

Aug 12, 2025

Surv-TCAV: Concept-Based Interpretability for Gradient-Boosted Survival Models on Clinical Tabular Data

DOI

<https://dx.doi.org/10.17504/protocols.io.n92ld6pxng5b/v1>

Emmanuel Pio Pastore¹

¹Department of Biology, Ecology and Earth Science, University of Calabria, 87036 Rende, Italy



Emmanuel Pio Pastore

University of Calabria

Create & collaborate more with a free account

Edit and publish protocols, collaborate in communities, share insights through comments, and track progress with run records.

Create free account

OPEN  ACCESS



DOI: <https://dx.doi.org/10.17504/protocols.io.n92ld6pxng5b/v1>

Protocol Citation: Emmanuel Pio Pastore 2025. Surv-TCAV: Concept-Based Interpretability for Gradient-Boosted Survival Models on Clinical Tabular Data. **protocols.io** <https://dx.doi.org/10.17504/protocols.io.n92ld6pxng5b/v1>

License: This is an open access protocol distributed under the terms of the **[Creative Commons Attribution License](#)**, which permits unrestricted use, distribution, and reproduction in any medium, provided the original author and source are credited

Protocol status: In development

We are still developing and optimizing this protocol

Created: August 12, 2025

Last Modified: August 12, 2025

Protocol Integer ID: 224535

Keywords: survival analysis, interpretability, concept activation vectors, gradient boosting, clinical AI, XGBoost, TreeSHAP, interpretable survival analysis on clinical tabular data, interpretable survival analysis, boosted survival model, survival models on clinical tabular data, survival models on structured clinical data, based interpretability, clinical tabular data, interpretability for gradient, structured clinical data, interpretability, testing with concept activation vector, concept activation vector, boosted accelerated failure time, primary biliary cirrhosis cohort, consistent with clinical expectation, tcav quantify, level directional sensitivity through an aft adaptation, tcav

Abstract

This protocol delivers interpretable survival analysis on clinical tabular data by pairing gradient-boosted accelerated failure time (AFT) modeling with concept-based explanations. The model is trained on the Primary Biliary Cirrhosis cohort (*PBC-276*, complete cases over 17 covariates) using XGBoost's AFT objective. Discrimination is assessed via the concordance index and calibration via integrated Brier score with inverse probability of censoring weights. Interpretability operates at two levels: feature-level summaries with TreeSHAP and concept-level directional sensitivity through an AFT adaptation of Testing with Concept Activation Vectors (TCAV), termed *Surv-TCAV*. *Surv-TCAV* quantifies how clinician-defined concepts (*cholestasis*, *coagulopathy*, *low albumin*, *older age*, *clinical complications*) perturb the model's predicted location parameter μ of $\log T$, reporting standardized effects and bootstrap confidence intervals. With 25 independent 80/20 train/validation partitions and fixed boosting rounds (no early stopping), the protocol yields a validation C-index of 0.837 ± 0.040 and IBS of 0.321 ± 0.023 . *Surv-TCAV* indicates negative directional effects for *low albumin*, *older age*, and *clinical complications*, consistent with clinical expectations. This is the first protocol applying concept-based interpretability to boosted survival models on structured clinical data, with full reproducibility. Source code: <https://github.com/emmanuel6474/surv-tcav-pbc>.

Image Attribution

Figure 1: Validation diagnostics: distribution of C-index across 25 validations and IBS vs. C-index scatter on validation splits.

Figure 2: TreeSHAP summary plot for the last trained XGBoost-AFT model, showing global feature attributions.

Figure 3: *Surv-TCAV* directional effects on μ for clinical concepts on the last split, with 95% bootstrap confidence intervals.

Guidelines

Preprocessing. Categorical variables (sex, discrete signs, and optionally trt, stage) are one-hot encoded with unknown-handling. Numerical variables pass through. Complete-case selection over the 17 features is applied.

Model. XGBoost with AFT objective models the log-time outcome with right-censoring. The **extreme** (Gumbel) loss distribution is used with fixed hyperparameters (Table 1). For each partition, the model is fit on 80% and evaluated on 20%.

Metrics. Discrimination is measured by the C-index [12]; calibration by integrated Brier score with IPCW [13, 14]. We report mean and standard deviation across 25 validations (Table 2).

Interpretability. TreeSHAP summarizes feature influence for the last trained model [6]. Surv-TCAV adapts TCAV to AFT: for each concept, a direction in standardized numerical-feature space is learned via a balanced logistic probe separating concept-positive from negative cases [7]. We estimate the directional derivative of μ along this vector using small steps h with Gaussian smoothing and Monte Carlo averaging, summarize effects as $\Delta\mu$ and standardized $\Delta\mu/SD(\mu)$, and compute bootstrap confidence intervals.

Hyperparameters.

Parameter	Value

Objective	survival:aft
AFT loss distribution	extreme (Gumbel)
AFT scale	0.4
Learning rate	0.02
Max depth	10
Min child weight	5.0
Subsample	0.6
Colsample by tree	0.6
L2 regularization λ	4.0
L1 regularization α	0.0
Tree method	hist (or gpu_hist if GPU)
Number of boosting rounds	800

Validation performance.

Metric	Value

Concordance index (C-index)	0.837 $\pm$ 0.040
Integrated Brier Score (IBS)	0.321 $\pm$ 0.023

Surv-TCAV effect summary.

Concept	$\Delta\mu$ (95% CI)	$\Delta\mu/SD(\mu)$ (95% CI)	Positivity rate
Cholestasis	-808.77 (-1937.45, +277.40)	-0.247 (-0.591, +0.085)	0.12
Coagulopathy	-197.62 (-441.47, +78.61)	-0.060 (-0.135, +0.024)	0.26
Low albumin	-564.38 (-994.48, -100.71)	-0.172 (-0.304, -0.031)	0.25
Older age	-603.51 (-1030.31, -172.03)	-0.184 (-0.315, -0.053)	0.25
Clinical complications	-399.41 (-618.53, -191.14)	-0.122 (-0.189, -0.058)	0.34

Materials

Experiments use Python 3 with **xgboost**, **pandas**, **numpy**, **scikit-learn**, **lifelines**, **matplotlib**, and **shap** [16, 19, 18]. The reference implementation, including figure generation, is available at <https://github.com/emmanuel6474/surv-tcav-pbc>.

Before start

The Mayo Clinic Primary Biliary Cirrhosis dataset is distributed in the **survival** R package [20, 21] and mirrored via *Rdatasets* [22]. The cohort includes baseline labs, signs, staging, follow-up times, and censoring indicators. Following common practice, complete cases over 17 covariates define the *PBC-276* cohort used here. Data for replication are fetched from <https://vincentarelbundock.github.io/Rdatasets/csv/survival/pbc.csv>. Users should verify licensing and cite original sources [20, 21, 22].

Protocol Overview

- 1 Partition the dataset into 25 independent random splits, each with 80% training and 20% validation, stratified by the event indicator. Do not include a test set; aggregate metrics on validation splits to match literature reporting for PBC benchmarks. Use a fixed number of boosting rounds with no early stopping to avoid leakage from validation to training.
- 2 Preprocess the data: One-hot encode categorical variables (sex, discrete signs, and optionally trt, stage) with unknown-handling. Allow numerical variables to pass through unchanged. Apply complete-case selection over the 17 features.
- 3 Train the model: Use XGBoost with the AFT objective to model the log-time outcome with right-censoring. Use the extreme (Gumbel) loss distribution and fixed hyperparameters (see Table 1). For each partition, fit the model on 80% of the data and evaluate on the remaining 20%.
- 4 Evaluate metrics: Measure discrimination using the C-index and calibration using the integrated Brier score with IPCW. Report mean and standard deviation across the 25 validations (see Table 2).
- 5 Interpret results: Use TreeSHAP to summarize feature influence for the last trained model. Adapt TCAV to AFT for concept-level interpretability: for each concept, learn a direction in standardized numerical-feature space via a balanced logistic probe separating concept-positive from negative cases. Estimate the directional derivative of μ along this vector using small steps h with Gaussian smoothing and Monte Carlo averaging. Summarize effects as $\Delta\mu$ and standardized $\Delta\mu/SD(\mu)$, and compute bootstrap confidence intervals.

Concept Definitions and Surv-TCAV Settings

- 6 Define concepts a priori on training data as follows: *cholestasis* (bilirubin and alkaline phosphatase $\geq$ training 75th percentile), *coagulopathy* (prothrombin time $\geq$ training 75th percentile), *low albumin* (albumin $\leq$ training 25th percentile), *older age* (age $\geq$ training 75th percentile), and *clinical complications* (any of ascites, edema, spider angioma positive).
- 7 Fit probes for each concept using l2-regularized logistic regressions with class balancing. Compute directional derivatives using step size $h = 0.02$, Gaussian smoothing $\sigma = 0.04$, and $K = 80$ Monte Carlo draws.
- 8 Summarize effects on the last split with 95% bootstrap confidence intervals. Effects are reported as $\Delta\mu$ and standardized $\Delta\mu/SD(\mu)$. The positivity rate is the fraction of validation samples positive for the concept.

Reproducibility and Sharing

- 9 Access the complete reference implementation, which downloads the dataset, builds the PBC-276 cohort, executes the 25×80/20 protocol, computes C-index and IBS, generates TreeSHAP, and performs Surv-TCAV, at <https://github.com/emmanuel6474/surv-tcav-pbc>. The repository includes an executable script with fixed hyperparameters and seeds, a pinned environment, and figure generation.

Figures

10

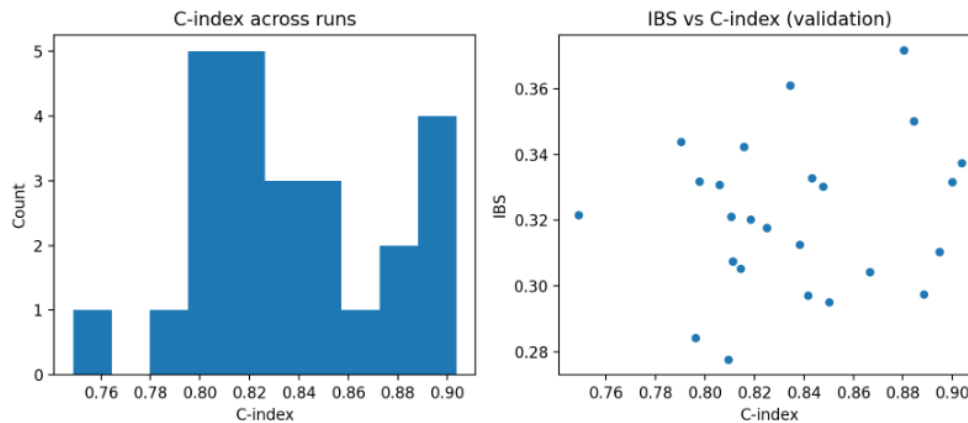


Figure 1: Validation diagnostics: distribution of C-index across 25 validations and IBS vs. C-index scatter on validation splits.

11

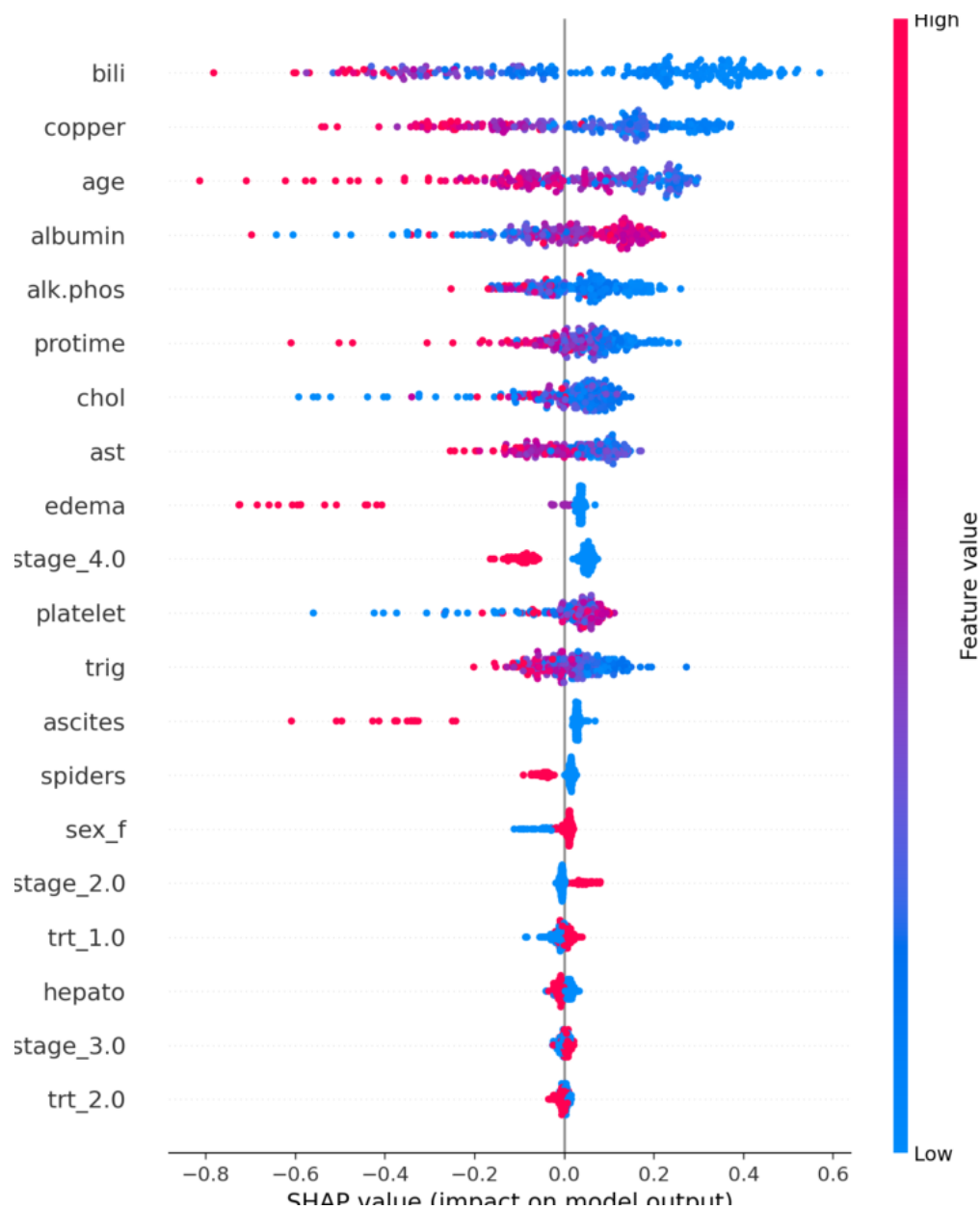


Figure 2: TreeSHAP summary plot for the last trained XGBoost-AFT model, showing global feature attributions.

12

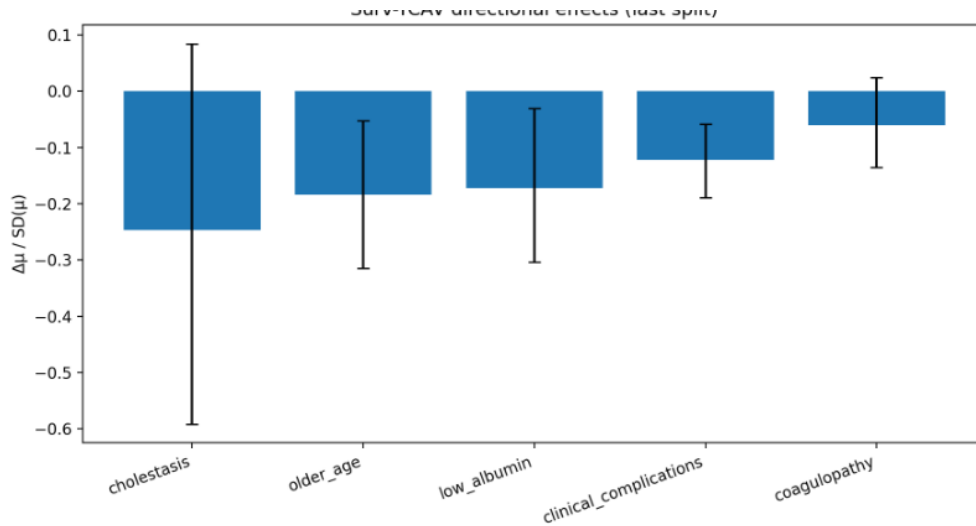


Figure 3: Surv-TCAV directional effects on μ for clinical concepts on the last split, with 95% bootstrap confidence intervals.

Limitations and Extensions

- 13 Concept definitions depend on available variables and clinical judgment. Surv-TCAV evaluates directional effects on μ and does not substitute for causal inference. Extensions include automated concept discovery [8] and concept bottlenecks [9]. The AFT distributional choice can vary; the present protocol uses *extreme* (Gumbel). If early stopping is desired, an external validation fold is required to avoid leakage.

Conclusion

- 14 This protocol provides a reproducible, fair-comparison setup for boosted survival modeling on PBC-276 with clinically aligned interpretability. It reports strong discrimination and transparent diagnostics, and quantifies how high-level clinical patterns influence predicted survival.

Data and Code Availability

- 15 Dataset: <https://vincentarelbundock.github.io/Rdatasets/csv/survival/pbc.csv> [22]. R package source: [21]. Code: <https://github.com/emmanuel6474/surv-tcav-pbc>.

Protocol references

- [1] F. Doshi-Velez and B. Kim. Towards a rigorous science of interpretable machine learning. arXiv:1702.08608, 2017.
- [2] C. Rudin. Stop explaining black box machine learning models for high stakes decisions and use interpretable models instead. *Nature Machine Intelligence*, 1(5):206–215, 2019.
- [3] J. Amann et al. Explainability for artificial intelligence in healthcare: a multidisciplinary perspective. *BMC Medical Informatics and Decision Making*, 20(1):310, 2020.
- [4] M. T. Ribeiro, S. Singh, and C. Guestrin. “Why Should I Trust You?”: Explaining the Predictions of Any Classifier. In *KDD*, 2016.
- [5] S. M. Lundberg and S.-I. Lee. A unified approach to interpreting model predictions. In *NeurIPS*, 2017.
- [6] S. M. Lundberg, G. Erion, and S.-I. Lee. From local explanations to global understanding with explainable AI for trees. *Nature Machine Intelligence*, 2:56–67, 2020.
- [7] B. Kim, M. Wattenberg, J. Gilmer, et al. Interpretability Beyond Feature Attribution: Quantitative Testing with Concept Activation Vectors (TCAV). In *ICML*, 2018.
- [8] A. Ghorbani, J. Wexler, J. Zou, and B. Kim. Towards Automatic Concept-Based Explanations. In *NeurIPS*, 2019.
- [9] P. W. Koh et al. Concept Bottleneck Models. In *ICML*, 2020.
- [10] E. L. Kaplan and P. Meier. Nonparametric estimation from incomplete observations. *JASA*, 53(282):457–481, 1958.
- [11] D. R. Cox. Regression models and life-tables. *JRSS B*, 34(2):187–220, 1972.
- [12] F. E. Harrell Jr., K. L. Lee, and D. B. Mark. Multivariable prognostic models: issues and measures. *Statistics in Medicine*, 15(4):361–387, 1996.
- [13] E. Graf, C. Schmoor, W. Sauerbrei, and M. Schumacher. Assessment and comparison of prognostic classification schemes for survival data. *Statistics in Medicine*, 18(17–18):2529–2545, 1999.
- [14] T. A. Gerds and M. Schumacher. Consistent estimation of the expected Brier score in general survival models with right-censoring. *Biometrical Journal*, 48(6):1029–1040, 2006.
- [15] J. H. Friedman. Greedy function approximation: A gradient boosting machine. *Annals of Statistics*, 29(5):1189–1232, 2001.

- [16] T. Chen and C. Guestrin. XGBoost: A scalable tree boosting system. In *KDD*, 2016.
- [17] XGBoost Documentation. Survival Analysis with AFT loss. <https://xgboost.readthedocs.io>. Accessed 2025.
- [18] C. Davidson-Pilon et al. lifelines: survival analysis in Python. *Journal of Open Source Software*, 4(40):1317, 2019.
- [19] F. Pedregosa et al. Scikit-learn: Machine Learning in Python. *JMLR*, 12:2825–2830, 2011.
- [20] T. M. Therneau and P. M. Grambsch. *Modeling Survival Data: Extending the Cox Model*. Springer, 2000.
- [21] T. M. Therneau. A Package for Survival Analysis in R. <https://CRAN.R-project.org/package=survival>. Accessed 2025.
- [22] V. Arel-Bundock. Rdatasets: Datasets from R packages. <https://vincentarelbundock.github.io/Rdatasets>. Accessed 2025.