choi et al., 2025). Higher urinary 2-hydroxyisobutyrate and valine were associated with better survival, whereas prior studies showed mixed results linking branched-chain–related metabolites and CRC progression (Shan et al., 2019; E. Xu et al., 2023; Y. Zhou et al., 2025). Differences in study design, biospecimen (urine *vs.* blood or tissue), metabolomics platforms and renal clearance may partly explain these discrepancies. The positive association of a uridine-like metabolite with poorer survival is consistent with reports that elevated urinary or circulating nucleosides in CRC reflect increased RNA turnover and tumor cell proliferation (Choi et al., 2025; Rais et al., 2024). Higher urinary indolelactate was associated with increased

hazard in our data, in contrast to experimental work describing indole-3-lactic acid as a locally anti-inflammatory and tumor-suppressive agent (Y. Liu et al., [2023](#); S. Z. Zhou et al., [2025](#)). This discrepancy likely reflects differences between mucosal and systemic/urinary levels and should be interpreted with caution. Xylitol and related polyol signals are relatively understudied in CRC but align with emerging evidence that polyol pathway alterations affect redox status, microbiota, and tumor metabolism (Cannon et al., [2024](#); Sahasakul et al., [2022](#)).

Several limitations should be acknowledged. The sample size and number of deaths limited model complexity and increased the risk of overoptimism despite cross-validation, motivating the use of Cox models with linear predictors. Although data on relevant factors such as comorbidities, lifestyle variables, and tumor molecular features were available in the ColoCare Study, these covariates were not included in our models to minimize model complexity and the risk of overfitting. As a result, some informative predictors may have been excluded, potentially limiting predictive performance. Future studies with larger sample sizes and more events should evaluate the incremental performance of incorporating these variables. Furthermore, the analytical depth was constrained by the technology available at the time of sample acquisition. The single-quadrupole GC-MS instrument necessitated a trade-off between scan rates and dwell times, limiting sensitivity and metabolome coverage compared to modern GC-TOF-MS instruments. In addition, MSEA was performed on a parsimonious set of 3–5 metabolites. This small input size limits statistical power, increasing the risk of false-positive overrepresentation. Finally, this was a single-center study without an independent validation cohort; therefore, the findings should be considered exploratory and require external validation. However, our multi-model framework reduces reliance on any single model, and simulation study demonstrates that the selected metabolite features are robust under the empirical correlation structure of the metabolomic

data and the observed censoring rate.

6 Conclusions

This study demonstrates that preoperative urinary metabolites improve CRC OS prediction beyond tumor stage and age at diagnosis. Penalized Cox models incorporating urinary metabolites improved discrimination compared with a stage- and age-only reference model. Across penalized models, a small metabolite panel, including indolelactate, 2-hydroxyisobutyrate, and a uridine-like metabolite, emerged as robust signals linked to relevant metabolic and microbiome-related pathways.

A complementary simulation study, leveraging the empirical metabolite correlation structure, suggested that these findings are consistent with metabolite effects of limited magnitude under realistic sample size and censoring conditions. Validation in larger, multi-center cohorts is needed to establish the clinical utility of urine-based metabolite signatures for risk stratification and personalized CRC management.

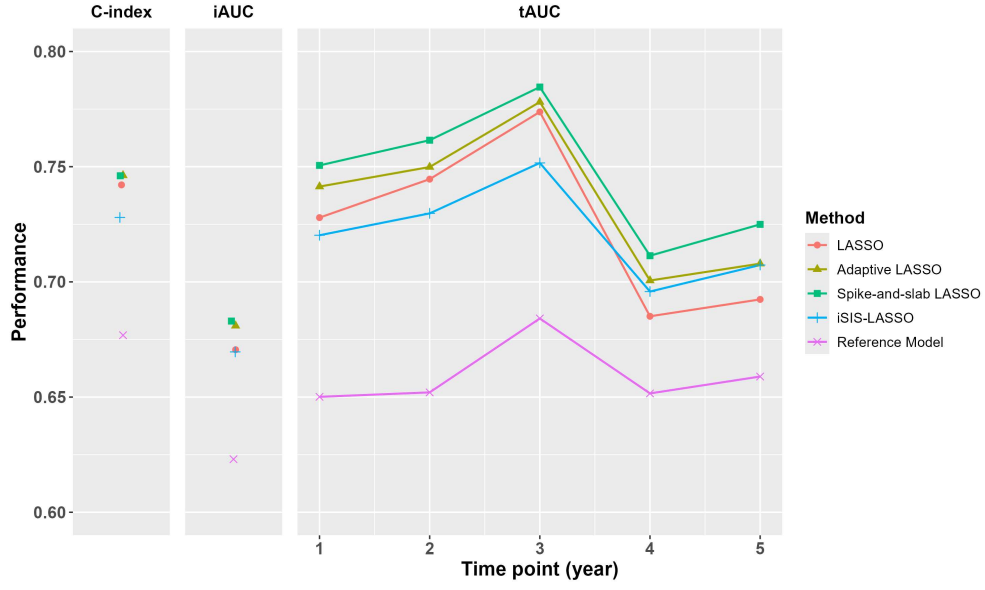


Fig 1: Model discrimination assessment

Figure 1. Discrimination performance of models. Performance metrics were evaluated using nested cross-validation. Panels (left to right) display the mean performance across folds for: the concordance index (C-index), the integrated AUC (iAUC) over the total follow-up period, and the time-dependent AUC (tAUC) at 1st, 2nd, 3rd, 4th, and 5th year. Models compared include LASSO, adaptive LASSO, spike-and-slab LASSO, iSIS-LASSO, and a clinical reference model. Higher values represent superior discrimination performance.

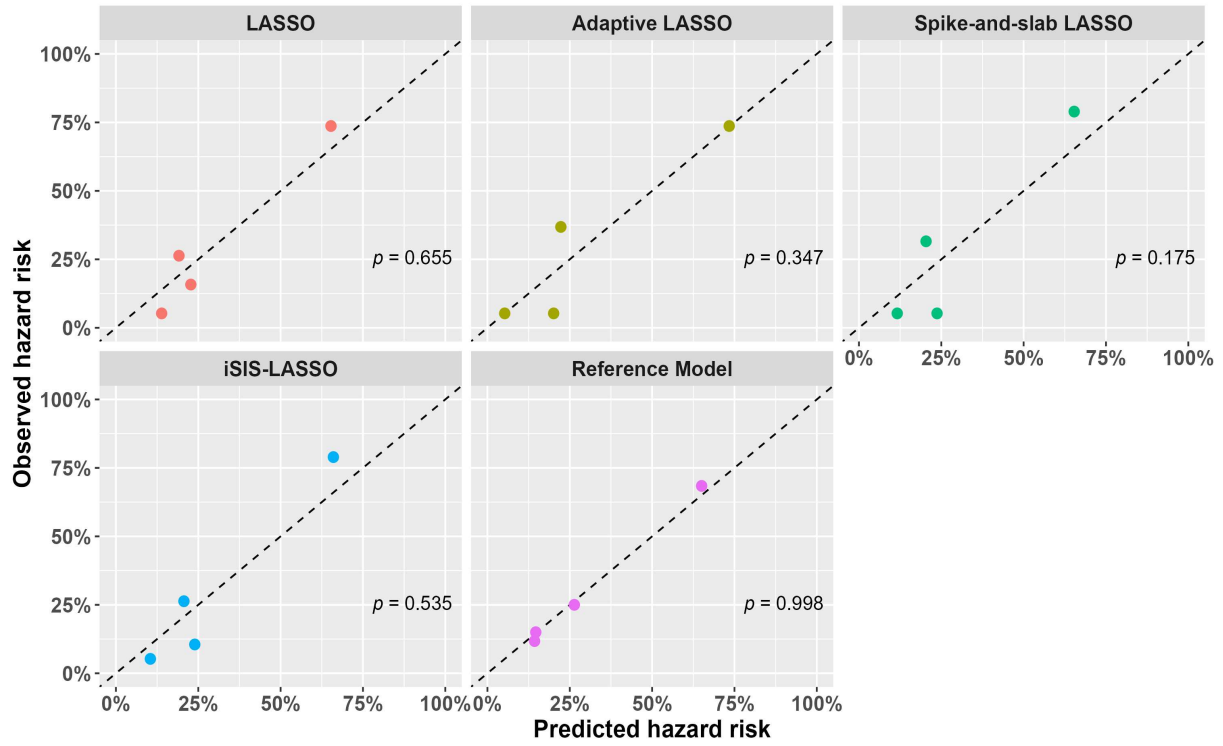


Fig 2: Model calibration assessment

Figure 2. Calibration of final prognostic models fit on all study participants. Each panel shows, for a given modeling approach (LASSO, adaptive LASSO, spike-and-slab LASSO, iSIS-LASSO, and the clinical reference model), the observed *vs.* predicted hazard risk across patient risk groups, where groups are defined by quartiles of the model's risk score (linear predictor). The dashed diagonal line indicates perfect calibration, and the p-value in each panel is from a goodness-of-fit test comparing predicted and observed risks across quartiles of risk.

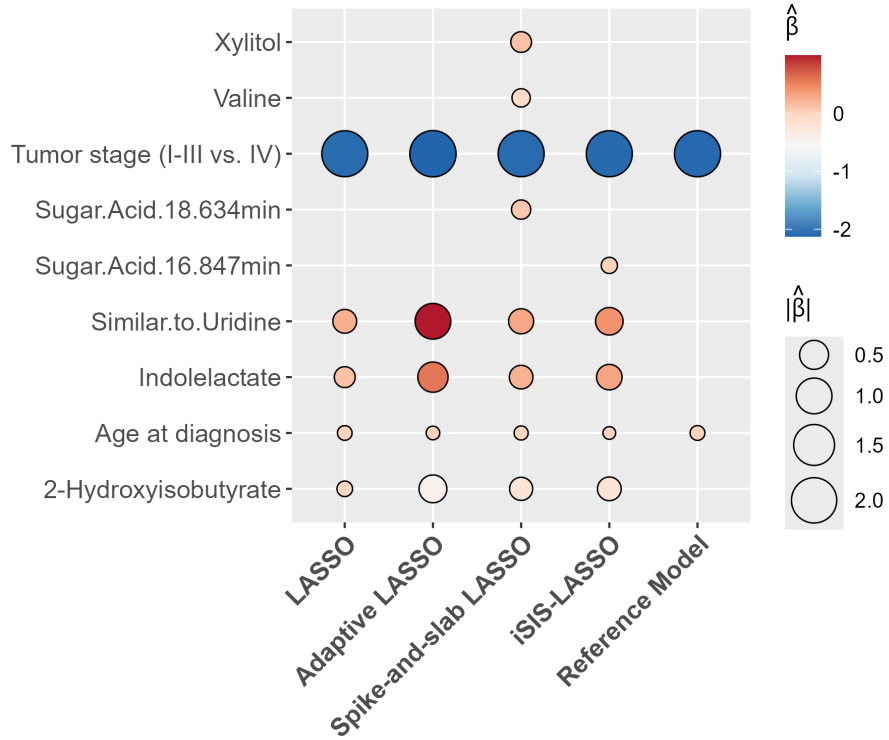


Fig 3: Model coefficient estimation from final models

Figure 3. Model coefficients of final prognostic models fit on all study participants. Columns correspond to modeling approaches (LASSO, adaptive LASSO, spike-and-slab LASSO, iSIS-LASSO and the clinical reference model), and rows represent to clinical and metabolite predictors in corresponding models. Each bubble indicates that a corresponding variable is included in the final model, with fill color representing the sign and magnitude of the estimated log-hazard ratio ($\hat{\beta}$; red = positive, blue = negative) and bubble size proportional to $|\hat{\beta}|$.

Table 1: Patient characteristics at baseline

Overall (N=76)	
Age at diagnosis	
Mean (SD)	65.4 (12.3)
Median [Min, Max]	66.0 [35.0, 90.0]
Sex	
Female	26 (34.2%)
Male	50 (65.8%)
Tumor site	
Colon	35 (46.1%)
Rectum	41 (53.9%)
Tumor stage	
I	11 (14.5%)
II	26 (34.2%)
III	23 (30.3%)
IV	16 (21.1%)
Neoadjuvant	
No	51 (67.1%)
Yes	25 (32.9%)
Adjuvant	
No	48 (63.2%)
Yes	28 (36.8%)

Table 2: Model coefficients for selected features

Feature	LASSO	Adaptive LASSO	Spike-and-slab	iSIS-LASSO	Reference Model
Xylitol	-	-	0.124	-	-
Valine	-	-	-0.061	-	-
Tumor stage (I-III vs. IV)	-2.084	-2.123	-2.085	-2.088	-2.098
Sugar.Acid.18.634min	-	-	0.083	-	-
Sugar.Acid.16.847min	-	-	-	0.025	-
Similar.to.Uridine	0.235	0.999	0.291	0.419	-
Indolelactate	0.130	0.583	0.229	0.313	-
Age at diagnosis	0.009	-0.003	0.005	0.002	0.010
2-Hydroxyisobutyrate	-0.017	-0.433	-0.208	-0.232	-

Note: The reference model includes only tumor stage and age at diagnosis; penalized models are adjusted for tumor stage and age at diagnosis.

Table 3: Calibration assessment – observed and expected event rates across predicted risk quartiles for each modeling approach

Method	Quartile 1 (O / E)	Quartile 2 (O / E)	Quartile 3 (O / E)	Quartile 4 (O / E)
LASSO	0.053 / 0.139	0.158 / 0.227	0.263 / 0.191	0.737 / 0.653
Adaptive LASSO	0.053 / 0.052	0.053 / 0.201	0.368 / 0.223	0.737 / 0.734
Spike-and-slab LASSO	0.053 / 0.116	0.053 / 0.237	0.316 / 0.204	0.789 / 0.654
iSIS-LASSO	0.053 / 0.105	0.105 / 0.239	0.263 / 0.206	0.789 / 0.660
Reference Model	0.150 / 0.147	0.250 / 0.264	0.118 / 0.143	0.684 / 0.650

Note: O/E denotes the ratio of observed (O) and expected (E) event rates within each predicted risk quartile; E was computed as the mean predicted event probability in that quartile.

Disclosures

The authors declare no financial or other competing interests that could have influenced the objectivity of this research or the preparation of this manuscript, except as noted below.

D.L. is currently an employee of Roche (Roche Pharma AG, Emil-Barell-Str. 1, 79639 Grenzach-Wyhlen, Germany). This work was conducted while D.L. was affiliated with NCT/DKFZ and is unrelated to his current employment.

C.M.U., as cancer center director, oversees research funded by several pharmaceutical companies but has not received direct personal funding.

Data Availability

The ColoCare Study data are available from `colocarestudy_admin@hci.utah.edu` on reasonable request and as described on the ColoCare website (<https://uofuhealth.utah.edu/huntsman/labs/colocare-consortium/>). Our data sharing procedures are available online (<https://uofuhealth.utah.edu/huntsman/labs/colocare-consortium/data-sharing/new-projects.php>). For any additional questions, please contact the ColoCare Study Administrator Team (`colocarestudy_admin@hci.utah.edu`).

References

- Atkinson, E., Crowson, C., Pedersen, R., & Therneau, T. (2008). Poisson models for person-years and expected rates. *Statistics in Medicine*, 27(24), 4817–4830. <https://doi.org/10.1002/sim.3435>
- Benjamini, Y., & Hochberg, Y. (1995). Controlling the false discovery rate: A practical and powerful approach to multiple testing. *Journal of the Royal Statistical Society: Series B (Methodological)*, 57(1), 289–300. <https://doi.org/10.1111/j.2517-6161.1995.tb02031.x>
- Berry, G. (1983). The analysis of mortality by the subject-years method. *Biometrics*, 39(1), 173–184. <https://www.jstor.org/stable/2530817>
- Bertini, I., Cacciatore, S., Jensen, B. V., Schou, J. V., Johansen, J. S., Kruhøffer, M., Luchinat, C., Nielsen, D. L., & Turano, P. (2012). Metabolomic nmr fingerprinting to identify and predict survival of patients with metastatic colorectal cancer. *Cancer Research*, 72(1), 356–364. <https://doi.org/10.1158/0008-5472.CAN-11-1543>
- Bouatra, S., Aziat, F., Mandal, R., Guo, A. C., Wilson, M. R., Knox, C., Bjorndahl, T. C., Krishnamurthy, R., Saleem, F., Liu, P., Dame, Z. T., Poelzer, J., Huynh, J., Yallou, F. S., Psychogios, N., Dong, E., Bogumil, R., Roehring, C., & Wishart, D. S. (2013). The human urine metabolome. *PLOS ONE*, 8(9), e73076. <https://doi.org/10.1371/journal.pone.0073076>
- Breslow, N. E. (1975). Analysis of survival data under the proportional hazards model. *International Statistical Review / Revue Internationale de Statistique*, 43(1), 45–57. <https://doi.org/10.2307/1402659>

- Brezmes, J., Llambrich, M., Cumeras, R., & Gumà, J. (2022). Urine NMR metabolomics for precision oncology in colorectal cancer. *International Journal of Molecular Sciences*, 23(19), 11171. <https://doi.org/10.3390/ijms231911171>
- Brown, M. L., Cai, X., & Simon, M. C. (2026). Branched chain amino acids and their aberrant metabolism in cancer. *Trends in Cancer*, 12(1), 34–47. <https://doi.org/10.1016/j.trecan.2025.10.004>
- Cannon, M., Cosantino, A., Tran, L., Chandel, N. S., & Ghoreishi, N. (2025). Effects of xylitol on tumor progression in syngeneic mice cancer models. *Nutraceuticals*, 5(1), 4. <https://doi.org/10.3390/nutraceuticals5010004>
- Cannon, M., Dempsey, E., Cosantino, A., Chandel, N., & Ghoreishi-Haack, N. (2024). Cancer cell line inhibition by osmotic pump-administered xylitol in a syngeneic mouse model. *Research Square*. <https://doi.org/10.21203/rs.3.rs-3977059/v1>
- Cawley, G. C., & Talbot, N. L. C. (2010). On over-fitting in model selection and subsequent selection bias in performance evaluation. *Journal of Machine Learning Research*, 11, 2079–2107. <https://www.jmlr.org/papers/v11/cawley10a.html>
- Cheng, Y., Xie, G. X., Chen, T. L., Qiu, Y. P., Zou, X., Zheng, M. H., Tan, B. B., Feng, B., Dong, T. T., He, P. A., Zhao, L. J., Zhao, A. H., Xu, L. X., Zhang, Y., & Jia, W. (2012). Distinct urinary metabolic profile of human colorectal cancer. *Journal of Proteome Research*, 11(2), 1354–1363. <https://doi.org/10.1021/pr201001a>
- Choi, K. M., Berard, B. A., Yoon, J. H., & Kim, D. (2025). Uridine as a hub in cancer metabolism and rna biology. *Experimental and Molecular Medicine*, 57(8), 1651–1662. <https://doi.org/10.1038/s12276-025-01402-7>

- Crowson, C. S., Atkinson, E. J., & Therneau, T. M. (2016). Assessing calibration of prognostic risk scores. *Statistical Methods in Medical Research*, 25(4), 1692–1706. <https://doi.org/10.1177/0962280213497434>
- De Bin, R., Boulesteix, A.-L., Benner, A., Becker, N., & Sauerbrei, W. (2020). Combining clinical and molecular data in regression prediction models: Insights from a simulation study. *Briefings in Bioinformatics*, 21(6), 1904–1919. <https://doi.org/10.1093/bib/bbz136>
- Dunn, W. B., Broadhurst, D., Begley, P., Zelena, E., Francis-McIntyre, S., Anderson, N., Brown, M., Knowles, J. D., Halsall, A., Haselden, J. N., Nicholls, A. W., Wilson, I. D., Kell, D. B., & Goodacre, R. (2011). Procedures for large-scale metabolic profiling of serum and plasma using gas chromatography and liquid chromatography coupled to mass spectrometry. *Nature Protocols*, 6(7), 1060–1083. <https://doi.org/10.1038/nprot.2011.335>
- Erben, V., Bhardwaj, M., Schrotz-King, P., & Brenner, H. (2018). Metabolomics biomarkers for detection of colorectal neoplasms: A systematic review. *Cancers*, 10(8). <https://doi.org/10.3390/cancers10080246>
- Fan, J., Feng, Y., & Wu, Y. (2010). High-dimensional variable selection for cox's proportional hazards model. *Annals of Statistics*, 38(4), 2307–2334. <https://doi.org/10.1214/10-AOS798>
- Fan, J., & Li, R. (2001). Variable selection via nonconcave penalized likelihood and its oracle properties. *Journal of the American Statistical Association*, 96(456), 1348–1360. <https://doi.org/10.1198/016214501753382273>
- Fan, J., & Lv, J. (2008). Sure independence screening for ultrahigh dimensional feature space. *Journal of the Royal Statistical Society: Series B*, 70(5), 849–911. <https://doi.org/10.1111/j.1467-9868.2008.00674.x>

- Fan, J., Samworth, R., & Wu, Y. (2009). Ultrahigh dimensional feature selection: Beyond the linear model. *Journal of Machine Learning Research*, 10, 2013–2038. <https://www.jmlr.org/papers/v10/fan09a.html>
- Farshidfar, F., Weljie, A. M., Kopciuk, K. A., Hilsden, R., McGregor, S. E., Buie, W. D., MacLean, A., Vogel, H. J., & Bathe, O. F. (2016). A validated metabolomic signature for colorectal cancer: Exploration of the clinical value of metabolomics. *British Journal of Cancer*, 115(7), 848–857. <https://doi.org/10.1038/bjc.2016.243>
- Fiehn, O. (2002). Metabolomics—the link between genotypes and phenotypes. *Plant Molecular Biology*, 48(1–2), 155–171.
- Gold, A., Choueiry, F., Jin, N., Mo, X., & Zhu, J. (2022). The application of metabolomics in recent colorectal cancer studies: A state-of-the-art review. *Cancers*, 14(3), 725.
- Griffin, J. L., & Shockcor, J. P. (2004). Metabolic profiles of cancer cells. *Nature Reviews Cancer*, 4(7), 551–561. <https://doi.org/10.1038/nrc1390>
- Guo, C., Xie, C., Chen, Q., Cao, X., Guo, M., Zheng, S., & Wang, Y. (2018). A novel malic acid-enhanced method for the analysis of 5-methyl-2'-deoxycytidine, 5-hydroxymethyl-2'-deoxycytidine, 5-methylcytidine and 5-hydroxymethylcytidine in human urine using hydrophilic interaction liquid chromatography-tandem mass spectrometry. *Analytica Chimica Acta*, 1034, 110–118. <https://doi.org/10.1016/j.aca.2018.06.081>
- Harrell, F. E., Lee, K. L., & Mark, D. B. (1996). Multivariable prognostic models: Issues in developing models, evaluating assumptions and adequacy, and measuring and reducing errors. *Statistics in Medicine*, 15(4), 361–387. [https://doi.org/10.1002/\(SICI\)1097-0258\(19960229\)15:4<361::AID-SIM168>3.0.CO;2-4](https://doi.org/10.1002/(SICI)1097-0258(19960229)15:4<361::AID-SIM168>3.0.CO;2-4)

- Hastie, T., Tibshirani, R., & Friedman, J. (2009). *The elements of statistical learning: Data mining, inference, and prediction* (2nd). Springer. <https://doi.org/10.1007/978-0-387-84858-7>
- Heagerty, P. J., Lumley, T., & Pepe, M. S. (2000). Time-dependent roc curves for censored survival data and a diagnostic marker. *Biometrics*, 56(2), 337–344. <https://doi.org/10.1111/j.0006-341X.2000.00337.x>
- Heagerty, P. J., & Zheng, Y. (2005). Survival model predictive accuracy and roc curves. *Biometrics*, 61(1), 92–105. <https://doi.org/10.1111/j.0006-341X.2005.030814.x>
- Huang, H., Luo, Z., Qi, S., Huang, J., Xu, P., Wang, X., Gao, L., Li, F., Wang, J., Zhao, W., Gu, W., Chen, Z., Dai, L., Dai, J., & Zhao, Y. (2018). Landscape of the regulatory elements for lysine 2-hydroxyisobutyrylation pathway. *Cell Research*, 28(1), 111–125. <https://doi.org/10.1038/cr.2017.149>
- International Agency for Research on Cancer. (2024). Colorectal cancer [Accessed: 2026-04-07]. <https://www.iarc.who.int/cancer-type/colorectal-cancer/>
- Jardillier, R., Koca, D., Chatelain, F., & Guyon, L. (2022). Prognosis of lasso-like penalized cox models with tumor profiling improves prediction over clinical data alone and benefits from bi-dimensional pre-screening. *BMC Cancer*, 22(1), 1045. <https://doi.org/10.1186/s12885-022-10117-1>
- Jiménez, B., Mirnezami, R., Kinross, J., Cloarec, O., Keun, H. C., Holmes, E., Goldin, R. D., Ziprin, P., Darzi, A., & Nicholson, J. K. (2013). 1h hr-mas nmr spectroscopy of tumor-induced local metabolic field-effects enables colorectal cancer staging and prognostication. *Journal of Proteome Research*, 12(2), 959–968. <https://doi.org/10.1021/pr3010106>

- Kamarudin, A. N., Cox, T., & Kolamunnage-Dona, R. (2017). Time-dependent roc curve analysis in medical research: Current methods and applications. *BMC Medical Research Methodology*, 17(1), 53. <https://doi.org/10.1186/s12874-017-0332-6>
- Kneuert, P. J., Chang, G. J., Hu, C.-Y., Rodriguez-Bigas, M. A., Eng, C., Vilar, E., Skibber, J. M., Feig, B. W., Cormier, J. N., & You, Y. N. (2015). Overtreatment of young adults with colon cancer: More intense treatments with unmatched survival gains. *JAMA Surgery*, 150(5), 402–409. <https://doi.org/10.1001/jamasurg.2014.3572>
- Koncina, E., Haan, S., Rauh, S., & Letellier, E. (2020). Prognostic and predictive molecular biomarkers for colorectal cancer: Updates and challenges. *Cancers*, 12(2), 319. <https://doi.org/10.3390/cancers12020319>
- Lawler, M., Alsina, D., Adams, R. A., Anderson, A. S., Brown, G., Fearnhead, N. S., Fenwick, S. W., Halloran, S. P., Hochhauser, D., Hull, M. A., Koelzer, V. H., McNair, A. G. K., Monahan, K. J., Näthke, I., Norton, C., Novelli, M. R., Steele, R. J. C., Thomas, A. L., Wilde, L. M., & Tomlinson, I. (2018). Critical research gaps and recommendations to inform research prioritisation for more effective prevention and improved outcomes in colorectal cancer. *Gut*, 67(1), 179–193. <https://doi.org/10.1136/gutjnl-2017-315333>
- Liesenfeld, D. B., Habermann, N., Toth, R., Owen, R. W., Frei, E., Staffa, J., Schrotz-King, P., Klika, K. D., & Ulrich, C. M. (2015). Changes in urinary metabolic profiles of colorectal cancer patients enrolled in a prospective cohort study (colocare). *Metabolomics*, 11(4), 998–1012. <https://doi.org/10.1007/s11306-014-0758-3>
- Liu, Y., Pei, Z., Pan, T., Wang, H., Chen, W., & Lu, W. (2023). Indole metabolites and colorectal cancer: Gut microbial tryptophan metabolism, host gut microbiome biomarkers, and

- potential intervention mechanisms. *Microbiological Research*, 272, 127392. <https://doi.org/10.1016/j.micres.2023.127392>
- Liu, Z. Q., Ciudad, M. T., & McGaha, T. L. (2025). New insights into tryptophan metabolism in cancer. *Trends in Cancer*, 11(7), 629–641. <https://doi.org/10.1016/j.trecan.2025.03.008>
- Magnanensi, J., Maumy-Bertrand, M., Meyer, N., & Bertrand, F. (2021). New developments in sparse pls regression. *Frontiers in Applied Mathematics and Statistics*, 7, 693126. <https://doi.org/10.3389/fams.2021.693126>
- Mehmood, T., Liland, K. H., Snipen, L., & Sæbø, S. (2012). A review of variable selection methods in partial least squares regression. *Chemometrics and Intelligent Laboratory Systems*, 118, 62–69. <https://doi.org/10.1016/j.chemolab.2012.07.010>
- Mehmood, T., Sæbø, S., & Liland, K. H. (2020). Comparison of variable selection methods in partial least squares regression. *Journal of Chemometrics*, 34(6), e3226. <https://doi.org/10.1002/cem.3226>
- Nakajima, T., Katsumata, K., Kuwabara, H., Soya, R., Enomoto, M., Ishizaki, T., Tsuchida, A., Mori, M., Hiwatari, K., Soga, T., Tomita, M., & Sugimoto, M. (2018). Urinary polyamine biomarker panels with machine-learning differentiated colorectal cancers, benign disease, and healthy controls. *International Journal of Molecular Sciences*, 19(3), 756. <https://doi.org/10.3390/ijms19030756>
- Ose, J., Gigic, B., Brezina, S., Lin, T., Peoples, A. R., Schobert, P. P., Baierl, A., van Roekel, E., Robinot, N., Gicquiau, A., Achaintre, D., Scalbert, A., van Duijnhoven, F. J. B., Holowatyj, A. N., Gumpenberger, T., Schrotz-King, P., Ulrich, A. B., Ulvik, A., Ueland, P. M., & Ulrich, C. M. (2023). Higher plasma creatinine is associated with an increased risk of

- death in patients with non-metastatic rectal but not colon cancer. *Cancers*, 15(13), 3391. <https://doi.org/10.3390/cancers15133391>
- Ose, J., Gigic, B., Hardikar, S., Lin, T., Himbert, C., Warby, C., Peoples, A., Lindley, C., Boehm, J., Schrotz-King, P., & Figueiredo, J. (2022). Presurgery adhesion molecules and angiogenesis biomarkers are differently associated with outcomes in colon and rectal cancer: Results from the colocare study. *Cancer Epidemiology, Biomarkers & Prevention*, 31(8), 1650–1660. <https://doi.org/10.1158/1055-9965.EPI-21-1357>
- Pang, Z., Lu, Y., Zhou, G., Hui, F., Xu, L., Viau, C., Spigelman, A. F., MacDonald, P. E., Wishart, D. S., Li, S., & Xia, J. (2024). Metaboanalyst 6.0: Towards a unified platform for metabolomics data processing, analysis and interpretation. *Nucleic Acids Research*, 52(W1), W398–W406. <https://doi.org/10.1093/nar/gkae253>
- Petrelli, F., Tomasello, G., Borgonovo, K., Ghidini, M., Turati, L., Dallera, P., Sgroi, G., & Barni, S. (2017). Prognostic survival associated with left-sided vs right-sided colon cancer: A systematic review and meta-analysis. *JAMA Oncology*, 3(2), 211–219. <https://doi.org/10.1001/jamaoncol.2016.4227>
- Phua, L. C., Chue, X. P., Koh, P. K., Cheah, P. Y., Ho, H. K., & Chan, E. C. (2014). Non-invasive fecal metabonomic detection of colorectal cancer. *Cancer Biology & Therapy*, 15(4), 389–397. <https://doi.org/10.4161/cbt.27625>
- Pinto, R. C. (2017). Chemometrics methods and strategies in metabolomics. *Advances in Experimental Medicine and Biology*, 965, 163–190. https://doi.org/10.1007/978-3-319-47656-8_7

- Pluskal, T., Castillo, S., Villar-Briones, A., & Oresic, M. (2010). Mzmine 2: Modular framework for processing, visualizing, and analyzing mass spectrometry-based molecular profile data. *BMC Bioinformatics*, *11*, 395. <https://doi.org/10.1186/1471-2105-11-395>
- Qiu, Y., Cai, G., Zhou, B., Li, D., Zhao, A., Xie, G., Li, H., Cai, S., Xie, D., Huang, C., Ge, W., Zhou, Z., Xu, L. X., Jia, W., Zheng, S., & Yen, Y. (2014). A distinct metabolic signature of human colorectal cancer with prognostic potential. *Clinical Cancer Research*, *20*(8), 2136–2146. <https://doi.org/10.1158/1078-0432.CCR-13-1939>
- Rais, T., Riaz, R., Siddiqui, T., Shakeel, A., Khan, A., & Zafar, H. (2024). Innovations in colorectal cancer treatment: Trifluridine and tipiracil with bevacizumab for improved outcomes – a review. *Frontiers in Oncology*, *14*, 1296765. <https://doi.org/10.3389/fonc.2024.1296765>
- Rocková, V., & George, E. I. (2014). Emvs: The em approach to bayesian variable selection. *Journal of the American Statistical Association*, *109*(506), 828–846. <https://doi.org/10.1080/01621459.2013.869223>
- Rosati, E., Rios Martini, G., Pogorelyy, M. V., Minervina, A. A., Degenhardt, F., Wendorff, M., Sari, S., Mayr, G., Fazio, A., Dowds, C. M., Hauser, C., Tran, F., von Schönfels, W., Pochhammer, J., Salnikova, M. A., Jaeckel, C., Gigla, J. B., Sabet, S. S., Hübenthal, M., & Franke, A. (2022). A novel unconventional t cell population enriched in crohn’s disease. *Gut*, *71*(11), 2194–2204. <https://doi.org/10.1136/gutjnl-2021-325373>
- Royston, P., & Altman, D. G. (2013). External validation of a Cox prognostic model: Principles and methods. *BMC Medical Research Methodology*, *13*(1), 33. <https://doi.org/10.1186/1471-2288-13-33>
- Sahasakul, Y., Angkhasirisap, W., Lam-Ubol, A., Aursalung, A., Sano, D., Takada, K., & Trachootham, D. (2022). Partial substitution of glucose with xylitol prolongs survival and

- suppresses cell proliferation and glycolysis of mice bearing orthotopic xenograft of oral cancer. *Nutrients*, 14(10), 2023. <https://doi.org/10.3390/nu14102023>
- Scheele, J., Schmidt, S. A., Tenzer, S., Henne-Bruns, D., & Kornmann, M. (2018). Overstaging: A challenge in rectal cancer treatment. *Visceral Medicine*, 34(4), 301–306. <https://doi.org/10.1159/000488652>
- Schmidt, D. R., Patel, R., Kirsch, D. G., Lewis, C. A., Vander Heiden, M. G., & Locasale, J. W. (2021). Metabolomics in cancer research and emerging applications in clinical oncology. *CA: A Cancer Journal for Clinicians*, 71(4), 333–358. <https://acsjournals.onlinelibrary.wiley.com/doi/10.3322/caac.21670>
- Seum, T., Cardoso, R., Stevenson-Hoare, J., Holleczech, B., Schöttker, B., Hoffmeister, M., & Brenner, H. (2025). Exploring metabolomics for colorectal cancer risk prediction: Evidence from the uk biobank and esther cohorts. *BMC Medicine*, 23(1), 283. <https://doi.org/10.1186/s12916-025-04107-w>
- Shan, Y. L., Gao, Y., Jin, W., Fan, M. M., Wang, Y., Gu, Y. H., Shan, C. X., Sun, L. J., Li, X., Yu, B., Luo, Q., & Xu, Q. (2019). Targeting h1bch to reprogram valine metabolism for the treatment of colorectal cancer. *Cell Death & Disease*, 10, 618. <https://doi.org/10.1038/s41419-019-1832-6>
- Shida, D., Inoue, M., Tanabe, T., Moritani, K., Tsukamoto, S., Yamauchi, S., Sugihara, K., & Kanemitsu, Y. (2020). Prognostic impact of primary tumor location in stage iii colorectal cancer—right-sided colon versus left-sided colon versus rectum: A nationwide multicenter retrospective study. *Journal of Gastroenterology*, 55(10), 958–968. <https://doi.org/10.1007/s00535-020-01706-7>

- Siegel, R. L., Giaquinto, A. N., & Jemal, A. (2026). Cancer statistics, 2026. *CA: A Cancer Journal for Clinicians*, 76(1), 7–33. <https://doi.org/10.3322/caac.70043>
- Siegel, R. L., Miller, K. D., Wagle, N. S., & Jemal, A. (2026). Colorectal cancer statistics, 2026. *CA: A Cancer Journal for Clinicians*, 76(3), 154–172. <https://doi.org/10.3322/caac.70067>
- Simon, N., Friedman, J., Hastie, T., & Tibshirani, R. (2011). Regularization paths for cox's proportional hazards model via coordinate descent. *Journal of Statistical Software*, 39(5), 1–13. <https://doi.org/10.18637/jss.v039.i05>
- Society, A. C. (2024). *Cancer facts & figures 2024*. <https://www.cancer.org/research/cancer-facts-statistics.html>
- Stone, M. (1974). Cross-validatory choice and assessment of statistical predictions. *Journal of the Royal Statistical Society: Series B (Methodological)*, 36(2), 111–133. <https://doi.org/10.1111/j.2517-6161.1974.tb00994.x>
- Sumner, L. W., Amberg, A., Barrett, D., Beale, M. H., Beger, R., Daykin, C. A., Fan, T. W., Fiehn, O., Goodacre, R., Griffin, J. L., Hankemeier, T., Hardy, N., Harnly, J., Higashi, R., Kopka, J., Lane, A. N., Lindon, J. C., Marriott, P., Nicholls, A. W., & Viant, M. R. (2007). Proposed minimum reporting standards for chemical analysis chemical analysis working group (cawg) metabolomics standards initiative (msi). *Metabolomics*, 3(3), 211–221. <https://doi.org/10.1007/s11306-007-0082-2>
- Sun, J., & Xia, Y. (2024). Pretreating and normalizing metabolomics data for statistical analysis. *Genes & Diseases*, 11(3), 100979. <https://doi.org/10.1016/j.gendis.2023.04.018>
- Tang, Z., Shen, Y., Zhang, X., & Yi, N. (2017). The spike-and-slab lasso cox model for survival prediction and associated genes detection. *Bioinformatics*, 33(18), 2799–2807. <https://doi.org/10.1093/bioinformatics/btx300>

- Tibshirani, R. (1996). Regression shrinkage and selection via the lasso. *Journal of the Royal Statistical Society: Series B*, 58(1), 267–288. <https://www.jstor.org/stable/2346178>
- Trygg, J., Holmes, E., & Lundstedt, T. (2007). Chemometrics in metabonomics. *Journal of Proteome Research*, 6(2), 469–479. <https://doi.org/10.1021/pr060594q>
- Trygg, J., & Wold, S. (2002). Orthogonal projections to latent structures (o-pls). *Journal of Chemometrics*, 16(3), 119–128. <https://doi.org/10.1002/cem.695>
- Ulrich, C. M., Gigic, B., Böhm, J., Ose, J., Viskochil, R., Schneider, M., Colditz, G. A., Figueiredo, J. C., Grady, W. M., Li, C. I., Shibata, D., Siegel, E. M., Toriola, A. T., & Ulrich, A. (2019). The colocare study: A paradigm of transdisciplinary science in colorectal cancer outcomes. *Cancer Epidemiology, Biomarkers & Prevention*, 28(3), 591–601. <https://doi.org/10.1158/1055-9965.EPI-18-0773>
- Uno, H., Cai, T., Tian, L., & Wei, L. J. (2007). Evaluating prediction rules for t-year survivors with censored regression models. *Journal of the American Statistical Association*, 102(478), 527–537. <https://doi.org/10.1198/0162145070000000149>
- Varma, S., & Simon, R. (2006). Bias in error estimation when using cross-validation for model selection. *BMC Bioinformatics*, 7, 91. <https://doi.org/10.1186/1471-2105-7-91>
- Venäläinen, M. K., Roine, A. N., Häkkinen, M. R., Vepsäläinen, J. J., Kumpulainen, P. S., Kiviniemi, M. S., Lehtimäki, T., Oksala, N. K., & Rantanen, T. K. (2018). Altered polyamine profiles in colorectal cancer. *Anticancer Research*, 38(6), 3601–3607. <https://doi.org/10.21873/anticanres.12634>
- Waldmann, P., Meszaros, G., Gredler, B., Fuerst, C., & Sölkner, J. (2013). Evaluation of the lasso and the elastic net in genome-wide association studies. *Frontiers in Genetics*, 4, 270. <https://doi.org/10.3389/fgene.2013.00270>

- Widlak, M. M., Neal, M., Daulton, E., Thomas, C. L., Tomkins, C., Singh, B., Harmston, C., Wicaksono, A., Evans, C., Smith, S., Savage, R. S., Covington, J. A., & Arasaradnam, R. P. (2018). Risk stratification of symptomatic patients suspected of colorectal cancer using faecal and urinary markers. *Colorectal Disease*, 20(12), O335–O342. <https://doi.org/10.1111/codi.14431>
- Wold, H. O. A. (1975). Soft modelling by latent variables: The non-linear iterative partial least squares (nipals) approach. *Journal of Applied Probability*, 12, 117–142. <https://www.jstor.org/stable/3212861>
- World Health Organization. (2024). Colorectal cancer [Accessed: 2026-04-07]. <https://www.who.int/news-room/fact-sheets/detail/colorectal-cancer>
- Xu, E., Ji, B., Jin, K., & Chen, Y. (2023). Branched-chain amino acids catabolism and cancer progression: Focus on therapeutic interventions. *Frontiers in Oncology*, 13, 1220638. <https://doi.org/10.3389/fonc.2023.1220638>
- Xu, Y., Wang, X., Huang, Y., Ye, D., & Chi, P. (2022). A lasso-based survival prediction model for patients with synchronous colorectal carcinomas based on seer. *Translational Cancer Research*, 11(8), 2795–2809. <https://doi.org/10.21037/tcr-20-1860>
- Yi, N., Tang, Z., Zhang, X., & Guo, B. (2019). Bhglm: Bayesian hierarchical glms and survival models, with applications to genomics and epidemiology. *Bioinformatics*, 35(8), 1419–1421. <https://doi.org/10.1093/bioinformatics/bty803>
- Zaimenko, I., Jaeger, C., Brenner, H., Chang-Claude, J., Hoffmeister, M., Grötzing, C., Detjen, K., Burock, S., Schmitt, C. A., Stein, U., & Lisec, J. (2019). Non-invasive metastasis prognosis from plasma metabolites in stage ii colorectal cancer patients: The dachs study. *International Journal of Cancer*, 145(1), 221–231. <https://doi.org/10.1002/ijc.32076>

- Zhang, F., Zhang, Y., Zhao, W., Deng, K., Wang, Z., Yang, C., Ma, L., Openkova, M. S., Hou, Y., & Li, K. (2017). Metabolomics for biomarker discovery in the diagnosis, prognosis, survival and recurrence of colorectal cancer: A systematic review. *Oncotarget*, 8(21), 35460–35472. <https://doi.org/10.18632/oncotarget.16727>
- Zhou, S. Z., Wang, K., Huang, J. D., Xu, Z., Yuan, Q. G., Liu, L. X., Wang, Z. F., Miao, J., Wang, H., Wang, T. T., Guan, W. X., & Ding, C. (2025). Indole-3-lactic acid suppresses colorectal cancer via metabolic reprogramming. *Gut Microbes*, 17(1), 2508949. <https://doi.org/10.1080/19490976.2025.2508949>
- Zhou, Y., Kou, J., Li, W., Wang, Y., Su, X., & Zhang, H. (2025). Bcaa metabolism in cancer progression and therapy resistance: The balance between fuel and cell signaling. *Frontiers in Pharmacology*, 16, 1595176. <https://doi.org/10.3389/fphar.2025.1595176>
- Zou, H. (2006). The adaptive lasso and its oracle properties. *Journal of the American Statistical Association*, 101(476), 1418–1429. <https://doi.org/10.1198/016214506000000735>

Supplementary Material

Supplement 1a: Model Specification & Model Calibration

Table S1.1 summarizes the penalized Cox regression approaches used in this study, including LASSO, adaptive LASSO, spike-and-slab LASSO, and iSIS-LASSO. For each model, the table outlines the penalty structure, variable selection strategy, tuning procedure, and treatment of clinical covariates, highlighting the key distinctions in how high-dimensional metabolite data were incorporated into the models.

Table S1.2 provides a stepwise overview of the Poisson regression model-based calibration procedure. For each subject, the cumulative hazard over follow-up was computed, which under the Cox model corresponds to the expected number of events. These quantities were then aggregated within risk strata defined by the linear predictor. A Poisson regression framework was used to formally evaluate goodness-of-fit while appropriately accounting for censoring.

Table S1.1: Model specifications of penalized Cox modeling

Model	Description
LASSO	An ℓ_1 -penalized Cox proportional hazards model was fit including all metabolites together with tumor stage and age at diagnosis. Tumor stage and age were treated as unpenalized covariates, and penalization was applied only to metabolite coefficients. The penalty parameter λ was selected using stratified K-fold cross-validation, choosing the value that minimized the cross-validated partial likelihood deviance (“lambda.min”).
Adaptive LASSO	A two-step procedure was used. First, a standard LASSO Cox model was fit to obtain initial metabolite coefficient estimates $\tilde{\beta}_j$. Second, a weighted LASSO Cox model was fit using penalty weights $w_j = \frac{1}{ \tilde{\beta}_j ^\gamma}$, $\gamma = 1$, applied to metabolite coefficients. Tumor stage and age at diagnosis were unpenalized in both steps. The penalty parameter was selected using the same cross-validation procedure as in the standard LASSO.
Spike-and-slab LASSO	A spike-and-slab LASSO Cox model was fit, assigning metabolite coefficients a two-component spike-slab prior with fixed scales $ss = c(0.15, 0.5)$. Estimation proceeded via an EM algorithm, updating prior scales and inclusion probabilities and re-estimating coefficients using variable-specific penalties. Tumor stage and age at diagnosis were kept unpenalized by assigning them large prior scales and excluding them from spike-slab updates. The model was implemented using internal routines from the BhGLM R package (Yi et al., 2019), with minimal modifications to <code>.bmlasso()</code> and <code>.bmlasso.fit()</code> to allow adjusting clinical covariates.
iSIS-LASSO	An iterative SIS-LASSO Cox procedure with the conservative (intersection) variant was applied (Fan et al., 2009), consisting of repeated cycles of covariate-adjusted screening and penalized model fitting. At each iteration, the sample was randomly split into two halves, and metabolites were ranked within each half using partial likelihood from Cox models adjusted for tumor stage and age at diagnosis; previously selected metabolites were included as conditioning variables from the second iteration onward. The top m metabolites from each half were selected, increasing m until their intersection contained $d = \left\lfloor \frac{n}{\log(n)} \right\rfloor$, with smaller values (e.g. $d = \left\lfloor \frac{n}{2\log(n)} \right\rfloor$,) used if convergence issues arose (De Bin et al., 2020; Jardillier et al., 2022). The intersected metabolites were then entered into a LASSO Cox model (tumor stage and age at diagnosis unpenalized), and iterations (up to 50) continued until the selected set stabilized or the target size was reached.

Table S1.2: Calibration assessment procedure

Step	Description
(i) Individual-level expected event estimation	For each final model, we computed individual risk scores, using the linear predictor (lp_i). To estimate expected event incidences, we utilized the hazard function $h_i(t)$, which represents the instantaneous rate of event occurrence at time t conditional on survival up to that time. Specifically, we fit an offset Cox model, $h_i(t) = h_0(t) \exp(lp_i)$, where the relative risk $\exp(lp_i)$ remained fixed to allow for the estimation of the baseline hazard $h_0(t)$. This was implemented using the R command <code>survival::coxph(y ~ offset(lp))</code> . Based on this fit, we calculated the expected event incidences e_i for each subject, equivalent to the cumulative hazard over their respective follow-up period, using <code>stats::predict(fit, type = "expected")</code> .
(ii) Risk grouping	Subjects were ranked according to lp_i and divided into quartiles. Quartiles were chosen to reflect the limited number of observed events while ensuring sufficient expected and observed counts within each group.
(iii) Group-level observed and expected counts	Within each risk group g, observed event counts were calculated as $O_g = \sum_{i \in g} \delta_i,$ where δ_i is the event indicator. Group-level expected counts were obtained by aggregating individual expected values: $E_g = \sum_{i \in g} e_i.$
(iv) Poisson calibration modeling and goodness-of-fit test	Under perfect calibration, $O_g \approx E_g$ (i.e., $O_g/E_g \approx 1$). Calibration was formally evaluated using a Poisson regression model $O_g \sim \text{Poisson}(\mu_g), \quad \log(\mu_g) = \log(E_g) + \gamma_g,$ with $\log(E_g)$ included as an offset (Atkinson et al., 2008; Berry, 1983). In R, this was implemented using <code>glm(status ~ -1 + group + offset(log(e)), family="poisson")</code>. In this parameterization, $\exp(\gamma_g) = O_g/E_g$ represents the group-specific calibration factor (or standardized incidence ratio). Values equal to 1 indicate perfect calibration; values greater than or less than 1 indicate systematic underestimation or overestimation of risk, respectively. Overall calibration was assessed using a likelihood ratio goodness-of-fit test comparing models with and without group effects. A non-significant p-value indicates no evidence of systematic miscalibration across risk strata (Crowson et al., 2016).

Supplement 1b: Metabolite Set Enrichment Analysis

For metabolite set enrichment analysis, we constructed two model-specific input lists: 1) a core three-metabolite panel (2-hydroxyisobutyrate(HMDB0000729), indolelactate(HMDB0000671), uridine(HMDB0000296)), which includes metabolites that were repeatedly selected by the LASSO, adaptive LASSO, and iSIS-LASSO models; 2) the spike-and-slab LASSO model derived list (2-hydroxyisobutyrate , valine (HMDB0000883), xylitol(HMDB0242149), indolelactate, uridine). For enrichment analysis, uridine was used as a proxy for the uridine-like metabolite. Although this feature was annotated to Level 3 (putatively characterized compound class), its structural signature allowed for the selection of a representative identifier to facilitate mapping to established metabolic pathways. In contrast, two sugar acids (Sugar.Acid.16.847min and Sugar.Acid.18.634min) selected by the iSIS-LASSO and spike-and-slab LASSO models, respectively, were excluded from the MSEA. Unlike the uridine-like feature, these sugar acids lacked sufficient structural specificity to be assigned a unique representative identifier, making them ineligible for exact mapping in the MetaboAnalyst. The corresponding results based on the two specific input lists are presented in **Figure S1.1** and **Figure S1.2**.

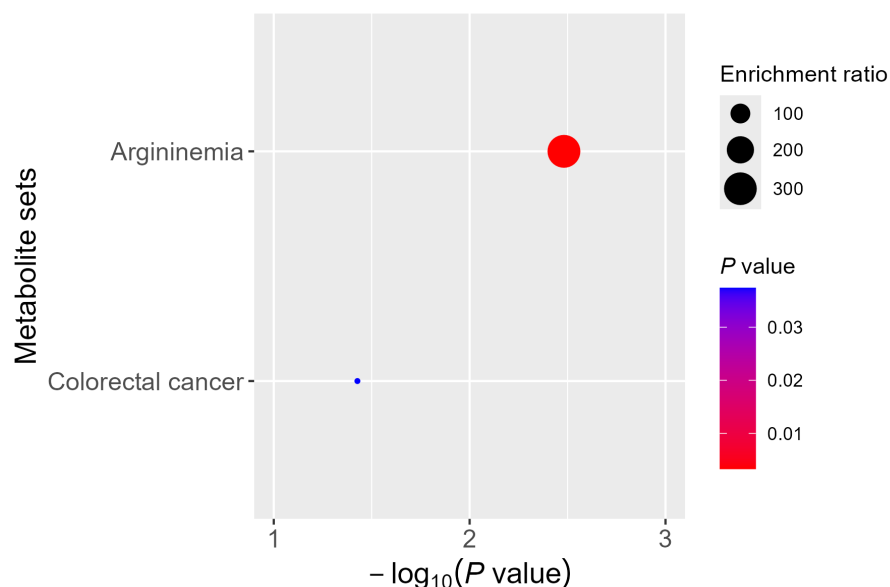


Fig S1.1: Enriched metabolite sets based on selected metabolites (Models: LASSO, adaptive LASSO, and iSIS-LASSO)

Figure S1.1. Metabolite set enrichment based on the core metabolite panel. Enrichment analysis was performed in MetaboAnalyst using the three metabolites (2-hydroxyisobutyrate, indolelactate, uridine) consistently selected by LASSO, adaptive LASSO, and iSIS-LASSO. An additional metabolite identified by iSIS-LASSO could not be mapped in MetaboAnalyst and was therefore excluded from the enrichment analysis. The x-axis shows $-\log_{10}(P)$ and the y-axis lists enriched metabolite sets. Bubble size is proportional to the enrichment ratio and color indicates the nominal P value. This panel yielded nominal enrichment of “Argininemia” and “Colorectal cancer”, but neither remained significant after multiple-testing correction.

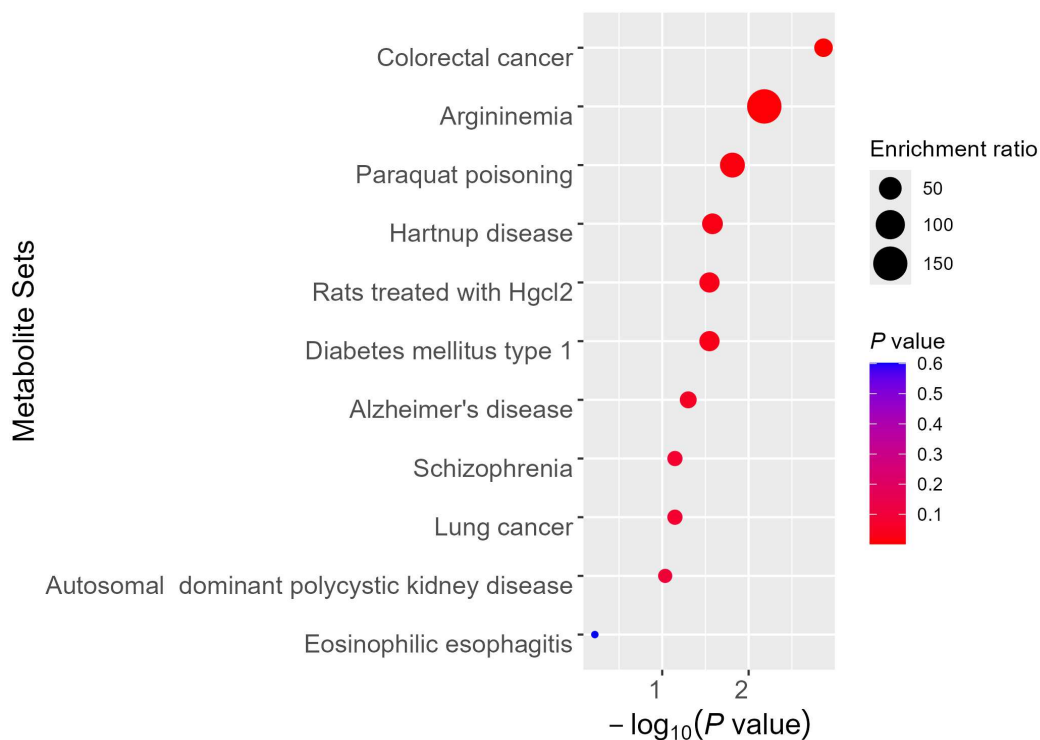


Fig S1.2: Enriched Metabolite Sets based on selected metabolites (Model: spike-and-slab LASSO)

Figure S1.2. Metabolite-set enrichment based on the spike-and-slab LASSO derived metabolite panel. Enrichment analysis was conducted using the five metabolites identified by the spike-and-slab LASSO model (2-hydroxyisobutyrate, valine, xylitol, indolelactate, and uridine). The x-axis shows $-\log_{10}(P)$ and the y-axis lists enriched metabolite sets. Bubble size reflects the enrichment ratio and color encodes the nominal P value. This panel showed broader but largely nominal enrichment across several disease-related sets including “Colorectal cancer”, however, no sets remained significant after multiple-testing correction.

Supplement 2: Sensitivity Analyses

In sensitivity analysis, we evaluated model performance with additionally adjustment for tumor site (colon vs. rectum), given accumulating evidence suggesting an association between tumor site and OS in CRC. For this purpose, tumor site was included as an additional unpenalized covariate in the clinical reference model and in each of the four penalized Cox models. All other aspects of model specification, tuning, and validation procedures were identical to those used in the primary analyses.

Model discrimination

In both the primary and tumor site adjusted sensitivity analyses, all penalized models demonstrated better discrimination than the clinical reference model in terms of the C-index, iAUC and tAUC (Figure S2.1).

When ranking models based on overall discrimination metrics (C-index and iAUC), spike-and-slab LASSO and adaptive LASSO consistently performed among the top methods in both analyses, with only modest differences between penalized models. The reference model showed the lowest performance across both metrics.

When comparing tAUCs among the penalized models, we focused on summary characteristics rather than pointwise rankings at individual time horizons. Across both the primary and tumor site adjusted sensitivity analyses, all penalized models achieved comparable peak tAUC values, with spike-and-slab LASSO and adaptive LASSO generally attaining the highest peak discrimination, followed closely by standard LASSO and iSIS-LASSO. Differences in peak tAUC across penalized methods were modest. Similarly, average tAUC performance across follow-up periods was broadly comparable among the penalized models. Although the relative ordering of tAUCs varied at

specific follow-up times and between analyses, the overall magnitude and temporal patterns of tAUCs were similar across penalized models and consistently exceeded those of the reference model.

Final model calibration

Across both the primary and tumor site-adjusted sensitivity analyses, all models showed broadly similar calibration patterns, with observed and expected event rates increasing across predicted risk groups, and agreement observed across all risk groups (Figure S2.2, Table S2.2). Calibration was largely preserved after additional adjustment for tumor site, with no systematic deterioration across modeling approaches.

In sensitivity analysis additionally adjusting for tumor site, iSIS-LASSO model, calibration was assessed using a median split rather than quartiles because the risk scores derived from this model yielded insufficient distinct values to support stable quartile-based grouping. Observed and expected event rates in the resulting two risk groups were comparable, indicating no substantial loss of calibration.

Selected features and corresponding coefficients in the final models

Across both the primary and tumor site adjusted sensitivity analyses, tumor stage and age at diagnosis were included in all models and exhibited highly stable coefficient estimates across modeling approaches. Tumor stage consistently showed a strong association with OS, with nearly identical effect sizes before and after tumor site adjustment, while age at diagnosis demonstrated smaller, directionally consistent effects with minimal change in magnitude.

Among the penalized models, the core set of selected metabolites remained largely consistent

after adjustment for tumor site. A uridine-like metabolite and indolelactate were retained across all penalized models in both analyses, with only modest attenuation of their coefficient magnitudes after tumor site adjustment, suggesting limited confounding by tumor site. Metabolites such as xylitol and valine, which were selected by the spike-and-slab LASSO, also showed directionally consistent effects with slight attenuation after tumor site adjustment. Adjustment for tumor site led to the selection of a small number of additional metabolites in certain penalized models (e.g., Threonate, Tartrate, and Myo-inositol were selected by the iSIS-LASSO model), but these changes were model specific and did not substantially alter the dominant metabolomic signals or their directions of association.

Together, the stability of feature selection, coefficient estimates, and relative effect sizes before and after tumor-site adjustment supports the robustness of the identified metabolomic predictors and indicates that their associations with OS are largely independent of tumor site. Notably, the estimated effect of tumor site itself was modest across all models and substantially smaller than the effects of the consistently selected metabolites: a uridine-like metabolite and Indolelactate.

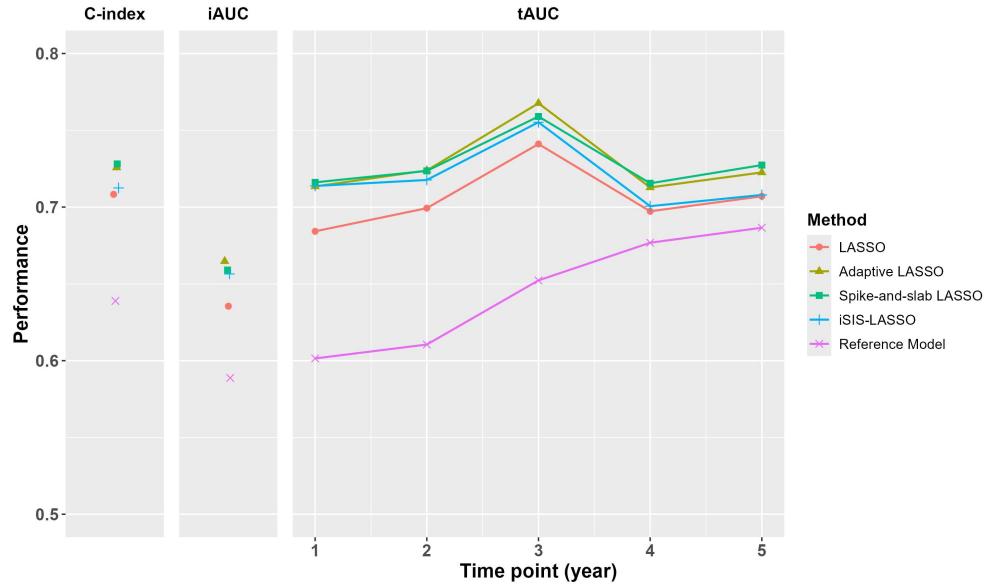


Fig S2.1: Model discrimination assessment (with additional adjustment for tumor site)

Figure S2.1. Discrimination performance of models (with additional adjustment for tumor site). Performance metrics were evaluated using nested cross-validation. Panels (left to right) display the mean performance across folds for: the concordance index (C-index), the integrated AUC (iAUC) over the total follow-up period, and the time-dependent AUC (tAUC) at 1st, 2nd, 3rd, 4th, and 5th year. Models compared include LASSO, adaptive LASSO, spike-and-slab LASSO, iSIS-LASSO, and a clinical reference model. Higher values represent superior discrimination performance. All penalized models continue to outperform the clinical reference model and no substantive changes in relative model performance ranking.

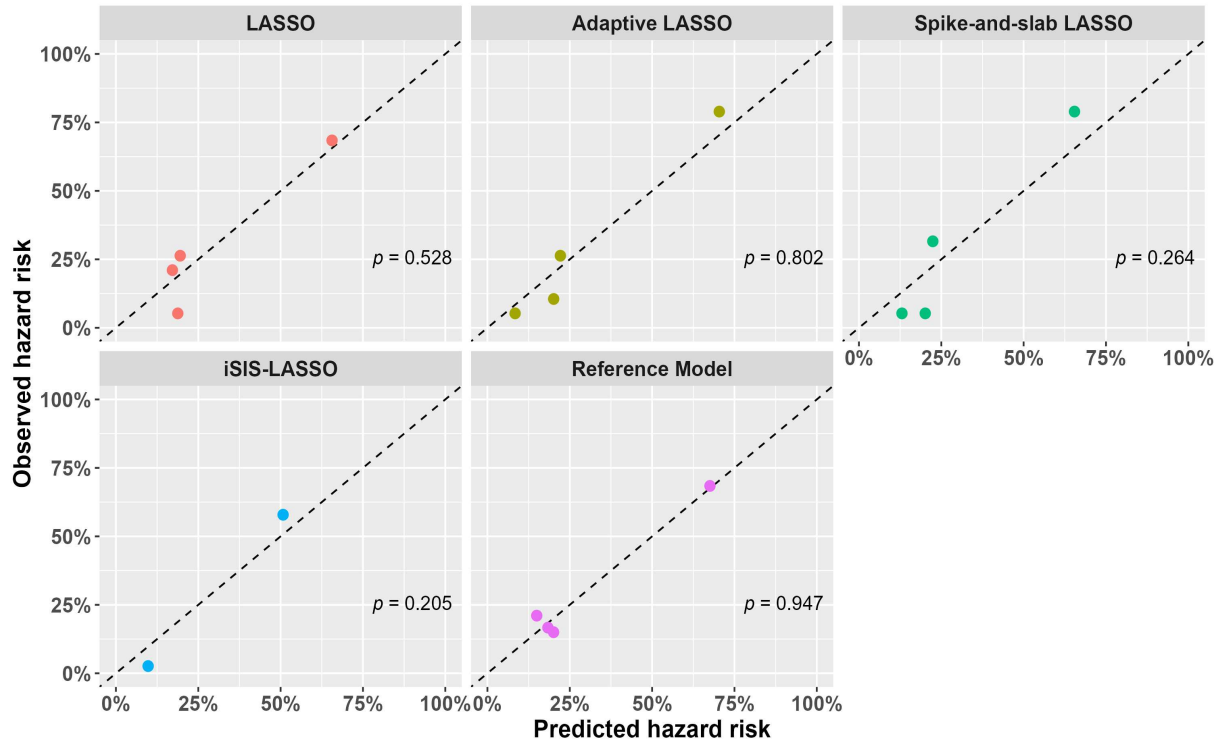


Fig S2.2: Model calibration assessment (with additional adjustment for tumor site)

Figure S2.2. Calibration of final prognostic models (with additional adjustment for tumor site) fit on all study participants. Each panel shows, for each modeling approach (LASSO, adaptive LASSO, spike-and-slab LASSO, iSIS-LASSO, and the clinical reference model), the observed vs. predicted hazard risk across patient risk groups, where groups are defined by quartiles of the model's risk score (linear predictor). For the iSIS-LASSO model, calibration was assessed using a median split rather than quartiles due to limited variability in predicted risk scores. The dashed diagonal line indicates perfect calibration, and the p-value in each panel is from a goodness-of-fit test comparing predicted and observed risks across quartiles of risk. Observed and expected event rates showing similar agreement across predicted risk groups and no evidence of systematic miscalibration after tumor site adjustment.

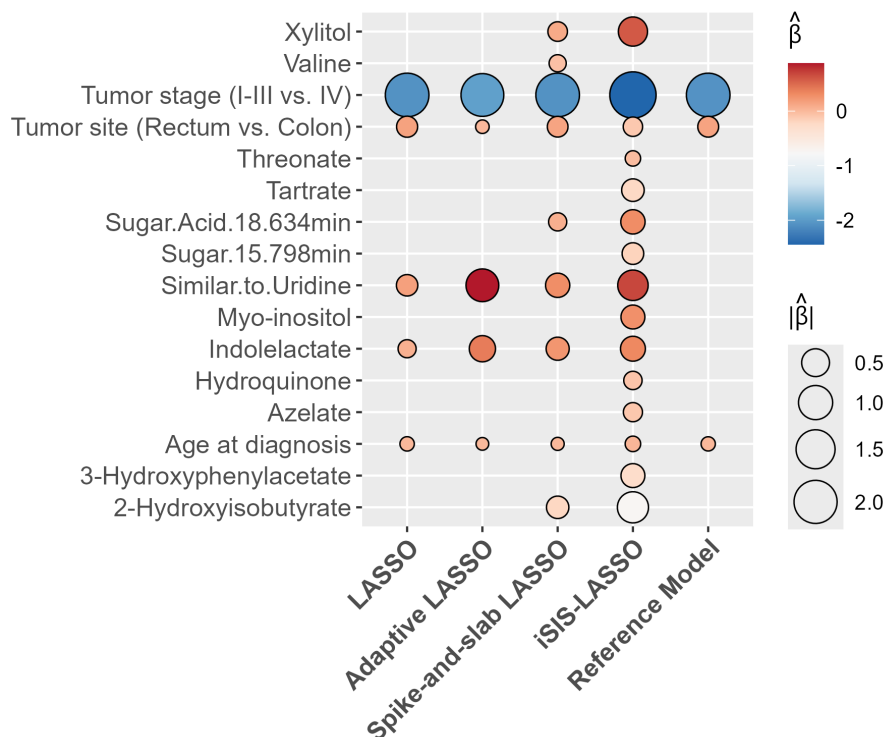


Fig S2.3: Model coefficient estimation from final models (with additional adjustment for tumor site)

Figure S2.3. Model coefficients of final prognostic models (with additional adjustment for tumor site) fit on all study participants. Columns correspond to modeling approaches (LASSO, adaptive LASSO, spike-and-slab LASSO, iSIS-LASSO and the clinical reference model), and rows represent to clinical and metabolite predictors in corresponding models. Each bubble indicates that a corresponding variable is included in the final model, with fill color representing the sign and magnitude of the estimated log-hazard ratio ($\hat{\beta}$; red = positive, blue = negative) and bubble size proportional to $|\hat{\beta}|$. Across penalized models, the core set of metabolites was largely preserved after adjustment for tumor site. A uridine-like metabolite and indolelactate were consistently selected in both analyses, across all penalized approaches, with stable effect directions and only modest attenuation of coefficient magnitudes after tumor-site adjustment, indicating limited confounding by tumor site.

Table S2.1: Model coefficients for selected features (with additional adjustment for tumor site)

Feature	LASSO	adpative LASSO	Spike-and-slab	iSIS-LASSO	Reference Model
Xylitol	–	–	0.107	0.594	–
Valine	–	–	-0.047	–	–
Tumor stage (I–III vs. IV)	-2.077	-1.961	-2.076	-2.444	-2.071
Tumor site (Rectum vs. Colon)	0.159	0.006	0.145	-0.101	0.150
Threonate	–	–	–	-0.021	–
Tartrate	–	–	–	-0.217	–
Sugar.Acid.18.634min	–	–	0.069	0.299	–
Sugar.15.798min	–	–	–	-0.179	–
Similar.to.Uridine	0.172	0.876	0.296	0.688	–
Myo-inositol	–	–	–	0.279	–
Indolelactate	0.067	0.401	0.241	0.329	–
Hydroquinone	–	–	–	-0.074	–
Azelate	–	–	–	-0.094	–
Age at diagnosis	0.011	0.005	0.005	0.022	0.010
3-Hydroxyphenylacetate	–	–	–	-0.279	–
2-Hydroxyisobutyrate	–	–	-0.211	-0.742	–

Note: The reference model includes only tumor stage, age at diagnosis and tumor site; penalized models are adjusted for tumor stage, age at diagnosis and tumor site.

Table S2.2: Model calibration assessment – observed and expected event rates across predicted risk groups for each modeling approach (with additional adjustment for tumor site)

Method	Quartile 1 (O / E)	Quartile 2 (O / E)	Quartile 3 (O / E)	Quartile 4 (O / E)
LASSO	0.053 / 0.188	0.211 / 0.171	0.263 / 0.195	0.684 / 0.656
Adaptive LASSO	0.053 / 0.084	0.105 / 0.201	0.263 / 0.221	0.789 / 0.704
Spike-and-slab LASSO	0.053 / 0.131	0.053 / 0.201	0.316 / 0.224	0.789 / 0.654
iSIS-LASSO	0.026 / 0.098		0.579 / 0.508	
Reference Model	0.150 / 0.201	0.211 / 0.150	0.167 / 0.184	0.684 / 0.676

Note: O/E denotes observed (O) and expected (E) event rates within predicted risk groups; E was calculated as the mean predicted event probability within each group. For LASSO, adaptive LASSO, spike-and-slab LASSO, and the reference model, predicted risks were grouped into quartiles. For iSIS-LASSO, predicted risks were grouped using a median split; values are shown centered across the first two and last two columns to reflect the two-group structure.

Supplement 3 - Simulation Study

Simulation steps

We conducted a simulation study to evaluate the performance of metabolomics-based penalized survival models under varying signal strength (three levels), multicollinearity among predictors (two levels), and censoring rates (three levels). Each scenario consists of 200 iterations. This simulation study aims to evaluate the robustness and finite-sample properties of the penalized Cox models under the empirical correlation structure of the metabolomics data. The main components of simulation procedure are summarized below.

(i) **Design matrix.** We fixed the covariate structure by using the empirical design matrix $\mathbf{X}$ from the ColoCare Study, with $n = 76$ patients (rows) and $p = 121$ metabolites (columns). This preserves the real-world correlation structure among metabolites.

(ii) **Construction of true signal sets.** To study the impact of multicollinearity, we predefined two disjoint sets of true signal metabolites:

- a *high-multicollinearity* set consisting of 3 metabolites with an average variance inflation factor (VIF) of 6.43;
- a *low-multicollinearity* set consisting of 3 metabolites with an average VIF of 1.03.

These sets were treated as the true prognostic metabolites in different simulation scenarios.

(iii) **Specification of true coefficients.** For each scenario, nonzero regression coefficients were assigned to the true signals on the log-hazard scale. We considered three target effect sizes $\beta \in \{0.2, 0.4, 0.7\}$ to approximate modest, moderate, and strong effects:

- $\exp(0.2) \approx 1.22$ (about a 20% increase in hazard)
- $\exp(0.4) \approx 1.49$ (about a 50% increase in hazard)
- $\exp(0.7) \approx 2.01$ (about a 100% increase in hazard)

For a given target level β , each true nonzero coefficient β_j was drawn independently from a uniform distribution

$$\beta_j \sim \text{Unif}(0.8 \beta, 1.2 \beta),$$

introducing mild variation in each iteration for effect sizes within each scenario. The sign of each nonzero coefficient was assigned to be consistent with the observed correlation structure among the signal metabolites (i.e., positively correlated signals were given the same sign, whereas negatively correlated signals were assigned opposite signs). All remaining metabolites were assigned true coefficients of zero.

(iv) **Generation of survival outcomes.** Given the design matrix $\mathbf{X}$ and the specified true coefficient vector, we generated right-censored survival data under three censoring levels: 30%, 50%, and 70%. Each combination of signal set (high vs. low multicollinearity), effect size, and censoring rate defined a distinct simulation scenario.

(1) Control of simulated median survival time (target 6 years)

We generated event times from a parametric proportional hazards model with a Weibull baseline. For subject i with covariate vector $\mathbf{X}_i$, the hazard function was

$$h(t \mid \mathbf{X}_i) = h_0(t) \exp(\mathbf{X}_i^\top \boldsymbol{\beta}),$$

where the baseline hazard followed a Weibull form

$$h_0(t) = \lambda \gamma t^{\gamma-1},$$

and the corresponding survival function is

$$S(t \mid \mathbf{X}_i) = \exp\{-\lambda t^\gamma \exp(\mathbf{X}_i^\top \boldsymbol{\beta})\},$$

with shape parameter fixed at $\gamma = 1.5$, reflecting increase of hazard (risk) over time.

Event times were drawn using inverse transform sampling. Let $U_i \sim \text{Unif}(0, 1)$, then

$$T_i = \left(-\frac{\log U_i}{\lambda \exp(\mathbf{X}_i^\top \boldsymbol{\beta})} \right)^{1/\gamma}.$$

To control the median survival time (reflecting the observed data: median survival time is 6 years), we chose the Weibull scale parameter λ so that the median survival equals a prespecified value m_i for each subject. Since

$$S(m_i \mid \mathbf{X}_i) = \exp\{-\lambda m_i^\gamma \exp(\mathbf{X}_i^\top \boldsymbol{\beta})\},$$

setting $S(m_i \mid \mathbf{X}_i) = 0.5$ implies

$$\lambda = \frac{\log 2}{m_i^\gamma \exp(\mathbf{X}_i^\top \boldsymbol{\beta})}.$$

In our simulations we fixed $m_i \equiv 6$ years for all subjects, so that the generated survival

times had a target median of 6 years across scenarios while preserving relative risk contrasts through β .

(2) Control of censoring rate

Independent time to event T_i and right-censoring times C_i were generated from a parametric distribution governed by a scalar tuning parameter θ (e.g., an exponential distribution with rate θ). Let the observed time and event indicator were

$$Y_i = \min(T_i, C_i), \quad \Delta_i = \mathbf{1}\{T_i \leq C_i\}.$$

Our goal was to choose θ to achieve a target censoring proportion $p^* \in \{0.3, 0.5, 0.7\}$. For a given value of θ , the empirical censoring rate is

$$\hat{p}(\theta) = \frac{1}{n} \sum_{i=1}^n \mathbf{1}\{C_i < T_i\} = \frac{1}{n} \sum_{i=1}^n (1 - \Delta_i).$$

We then searched for

$$\theta^* = \arg \min_{\theta > 0} |\hat{p}(\theta) - p^*|$$

using a simple binary search, exploiting the fact that $\hat{p}(\theta)$ increases monotonically with θ :

- (a) Choose bounds $0 < \theta_L < \theta_H$ such that $\hat{p}(\theta_L) \leq p^* \leq \hat{p}(\theta_H)$.
- (b) At each iteration, set $\theta_M = (\theta_L + \theta_H)/2$, regenerate censoring times C_i and recompute $\hat{p}(\theta_M)$.
- (c) If $\hat{p}(\theta_M) > p^*$ (too much censoring), update $\theta_H \leftarrow \theta_M$; otherwise update $\theta_L \leftarrow \theta_M$.

(d) Stop when $|\hat{p}(\theta_M) - p^*| \leq \varepsilon$ for a small tolerance ε (e.g. $\varepsilon = 0.01$), and take $\theta^* \approx \theta_M$.

This procedure yielded simulated datasets with the desired censoring proportions (30%, 50%, and 70%) while keeping the event-time distribution fixed across censoring scenarios.

(v) **Methods compared.** Across all scenarios, we applied the following methods with consistent model specification described in the method section of the main body of the manuscript:

- LASSO;
- adaptive LASSO;
- spike-and-slab LASSO;
- iSIS-LASSO.

(vi) **Performance metrics.** For each method and scenario, we evaluated:

- *Prediction performance:* Harrell's C-index, time-dependent AUC at 1st, 2nd, 3rd, 4th, and 5th year, and integrated AUC (iAUC) across all follow up time;
- *Feature selection accuracy:* recovery of the full true set (all true signals selected and no noise variables), F_1 score and recall (or true positive rate).
- *Coefficient estimation error:* the ℓ_2 -Norm of the difference between estimated and true coefficients, evaluated both on the full coefficient vector and restricted to the true support.

Simulated data quality check

To confirm that the simulated data reflected the prespecified simulation design, we performed several data quality checks across all scenarios. First, the absolute values of the six true log-hazard coefficients clustered tightly around their target magnitudes (0.2, 0.4, and 0.7) under both high- and low-multicollinearity settings, confirming that the signal strength was well controlled (Figure S3.1). Second, Kaplan–Meier estimates of the median survival time were centered around the target of 6 years for the 30% and 50% censoring scenarios, with greater dispersion only under 70% censoring, as expected when events are reduced (Figure S3.2). Third, the realized censoring proportions closely matched the prespecified levels (30%, 50%, 70%) across all effect sizes and correlation structures (Figure S3.3). Together, these checks indicate that the simulated datasets faithfully reflect the planned effect size, median survival time, and censoring rate.

Predictive model performance in the simulation study

In terms of model discrimination, the oracle model (fit with knowledge of the true signals) consistently achieved the highest C-index, iAUC and tAUC across effect sizes and censoring levels, as expected (Figure S3.4, S3.5, S3.6). For all methods, discrimination improved with increasing effect size and declined modestly as censoring increased, with broadly similar patterns under high and low multicollinearity. Across the four penalized approaches, differences in predictive performance were generally small in magnitude. Under low censoring (30%), spike-and-slab LASSO showed the highest discrimination, whereas under heavier censoring ($\geq 50\%$) with moderate-to-strong signals ($\beta > 0.4$), iSIS-LASSO most often exhibited the highest discrimination. When discrimination was examined at specific follow-up horizons, tAUC increased with follow-up time, peaked around 3 or 4 years, and then declined, with heavier censoring

modestly reducing performance. At 2–4 years, spike-and-slab LASSO tended to yield the highest tAUCs under 30% censoring, whereas iSIS-LASSO most often achieved the highest tAUCs when censoring exceeded 50%.

Feature-selection accuracy showed clearer separation between methods (Figure S3.7, S3.8, S3.9). Overall, the F_1 scores increased with larger effect size, lower multicollinearity and lower censoring, but remained far from perfect even in favorable settings. The adaptive LASSO model consistently outperformed all other penalized models when censoring rate is ≤ 0.5 , while the spike-and-slab LASSO performed the best when censoring rate is high (0.7) and under low multicollinearity. The iSIS-LASSO yielded the lowest F_1 scores, while achieved the best recall (also known as true positive rate or sensitivity) across most scenarios. Full-set recovery was essentially zero for all methods at the smallest effect size ($\beta = 0.2$), but increased at $\beta = 0.4$ and $\beta = 0.7$, particularly with low censoring and low multicollinearity. In these stronger-signal settings, iSIS-LASSO had the largest probability of recovering the complete true set, whereas LASSO, adaptive LASSO and spike-and-slab LASSO showed more modest recovery rates, which deteriorated as censoring increased.

Penalized methods showed comparable l_2 -Norm errors, with larger variability for procedures that rely on aggressive iterative screening (e.g., iSIS-LASSO) when signals were weak or censoring was heavy (Figure S3.10, S3.11).

While the penalized methods achieve similar discrimination, iSIS-LASSO provides the most reliable recovery of the true signals, at the cost of more variability in coefficient estimation and more false discovery rate, especially under challenging high-censoring or highly collinear settings.

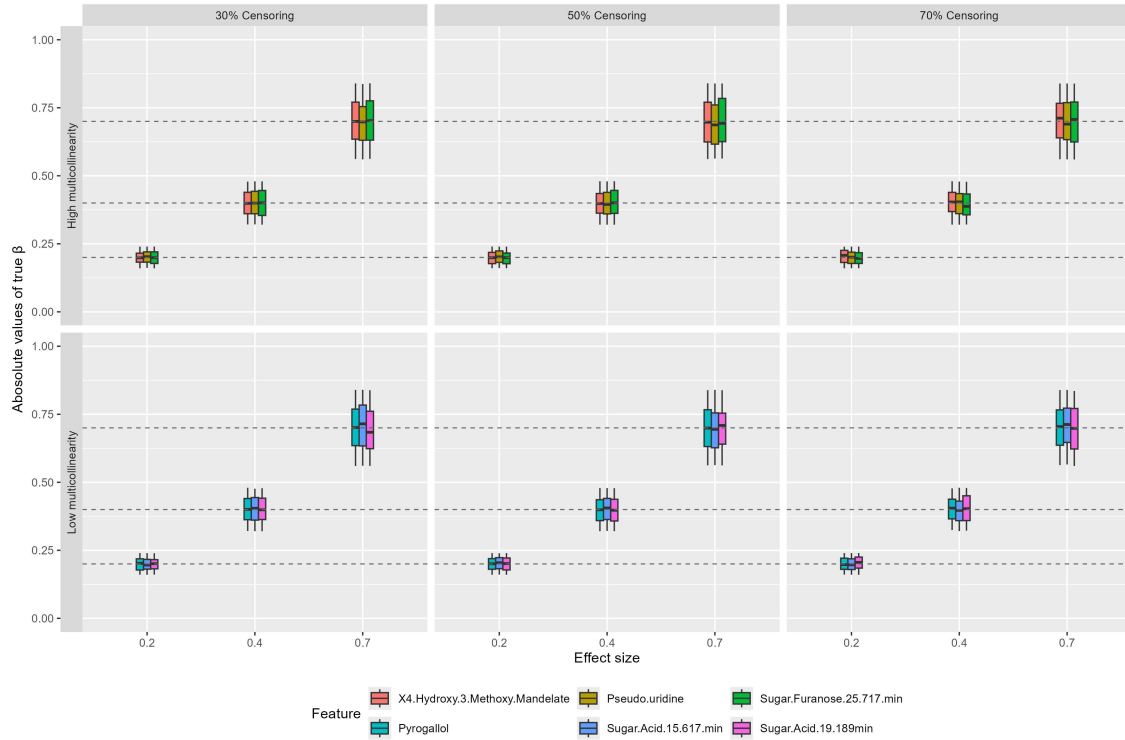


Fig S3.1: Distribution of simulated model coefficients

Figure S3.1. Distributions of the absolute value of true coefficients in log-hazard scale used in the simulation, by target effect size, multicollinearity setting, and censoring level. Each boxplot summarizes the distribution of 200 simulated draws of β for the three high-multicollinearity signals (top panels) and three low-multicollinearity signals (bottom panels) under effect sizes 0.2, 0.4, and 0.7 (x-axis), for censoring rates of 30%, 50%, and 70% (columns). Different colors denote individual metabolites within each set, and horizontal dashed lines indicate the nominal target effect sizes, showing that the generated coefficients are centered around the intended magnitudes across all scenarios.

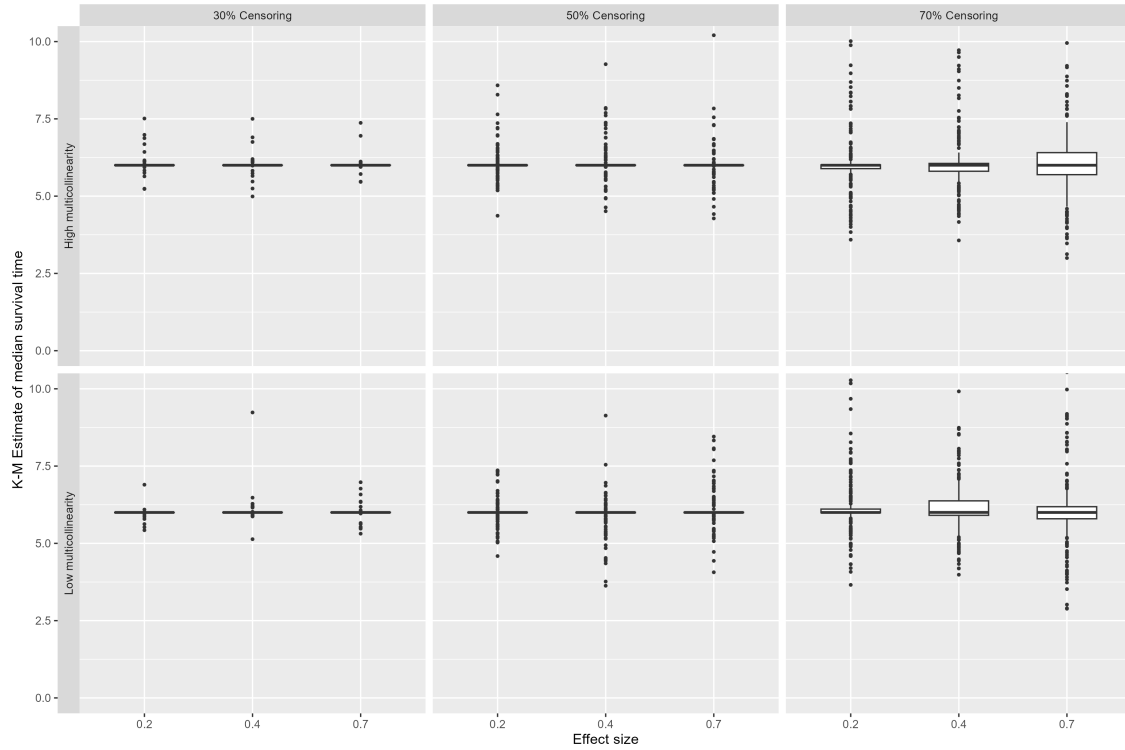


Fig S3.2: Distribution of median survival time of simulated survival data

Figure S3.2. Kaplan–Meier estimates of the median survival time across simulation settings. Each panel shows the distribution of simulated survival data over 200 replicates by effect size (0.2, 0.4, 0.7 on the x-axis) and censoring level (30%, 50%, 70% in columns), stratified by high vs. low multicollinearity of the true signal set (rows). Across all scenarios, the empirical medians cluster around the target value of 6 years, indicating successful control of the underlying Weibull baseline; variability increases with higher censoring, especially under 70% censoring, but without systematic drift away from the intended median.

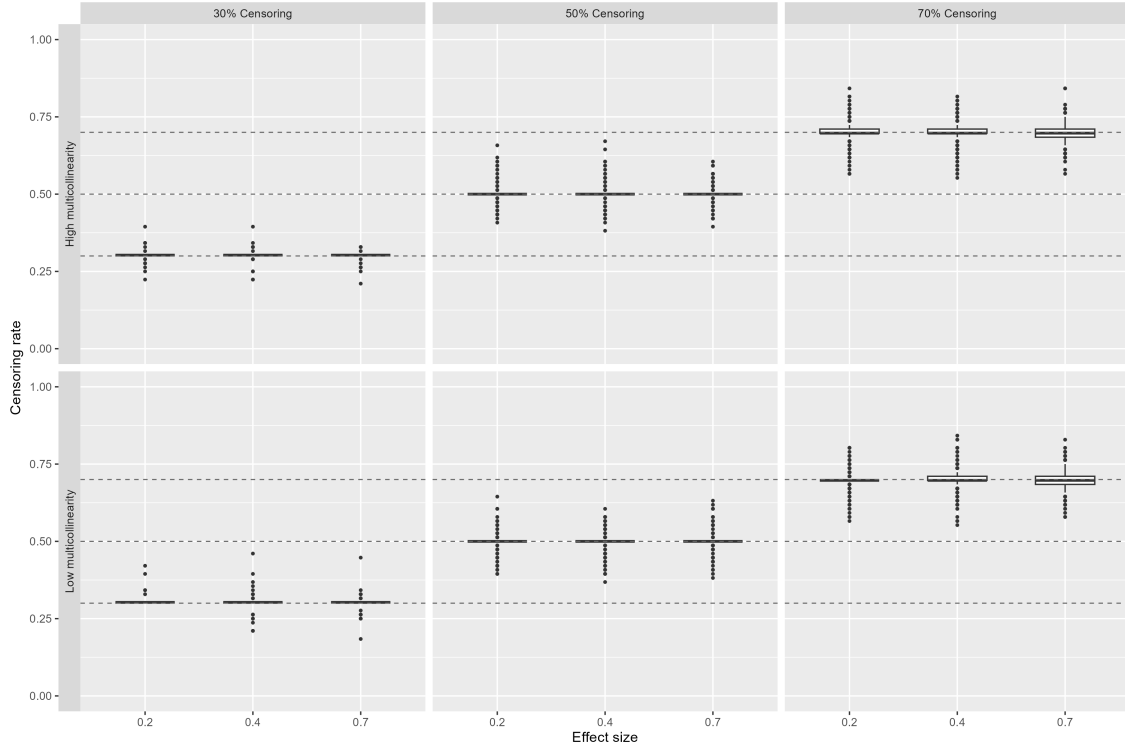


Fig S3.3: Distribution of censoring of simulated survival data

Figure S3.3. Empirical censoring rates across simulation scenarios. Each panel shows the distribution of observed censoring proportions over 200 simulated datasets by effect size (0.2, 0.4, 0.7 on the x-axis) and target censoring level (30%, 50%, 70% in columns), stratified by high vs. low multicollinearity of the true signals (rows). Horizontal dashed lines indicate the target censoring rates. Across all settings, the realized censoring proportions cluster closely around their targets with only modest variability, confirming that the binary-search achieved the desired censoring levels regardless of effect size or correlation structure.

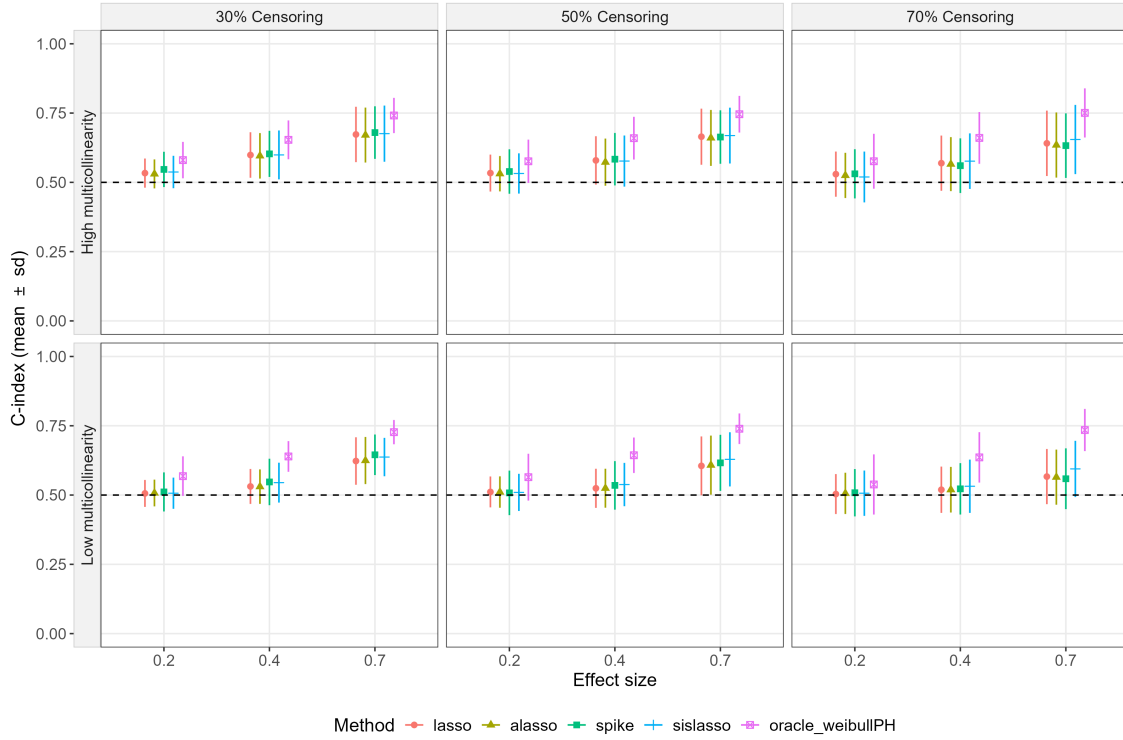


Fig S3.4: Model performance in simulation study: C-index

Figure S3.4. Mean C-indices (points) and simulation standard deviations (bars) for each method across effect sizes (0.2, 0.4, 0.7), censoring levels (30%, 50%, 70%; columns), and correlation structures (high vs. low multicollinearity; rows). The dashed line indicates no better than random prediction (C-index = 0.5). As expected, discrimination increased with larger effect sizes for all methods and was only modestly attenuated by heavier censoring. The oracle Weibull-PH model (fit with the true active set) provided the highest C-indices across all settings, representing an upper bound on achievable performance. The prediction performance did not differ significantly across all methods in the same scenario. Spike-and-slab LASSO had the highest performance at low censoring (30%). When higher censoring ($> 50\%$) and larger effect sizes (> 0.4) were present, iSIS-LASSO yielded the highest performance.

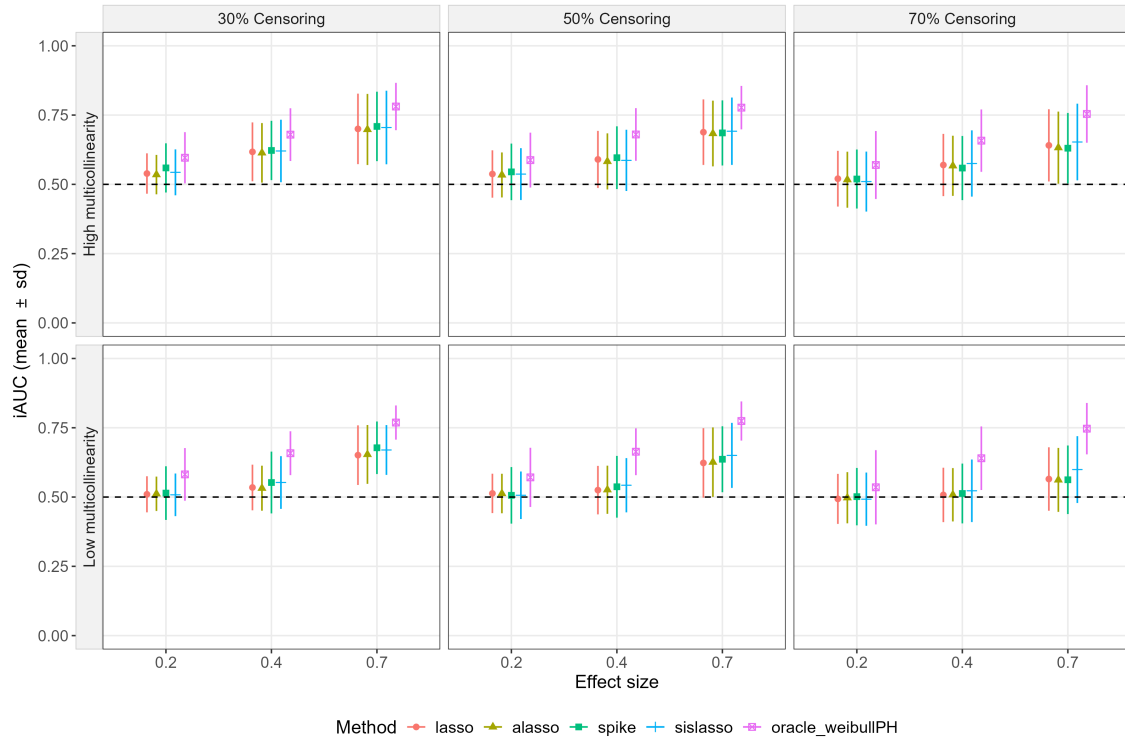


Fig S3.5: Model performance in simulation study: iAUC

Figure S3.5. Mean integrated time-dependent AUCs (iAUCs) and simulation standard deviations (sd; bars) for each method across effect sizes, censoring levels, and correlation structures. The dashed line corresponds to $iAUC = 0.5$. The patterns closely mirrored those for the C-index.

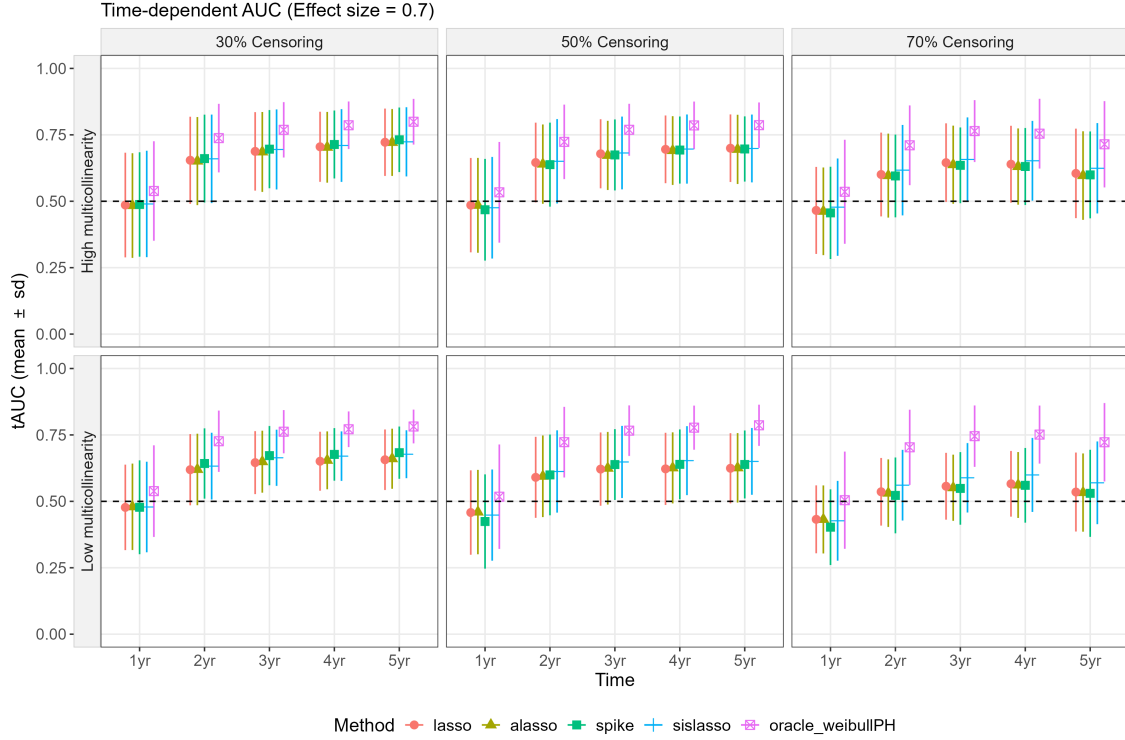


Fig S3.6: Model performance in simulation study: tAUC

Figure S3.6. Mean time-dependent AUCs (tAUCs; points) and simulation standard deviations (sd; bars) for each method at effect size 0.7 across censoring levels (30%, 50%, 70%; columns) and correlation structures (high vs. low multicollinearity; rows), evaluated at 1–5 years. The dashed line indicates no better than random discrimination (tAUC = 0.5). Discrimination was limited at 1 year but became evident after 2 years (consistent with results for $\beta < 0.7$). The performance increased with follow-up time, peaked around 3 or 4 years, and declined afterwards, with heavier censoring modestly reducing performance. At 2–4 years, spike-and-slab LASSO tended to perform best under 30% censoring, whereas when censoring exceeded 50%, iSIS-LASSO most often achieved the highest tAUCs. Overall differences among methods were small under the same scenario; the oracle Weibull-PH model (fit with true active set) provided an upper bound with consistently the highest tAUCs.

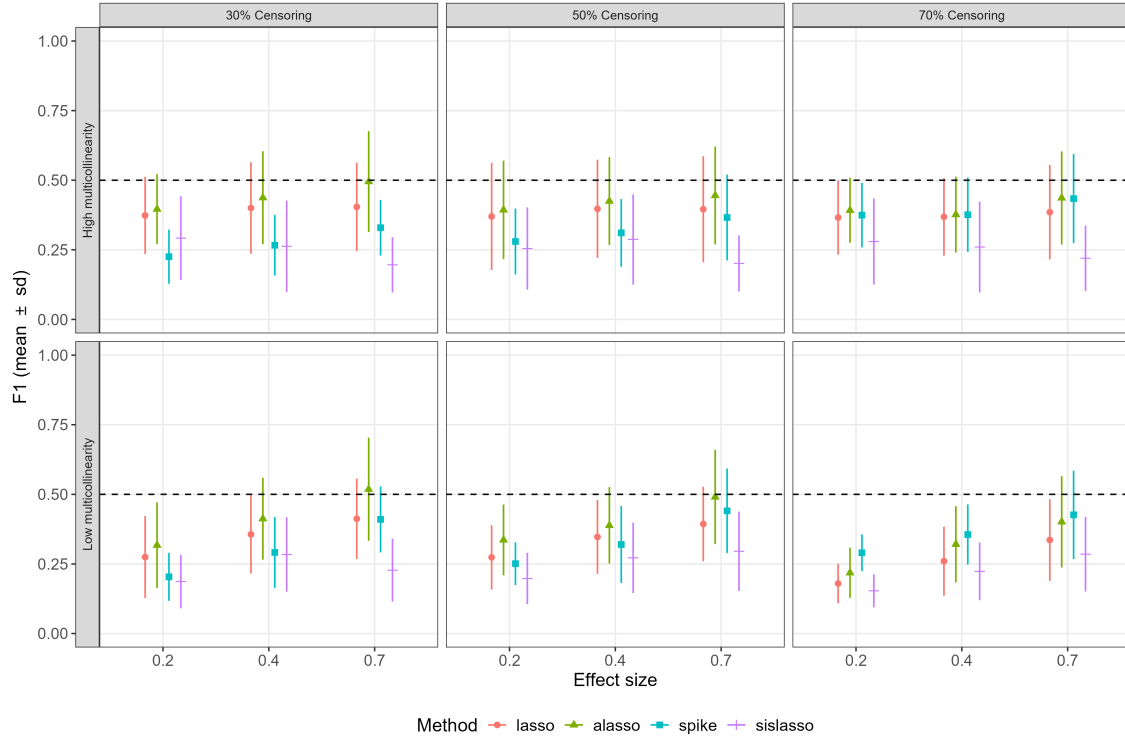


Fig S3.7: Model performance in simulation study: F_1 Score

Figure S3.7. F_1 -score for feature selection across simulation scenarios. Mean F_1 -scores (points) with ± 1 standard deviation (sd; bars) for recovery of the true signal set are shown by method, effect size (0.2, 0.4, 0.7), censoring rate (30%, 50%, 70%), and degree of multicollinearity (top row: high; bottom row: low). F_1 -scores increase with effect size and are generally higher under lower censoring and low multicollinearity. The adaptive LASSO consistently outperformed all other methods, whereas iSIS-LASSO yielded the worst performance across most scenarios, which might be related to its high false discovery rate.

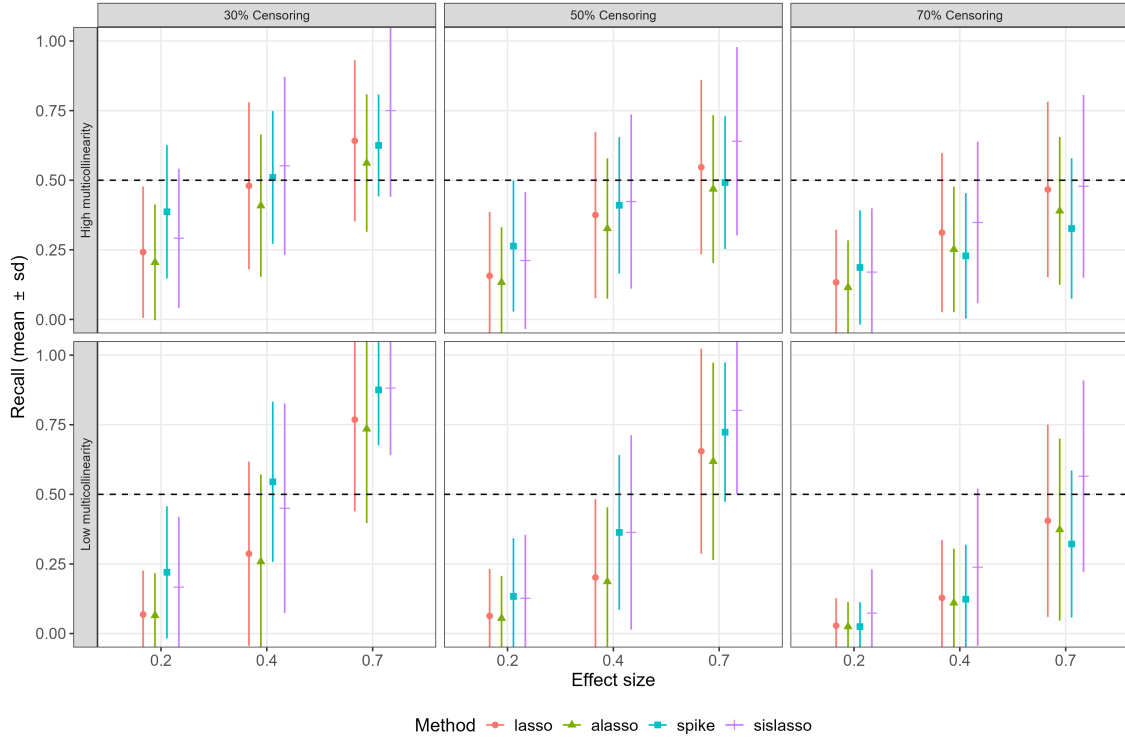


Fig S3.8: Model performance in simulation study: Recall of true signals

Figure S3.8. Recall of true signals across simulation scenarios. Mean recall (points) and simulation standard deviations (sd; bars) for each method across effect sizes (0.2, 0.4, 0.7), censoring levels (30%, 50%, 70%; columns), and correlation structures (high vs. low multicollinearity; rows). The dashed line marks a reference level of recall ($= 0.5$). Across all settings, recall increased with larger effect sizes, with little recovery at $\beta = 0.2$, moderate gains at $\beta = 0.4$, and the highest recall at $\beta = 0.7$. Heavier censoring generally reduced recall and increased variability, particularly under high multicollinearity. Spike-and-slab LASSO tended to achieve the highest recall under low censoring (30%), particularly with low multicollinearity, but its advantage diminished as censoring increased, whereas iSIS-LASSO most consistently delivered the top recall when $\beta \geq 0.4$ and censoring $\geq 50\%$.

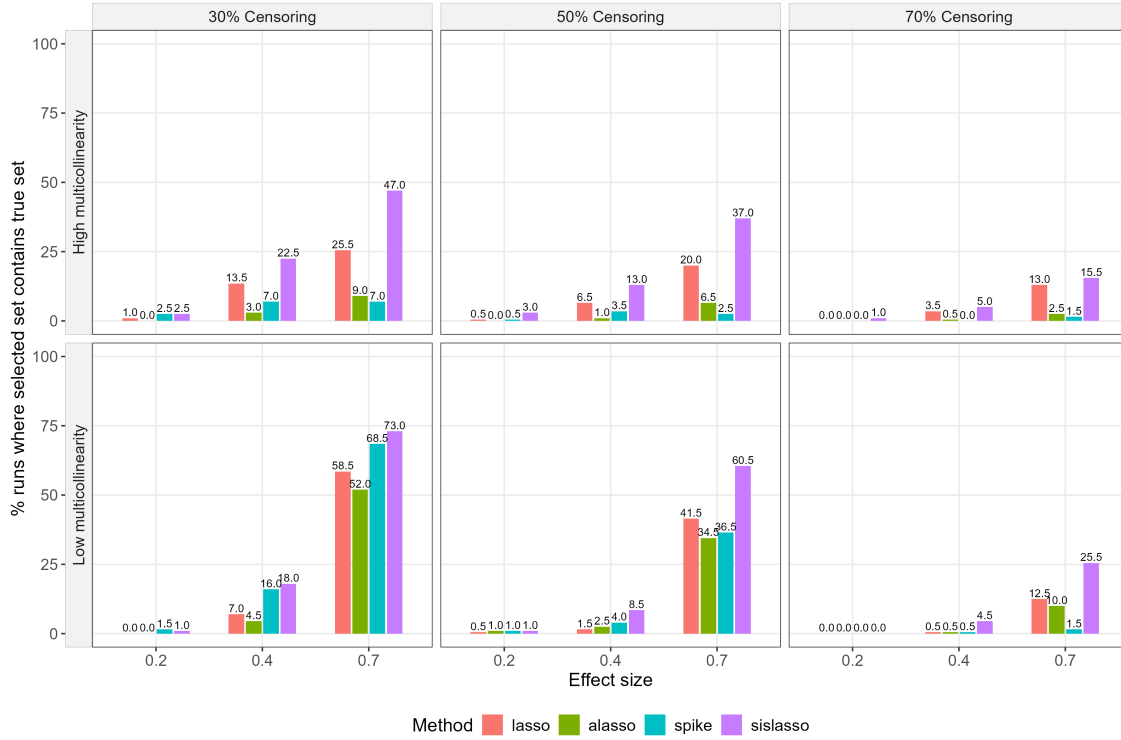


Fig S3.9: Model performance in simulation study: Proportion of full-set recovery

Figure S3.9. Full-set recovery of true signals. Bars show the percentage of simulation runs in which the selected feature set contained all three true signals (“full-set recovery”) for each method, stratified by effect size, censoring rate, and multicollinearity. Recovery is essentially zero for all methods at the smallest effect size ($\beta = 0.2$) and improves with larger effect sizes, especially under lower censoring and low multicollinearity. Across scenarios, iSIS-LASSO most frequently achieves full-set recovery, particularly at ($\beta = 0.7$), while the other methods show more modest probabilities of recovering all true signals, and performance deteriorates for all methods as censoring increases.

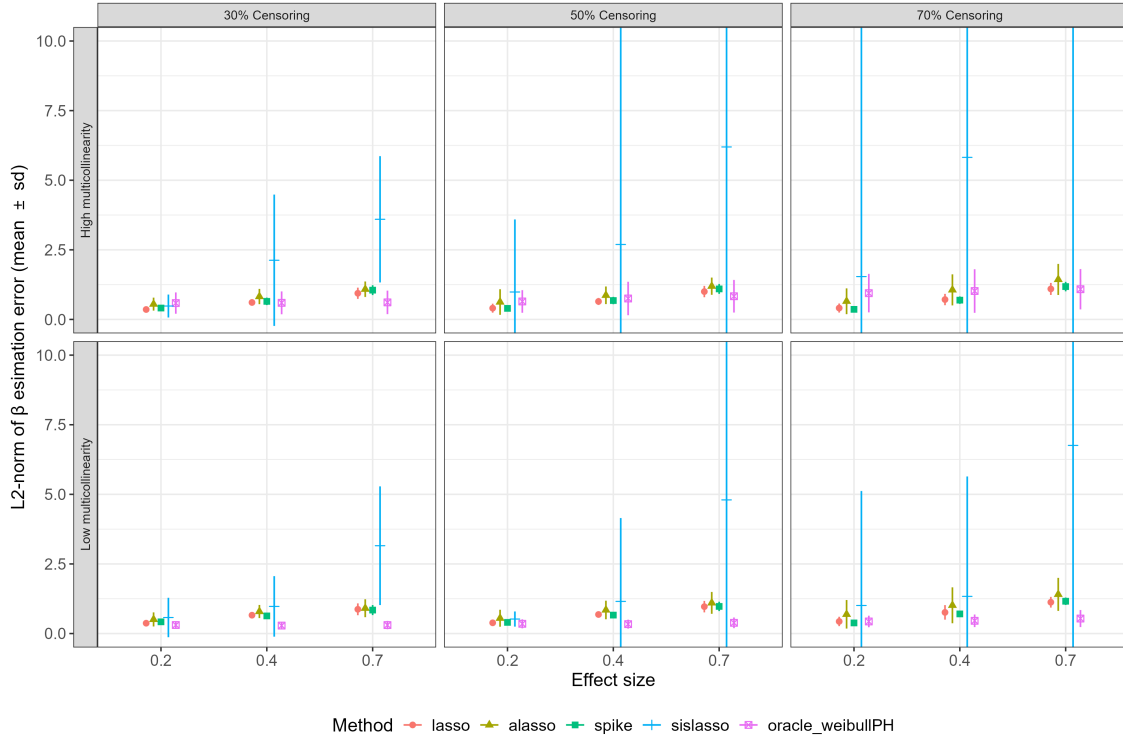


Fig S3.10: Model performance in simulation study: ℓ_2 -Norm of β estimation error, full supports

Figure S3.10. ℓ_2 -Norm of coefficient estimation error on the full support. This figure displays the mean ℓ_2 -Norm of the coefficient estimation error ($\|\hat{\beta} - \beta\|_2$) over the full parameter vector (including noise variables), with ± 1 standard deviation (sd) error bars, by method, effect size, censoring rate, and multicollinearity. The oracle model unsurprisingly yields the smallest estimation error across almost all scenarios. Among the penalized methods, LASSO, adaptive LASSO, and spike-and-slab LASSO show broadly similar and relatively small estimation errors, which increase with heavier censoring. In contrast, iSIS-LASSO exhibits much larger variability, particularly under high multicollinearity, which might be related to the high false discovery rate of selected features.

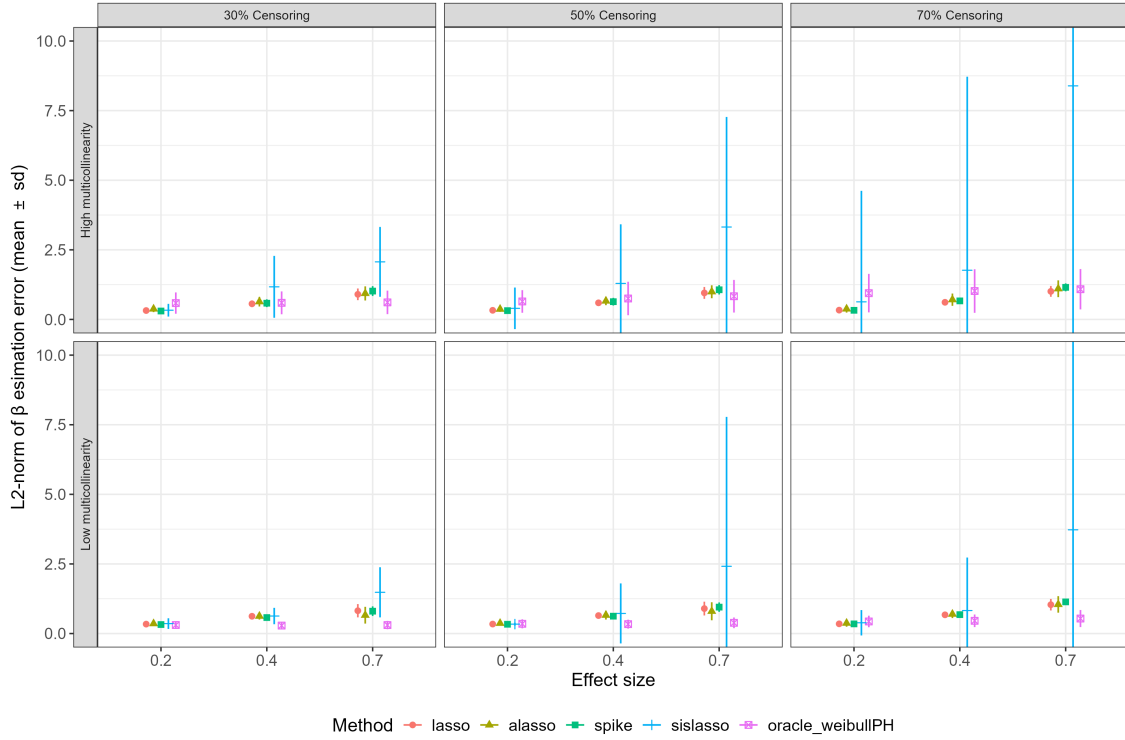


Fig S3.11: Model performance in simulation study: ℓ_2 -Norm of β estimation error, true supports

Figure S3.11. ℓ_2 -Norm of coefficient estimation error on the true support. Mean ℓ_2 -Norm of the estimation error restricted to the true non-zero coefficients with ± 1 standard deviation (sd) is shown for each method and simulation scenario. Errors are smaller than for the full-support metric, and the estimated model coefficient from oracle model remains closest to the true β throughout. The error is similar for LASSO, adaptive LASSO, and spike-and-slab LASSO across different scenarios, whereas iSIS-LASSO yielded the largest errors and variability, especially in high-multicollinearity settings.

List of Figures

Figure 1	Model discrimination assessment
Figure 2	Model calibration assessment
Figure 3	Model coefficient estimation from final models
Figure S1.1	Enriched metabolite sets based on selected metabolites (Models: LASSO, adaptive LASSO, and iSIS-LASSO)
Figure S1.2	Enriched Metabolite Sets based on selected metabolites (Model: spike-and-slab LASSO)
Figure S2.1	Model discrimination assessment (with additional adjustment for tumor site)
Figure S2.2	Model calibration assessment (with additional adjustment for tumor site)
Figure S2.3	Model coefficient estimation from final models (with additional adjustment for tumor site)
Figure S3.1	Distribution of simulated model coefficients
Figure S3.2	Distribution of median survival time of simulated survival data
Figure S3.3	Distribution of censoring of simulated survival data
Figure S3.4	Model performance in simulation study: C-index
Figure S3.5	Model performance in simulation study: iAUC
Figure S3.6	Model performance in simulation study: tAUC
Figure S3.7	Model performance in simulation study: F_1 Score
Figure S3.8	Model performance in simulation study: Recall of true signals
Figure S3.9	Model performance in simulation study: Proportion of full-set recovery
Figure S3.10	Model performance in simulation study: ℓ_2 -Norm of β estimation error, full supports

Figure S3.11 Model performance in simulation study: ℓ_2 -Norm of β estimation error, true supports

List of Tables

Table 1	Patient characteristics at baseline
Table 2	Model coefficients for selected features
Table 3	Calibration assessment – observed and expected event rates across predicted risk quartiles for each modeling approach
Table S1.1	Model specifications of penalized Cox modeling
Table S1.2	Calibration assessment procedure
Table S2.1	Model coefficients for selected features (with additional adjustment for tumor site)
Table S2.2	Model calibration assessment – observed and expected event rates across predicted risk groups for each modeling approach (with additional adjustment for tumor site)