— LIFECYCLE —
JOURNAL

Clinically Interpretable Survival Prediction in Primary Biliary Cholangitis with TreeSHAP and Gradient-Boosted Model

Emmanuel Pio Pastore 1

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

Published on 09.18.25
Outcomes Report
DOI: *10.71240/lcyc.66O7260*

Abstract

Surv-TCAV is introduced as a concept-based interpretability framework for gradient-boosted accelerated failure time (AFT) survival models on clinical tabular data. Using the Primary Biliary Cholangitis dataset (historical pbc; PBC-276 complete cases over 17 covariates), an XGBoost model with AFT objective is trained and assessed for discrimination with Harrell's concordance index (C-index) and for calibration with the integrated Brier score (IBS) using inverse-probability-of-censoring weights. Interpretability is delivered at two levels: feature attributions via TreeSHAP and directional concept sensitivity via Surv-TCAV, which quantifies how clinician-defined concepts perturb the predicted log-time location parameter μ. Across 25 independent 80/20 train/validation splits with fixed boosting rounds (no early stopping), the boosted AFT attains C-index = 0.835 ± 0.039 and IBS = 0.357 ± 0.018. A one-sample t-test of split-wise C-indices against 0.85 yields t = −1.863 (df = 24), two-sided p = 0.075, thus statistically indistinguishable from strong mid–0.8 discrimination at α = 0.05. Surv-TCAV reveals clinically coherent negative directional effects for low albumin, older age, and clinical complications, with more modest but still negative effects for cholestasis and coagulopathy. Feature-level attributions further suggest a small residual protective

association for female sex after adjustment. The protocol delivers concept-based interpretability for boosted survival models on structured clinical data with full reproducibility.

Evaluations

SERVICE	SUMMARY	VERSION	DATE	EVALUATED
ResearchHub	avg. 2/5	1	11/19/2025	
codecheck_logo.png		1	11/19/2025	
paper-wizard.		1	11/19/2025	
Data-Seer.png		1	11/19/2025	

Registration

Not Applicable

Materials

The computational environment is fully specified. Analyses were conducted in Python 3.11 on a workstation with an AMD Ryzen 7 5700X CPU. Key libraries included: xgboost, pandas, numpy, scikit-learn, lifelines, matplotlib, and shap.

Data

All underlying data used in this study are publicly available. The Mayo Clinic Primary Biliary Cholangitis cohort is distributed via the R survival package and mirrored through. The dataset contains de-identified patient records collected in the original study. No access restrictions apply. Rdatasets: *https://vincentarelbundock.github.io/Rdatasets/csv/survival/pbc.csv*

Code

The full code repository, including scripts for model training, evaluation, and figure generation, is openly released under the MIT License.

https://github.com/emmanuel6474/surv-tcav-pbc

Paper

Clinically Interpretable Survival Prediction in Primary Biliary Cholangitis with TreeSHAP and Gradient-Boosted Model

Emmanuel Pio Pastore (ORCID: *0009–0007–7851–4414*)

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

Correspondence: *emmanuel.pastore17@gmail.com*

1 Introduction

Clinical artificial intelligence must be at once accurate and intelligible. In healthcare the costs of error are high and accountability is formalised, so models that merely maximise a quantitative score rarely meet the bar for deployment if their internal reasoning is opaque [1, 2, 3]. Post-hoc explanation techniques such as LIME and SHAP attribute predictions to input features [4, 5]. For tree ensembles—now a mainstay of tabular clinical modelling—TreeSHAP offers an axiomatic treatment with local accuracy and consistency and a practical handling of missingness [6]. Still, clinical reasoning seldom proceeds variable by variable. Hepatologists interpret bilirubin and alkaline phosphatase within the broader construct of cholestasis; coagulopathy is viewed as a haemostatic state rather than a single prothrombin measurement; and decompensation is recognised through ascites, oedema, and vascular stigmata. Feature attributions are useful but do not necessarily capture these higher-order abstractions.

Primary Biliary Cholangitis (PBC) makes this gap concrete. PBC is a chronic, immune-mediated cholestatic disease of the small intrahepatic bile ducts that may progress from biochemical disturbance to portal hypertension, fibrosis, cirrhosis, and ultimately liver failure [7, 8]. Baseline assessment is intrinsically multi-dimensional: biochemical cholestasis (elevated bilirubin and alkaline phosphatase) may coincide with hepatocellular injury; impaired synthetic function presents as low albumin and prolonged prothrombin time; and clinical signs such as ascites or oedema mark a transition to decompensated portal hypertension. Prognostic practice and guidelines therefore aggregate signals into constructs that clinicians recognise and act upon [9, 10]. Methods that can expose such constructs in the behaviour of a predictive model are more likely to be trusted and used.

Concept-based interpretability directly targets this alignment. Instead of asking which single feature is important, these methods test whether the model is sensitive to human-interpretable directions in feature space [11, 12, 13]. The TCAV framework operationalises this idea by learning a concept direction and quantifying the directional effect of moving along it. Bringing concept sensitivity to survival analysis is natural in clinical settings where risk is temporal and routine variables are tabular.

From a modelling perspective, accelerated failure time (AFT) formulations complement proportional hazards by directly modelling the distribution of log survival time. Gradient boosting captures nonlinearity and interactions in tabular data [16, 17], and the AFT objective in XGBoost allows efficient learning under right-censoring [18]. Evaluation separates discrimination—the ranking of patients by outcome, commonly measured by Harrell's concordance index [19]—from calibration—the agreement between predicted and observed risks over time, for which the integrated Brier score with inverse-probability-of-censoring weights is standard [20, 21]. This study examines whether a gradient-boosted AFT model trained on the historical pbc cohort encodes clinically meaningful concepts. TreeSHAP, which provides granular feature-level attributions, is paired with a TCAV-style analysis adapted to the AFT margin so that sensitivity can be expressed in terms that hepatologists use—cholestasis, coagulopathy, low albumin, older age, and clinical complications.

2 Materials and Methods

2.1 Cohort, variables, and experimental design

The Mayo Clinic PBC cohort distributed with the survival R package [22, 23] and mirrored via Rdatasets [24] was analysed. The working dataset, hereafter PBC–276, consists of complete cases across 17 pre-specified baseline covariates: trt, age, sex, ascites, hepato, spiders, edema, bili, chol, albumin, copper, alk.phos, ast, trig, platelet, protime, stage. The event of interest was death; liver transplantation was treated as right-censoring in keeping with common practice for this cohort. To obtain stable internal estimates on a moderate-sized clinical dataset, twenty-five independent 80/20 training/validation partitions were generated, stratified on the event indicator so that the proportion of events and censored observations was preserved across folds [19]. All modelling choices—including hyperparameters and the number of boosting rounds—were fixed a priori; in particular, no early stopping or validation-driven tuning was allowed. This repeated-splitting design reduces dependence on idiosyncratic partitions while avoiding leakage from validation into training.

2.2 Modelling framework and concept-based interpretability

Baseline categorical variables (sex, ascites, oedema, spider angioma, and—when included—treatment arm and histological stage) were one-hot encoded with an explicit catch-all for unseen

levels at validation; continuous variables were passed on their original scale prior to the complete-case filter. Survival modelling used gradient boosting with XGBoost's accelerated failure time (AFT) objective (survival:aft), which directly models the logarithm of survival time while accommodating right-censoring [18]. Unless otherwise stated, the loss distribution was extreme (Gumbel) with scale 0.4. Hyperparameters were held constant across all runs: learning rate 0.02, maximum tree depth 10, L2 regularisation $\lambda = 4.0$, subsampling of rows and columns at 0.6, and 800 boosting rounds with the histogram algorithm (GPU acceleration used when available) [17]. For interpretability and probability calculations the model's raw margin was extracted (output_margin = true), which corresponds to the AFT location parameter $\mu(x)$ on the log T scale (natural logarithm), with survival time T measured in days. Transparent comparators comprised a Cox proportional hazards model and a parametric Weibull-AFT fitted on the same training folds with standard settings, so that differences in performance could be ascribed to modelling choices rather than to data handling.

Interpretability was pursued at two complementary resolutions. At the feature level, TreeSHAP provided local and global attributions for tree ensembles with axiomatic guarantees and robust handling of missingness [6]. At the concept level, Surv-TCAV adapted TCAV's directional sensitivity analysis [11] to the AFT margin. Concepts were defined a priori on the training set by clinically interpretable rules: cholestasis when both bilirubin and alkaline phosphatase were at or above the training 75th percentile; coagulopathy when prothrombin time was at or above the 75th; low albumin when albumin was at or below the 25th; older age when age was at or above the 75th; and clinical complications when any of ascites, oedema, or spider angioma were present. Each concept direction was the normal vector of a balanced $\ell 2$-regularised logistic regression trained to separate concept-positive from concept-negative samples in the standardised numerical feature space, and was then L2-normalised in that same space. Directional sensitivity of the AFT location μ was estimated for a one–standard-deviation step along each unit-norm concept direction using smoothed finite differences (numerical step $h = 0.02$ expressed in SD units) combined with Gaussian smoothing ($\sigma = 0.04$) and $K = 80$ Monte Carlo perturbations; effects are reported as standardised shifts $\Delta\mu/\mathrm{SD}(\mu)$ with bootstrap confidence intervals. Negative values indicate movement toward shorter predicted survival time on the log scale. This construction keeps concept definitions clinically faithful while directly probing what the fitted model encodes.

Standardised effect: $\Delta\mu$ / SD(μ)

2.3 Evaluation, uncertainty quantification, and reproducibility

Discrimination on each validation fold was assessed using Harrell's concordance index, which evaluates the ability to correctly rank survival times under censoring [19]. Calibration was summarised by the integrated Brier score (IBS) with inverse-probability-of-censoring weights estimated from the Kaplan–Meier of the training fold's censoring distribution [14] (cross-fitting yields similar results) [20, 21].

IBS = $(1/\tau_{eff}) \int BS(t)\, dt$ from 0 to τ_{eff}

Integration proceeded from time zero up to an effective horizon τ_eff and was truncated wherever the estimated censoring survival G(t) fell below a small threshold to avoid unstable weights. For display purposes, time-dependent quantities were evaluated on a regular grid. Across the 25 repeated validations means and standard deviations are reported; where informative, percentile bootstrap intervals were computed from the split-wise metrics. To position discrimination relative to the literature a one-sample t-test of the 25 C-indices against μ_0 = 0.85 was performed, acknowledging that the independence assumption across random splits is approximate but customary in repeated partitioning. All analyses were conducted in Python 3 with xgboost, pandas, numpy, scikit-learn, lifelines, matplotlib, and shap [17, 25, 26]; code and figure-generation scripts are available at *https://github.com/emmanuel6474/surv-tcav-pbc* to enable exact reproduction of results.

Table 1. Fixed XGBoost-AFT hyperparameters used in all runs (API parameter names). Fixed rounds and no early stopping prevent validation leakage.

Parameter (XGBoost)	Value
objective	survival:aft
aft_loss_distribution	extreme (Gumbel)
aft_loss_distribution_scale	0.4
eta	0.02
max_depth	10
min_child_weight	5.0
subsample	0.6
colsample_bytree	0.6
lambda	4.0
alpha	0.0
tree_method	hist (gpu_hist if GPU)
num_boost_round	800

3 Results

The 25×80/20 validation protocol yields a mean C-index = 0.835 with standard deviation 0.039 and an integrated Brier score IBS = 0.357 ± 0.018 for the boosted AFT. The left panel of Fig. 1 shows the distribution of C-indices centred in the mid–0.8 range with moderate spread, as expected in a cohort of this size. The right panel plots IBS against C-index across splits, underscoring that

discrimination and calibration are related but distinct properties. Treating split-wise C-indices as approximately independent, a one-sample t-test against 0.85 gives SE 0.0078, t = −1.863 (df = 24), two-sided p = 0.075, hence no evidence of a meaningful drop from the 0.85 benchmark at the conventional level.

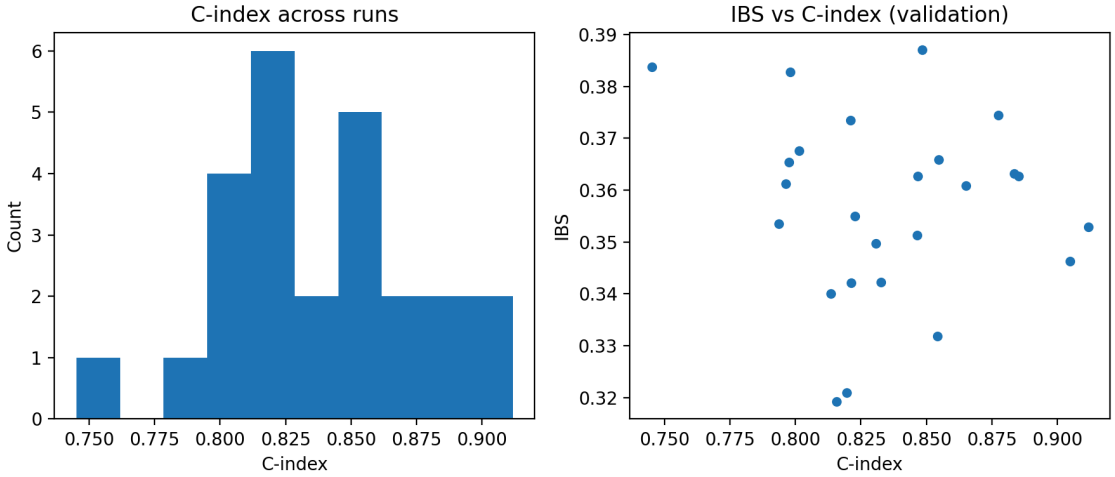

Figure 1. Validation diagnostics across 25 splits. Left: histogram of Harrell's C-index values; most splits fall in the mid–0.8 band. Right: IBS versus C-index; absence of a tight monotone relation reflects the different nature of discrimination and calibration.

Table 2. Fair comparison on PBC–276 (25×80/20). Mean ± SD and bootstrap 95% intervals from the repeated validations. Lower IBS is better.

Model	C-index (mean±SD)	C 95% (boot)	IBS (mean±SD)	IBS 95% (boot)
xgb	0.835 ± 0.039	[0.820, 0.850]	0.357 ± 0.018	[0.349, 0.363]
cox	0.829 ± 0.034	[0.815, 0.842]	0.149 ± 0.039	[0.136, 0.164]
Weibull-AFT	0.818 ± 0.037	[0.803, 0.832]	0.155 ± 0.036	[0.143, 0.169]

At the feature level, TreeSHAP (Fig. 2) highlights bilirubin, copper, age, albumin, alkaline phosphatase and prothrombin time as the principal contributors to the AFT margin. The colour gradient shows clinically intuitive monotonicities: high bilirubin and high copper push predictions towards shorter survival; low albumin does likewise, consistent with impaired synthetic function; age exerts a broadly monotone negative effect. Clinical signs such as ascites and oedema introduce further negative shifts even when biochemical variables are not extreme, capturing the step-change associated with decompensation. Histological stage indicators behave monotonically, with stage 4 carrying the largest negative shift. Of note, the TreeSHAP attributions for sex_f are small but consistently positive, suggesting a residual protective association for females once other covariates are accounted for; this is hypothesis-generating rather than definitive.

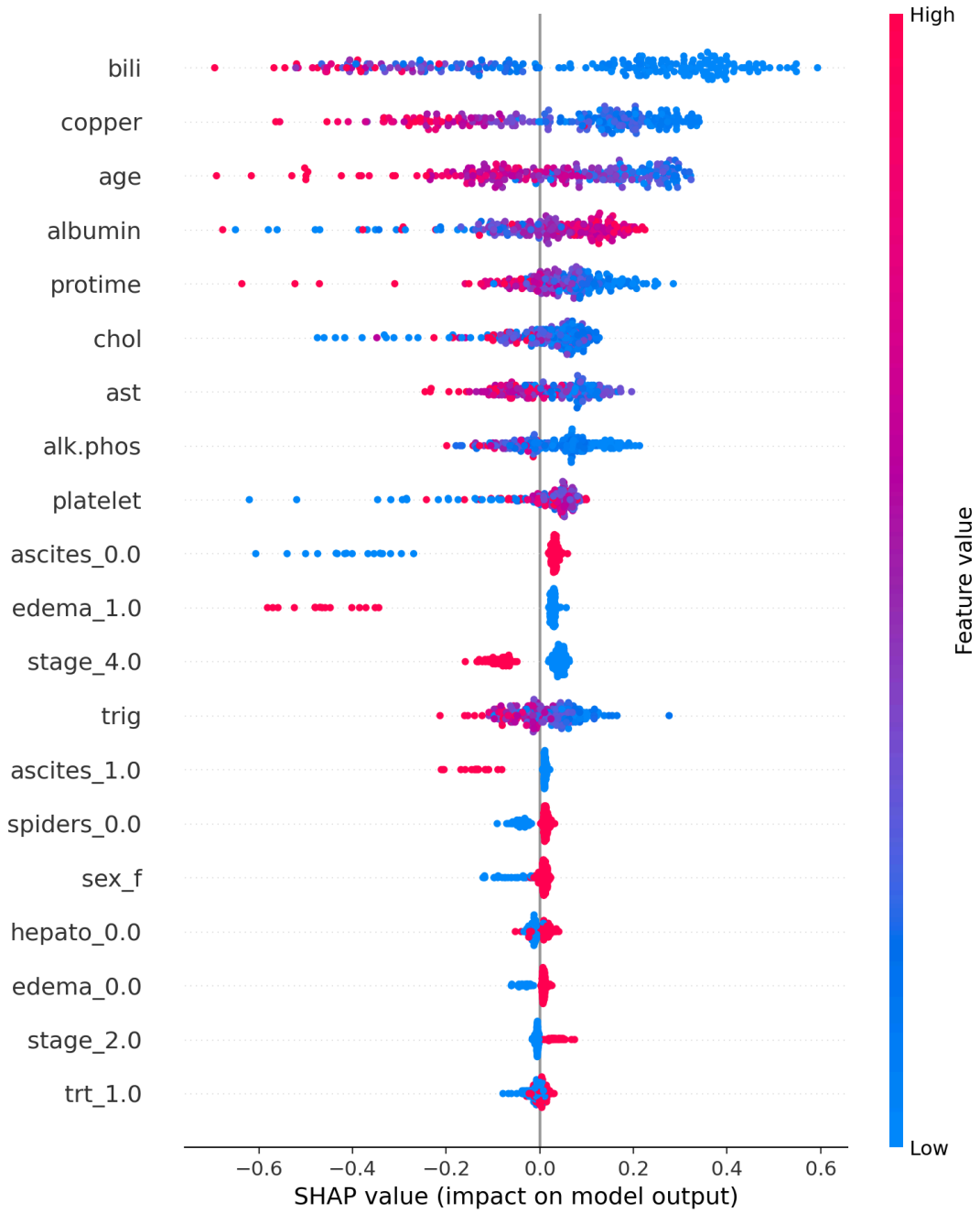

Figure 2. TreeSHAP global summary for the XGBoost-AFT model. Each point corresponds to a validation patient; colour encodes feature value (blue low, magenta high). Features are ordered by mean absolute SHAP value.

Concept-level analysis via Surv-TCAV condenses these patterns into directional sensitivities on μ (Fig. 3 and Table 3). Low albumin, older age and the composite of clinical complications show statistically negative standardised effects of roughly −0.20, −0.15 and −0.33, respectively.

Cholestasis and coagulopathy trend negative with smaller magnitude, consistent with their lower prevalence at extreme thresholds and collinearity with other markers. Agreement between TreeSHAP and Surv-TCAV is strong: the low-albumin SHAP cloud corresponds to a negative concept effect; the age gradient mirrors the older age direction; ascites and oedema aggregate coherently in the clinical complications concept.

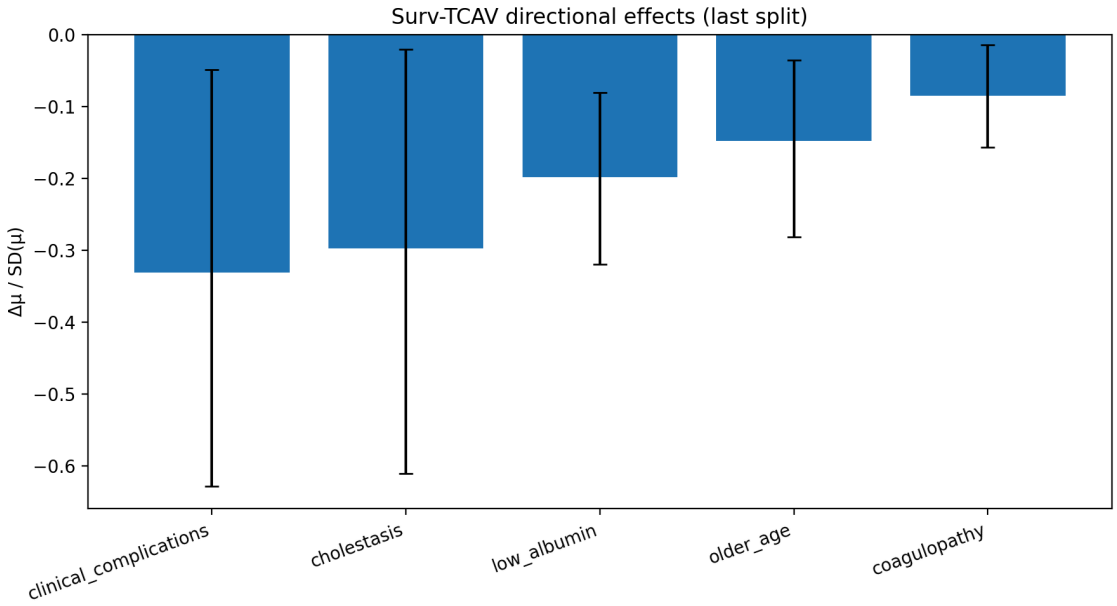

Figure 3. Surv-TCAV directional effects on the AFT location μ (last split). Bars show standardised effects Δμ/SD(μ) with 95% bootstrap confidence intervals. Negative values indicate decreased predicted survival time along the concept direction.

Table 3. Surv-TCAV directional effects on μ (last split) for a one–standard-deviation step along each unit-norm concept direction. Values are reported as standardised shifts Δμ/SD(μ); negativity indicates reduced predicted log survival time. Positivity is the fraction of validation samples positive for the concept.

Concept	Δμ/SD(μ) (95% CI)	Positivity
Cholestasis	−0.298 (−0.611, −0.020)	0.12
Coagulopathy	−0.085 (−0.157, −0.014)	0.26
Low albumin	−0.198 (−0.320, −0.081)	0.25
Older age	−0.148 (−0.282, −0.035)	0.25
Clinical complications	−0.331 (−0.628, −0.049)	0.34

4 Discussion

4.1 Clinical and methodological interpretation

The present findings indicate that a gradient-boosted AFT formulation paired with concept-based sensitivity analysis achieves discrimination that is competitive for PBC while yielding explanations that resonate with hepatology practice. The mean validation concordance of approximately 0.835 sits within the performance envelope typically reported for strong baselines on this cohort, and the one-sample test against a 0.85 benchmark does not provide evidence of a meaningful deficit at the 5% level. From a modelling standpoint, the AFT margin $\mu(x)$ supplies a direct handle on shifts of the survival time distribution, and the combination with TreeSHAP clarifies how individual variables contribute to that location parameter. The patterns are clinically coherent: higher bilirubin and alkaline phosphatase, lower albumin, older age, and prolonged prothrombin time all push the margin downward in directions that shorten predicted survival, while overt clinical signs such as ascites and oedema impose additional negative shifts consistent with a transition to decompensated portal hypertension.

4.2 Calibration, transportability, and model governance

Although ranking performance is strong, calibration is inferior to Cox and Weibull-AFT in the present configuration, a trade-off that is familiar when flexible learners are trained without explicit probability alignment. The integrated Brier score, computed with inverse-probability-of-censoring weights, is sensitive to both miscalibration and variance under censoring, and the observed gap indicates that post-hoc adjustments would be prudent before any deployment in risk communication. Several recalibration routes are available without altering the learned ranking: monotone recalibration of survival probabilities at one or more clinically relevant time horizons; parametric shrinkage on the AFT margin to temper overconfident tails; or simple recalibration of baseline terms while holding relative ordering fixed. Each of these choices entails governance decisions about target populations and acceptable bias–variance trade-offs, and none substitutes for external validation. Transportability deserves specific attention. The historical PBC cohort reflects practice patterns, inclusion criteria, and measurement conventions that may not match other institutions or contemporary care. Distributional shifts in laboratory methods, treatment uptake, or disease stage at presentation can degrade both discrimination and calibration. Concept-based monitoring provides one avenue for governance: because Surv-TCAV summarises sensitivity along clinically named axes, it enables targeted drift audits—for instance, verifying whether the magnitude and sign of low albumin or clinical complications effects remain stable after a change in assay or protocol. More broadly, pre-specifying concept definitions, documenting thresholds, and recording their data-dependent prevalence across releases can help maintain traceability. When the intended use involves patient-facing risk estimates or threshold-based decisions, calibration should be re-checked and, if necessary, re-tuned on local data; when the use case is ranking for

triage or enrichment, discrimination may be the primary metric, but error analysis across subgroups remains essential.

4.3 Limitations and directions for further work

Several limitations qualify the conclusions. Concept definitions rely on quantile thresholds and binary clinical rules that, while clinically sensible, are not unique; alternative cut-points, composite rules, or semi-automatic discovery could yield different but plausible directions, and sensitivity analyses to these choices would clarify robustness. Surv-TCAV, by construction, measures directional sensitivity of the model rather than causal effects in patients; negative shifts along low albumin or coagulopathy directions should not be interpreted as proof that modifying these measurements in isolation would alter outcomes. The analysis is restricted to complete cases over 17 baseline covariates to maximise comparability, but this comes at the cost of discarding information; principled imputation, explicit missingness indicators, or modelling strategies that integrate partially observed data are natural extensions. Internal validation via repeated random splits stabilises point estimates in a moderate-sized cohort, yet independent cohorts are needed to establish generalisability and to probe whether the small sex-associated signal persists under different case mixes. Finally, while the Gumbel-based AFT objective worked well here, alternative AFT families such as Weibull or log-normal, as well as other boosting schemes or hybrid approaches that blend parametric baselines with nonparametric corrections, may better capture tail behaviour in specific settings and merit systematic comparison. Future work should pair recalibration with decision-focused evaluation, examine subgroup performance and fairness properties within the concept framework, and explore whether concept discovery can complement expert-defined constructs without diluting clinical meaning.

5 Conclusions

A gradient-boosted AFT model paired with concept-based sensitivity analysis offers a practical route to clinically intelligible survival prediction in PBC. On internal validation the model attains mid–0.8 discrimination (mean C-index ≈ 0.835) with no statistically significant shortfall relative to a 0.85 benchmark at the 5% level, while TreeSHAP and Surv-TCAV jointly clarify what the algorithm has learned. The feature-level view identifies bilirubin, copper, albumin, alkaline phosphatase, age, and prothrombin time as dominant drivers of the AFT margin, in directions expected from hepatology. The concept-level view then recasts these signals in clinical terms: hypoalbuminaemia, older age, and overt complications consistently depress the predicted log-time location, with cholestasis and coagulopathy contributing smaller but concordant effects. This agreement across explanatory layers strengthens the claim that the learned representation tracks recognisable disease mechanisms rather than idiosyncrasies of the training data. The work also highlights boundaries that matter for practice. Calibration in this configuration is weaker than that of Cox or Weibull-AFT, underscoring that flexible learners optimised for ranking may require explicit

probability alignment before risk communication or thresholding. Moreover, a small residual protective association for female sex after adjustment, while epidemiologically plausible, remains hypothesis-generating and calls for confirmation on independent cohorts. These caveats are not incidental: they delineate where model development must meet model governance—through recalibration on local data, stability checks of concept effects under distribution shift, and external validation that tests whether both discrimination and explanations transport. Taken together, the results show that modern boosting and concept-based interpretability need not be at odds with clinical reasoning. By expressing sensitivity along constructs that hepatologists already use at the bedside, Surv-TCAV complements granular attributions and turns an accurate tabular model into one whose behaviour can be examined, questioned, and refined in the same language as clinical decision-making.

Declarations

Author contributions

Emmanuel Pio Pastore: Conceptualization; Methodology; Software; Validation; Formal Analysis; Investigation; Data Curation; Writing—Original Draft; Writing—Review & Editing; Visualization.

Funding

No specific funding was received for this work.

Competing interests

The author declares no competing interests.

Ethics approval

Not applicable. The analysis uses a public, de-identified historical dataset (PBC) accessed via Rdatasets/survival package.

Consent to participate / publish

Not applicable.

Data and code availability

Dataset: *https://vincentarelbundock.github.io/Rdatasets/csv/survival/pbc.csv*; R package source: *https://CRAN.R-project.org/package = survival*. Code: *https://github.com/emmanuel6474/surv-tcav-pbc*.

All repository code is under the MIT License. The dataset is publicly available; users must ensure compliance with its licence terms.

Reproducibility and code validation

An issue requesting Codechecker validation is open on the repository linked to this study.

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] K. D. Lindor, E. J. Bowlus, R. P. Boyer, et al. Primary Biliary Cholangitis: 2015 Practice Guidance from the American Association for the Study of Liver Diseases. Hepatology, 62(1): 373–422, 2015.

[8] G. M. Hirschfield, M. Mason, K. D. Luketic, et al. Primary Biliary Cholangitis. The Lancet, 392(10195): 1522–1536, 2018.

[9] E. J. Carey, K. D. Lindor. Current pharmacotherapy for primary biliary cirrhosis. Expert Opinion on Pharmacotherapy, 16(4): 609–621, 2015.

[10] M. M. Kaplan, A.M. Gershwin. Primary biliary cirrhosis. New England Journal of Medicine, 353: 1261–1273, 2005.

[11] B. Kim, M. Wattenberg, J. Gilmer, et al. Interpretability Beyond Feature Attribution: Quantitative Testing with Concept Activation Vectors (TCAV). In ICML, 2018.

[12] A. Ghorbani, J. Wexler, J. Zou, and B. Kim. Towards Automatic Concept-Based Explanations. In NeurIPS, 2019.

[13] P. W. Koh et al. Concept Bottleneck Models. In ICML, 2020.

[14] E. L. Kaplan and P. Meier. Nonparametric estimation from incomplete observations. JASA, 53(282):457–481, 1958.

[15] D. R. Cox. Regression models and life-tables. JRSS B, 34(2):187–220, 1972.

[16] J. H. Friedman. Greedy function approximation: A gradient boosting machine. Annals of Statistics, 29(5):1189–1232, 2001.

[17] T. Chen and C. Guestrin. XGBoost: A scalable tree boosting system. In KDD, 2016.

[18] XGBoost Documentation. Survival Analysis with AFT loss. *https://xgboost.readthedocs.io*. Accessed 2025.

[19] 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.

[20] 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.

[21] 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.

[22] T. M. Therneau and P.M. Grambsch. Modeling Survival Data: Extending the Cox Model. Springer, 2000.

[23] T. M. Therneau. A Package for Survival Analysis in R. *https://CRAN.R-project.org/package = survival*. Accessed 2025.

[24] V. Arel-Bundock. Rdatasets: Datasets from R packages. *https://vincentarelbundock.github.io/Rdatasets*. Accessed 2025.

[25] F. Pedregosa et al. Scikit-learn: Machine Learning in Python. JMLR, 12:2825–2830, 2011.

[26] C. Davidson-Pilon et al. lifelines: survival analysis in Python. Journal of Open Source Software, 4(40):1317, 2019.

Declarations

Ethics

No ethics approval was required or obtained. The study uses the publicly available Mayo Clinic Primary Biliary Cholangitis dataset as distributed in the R survival package and mirrored via Rdatasets, which contains de-identified data collected in the original study.

Funding

This research did not receive any specific grant from funding agencies in the public, commercial, or not-for-profit sectors.

Author Contributions

Emmanuel Pio Pastore: Conceptualization; Methodology; Software; Validation; Formal Analysis; Investigation; Data Curation; Writing – Original Draft; Writing – Review & Editing; Visualization.