— LIFECYCLE —
JOURNAL

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

Emmanuel Pio Pastore 1

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

Published on 02/11/2026
Outcomes Report
DOI: *10.71240/lcyc.732991*

Abstract

Survival prediction tools aid management of chronic liver diseases by informing decisions on monitoring and transplantation. In primary biliary cholangitis (PBC), classical risk scores discriminate well but are hard to interpret and treat variables in isolation, whereas clinicians think in terms of broader constructs such as cholestasis or coagulopathy. We introduce Surv-TCAV, a concept-based interpretability method for accelerated failure-time (AFT) survival models, and apply it to gradient-boosted trees on the Mayo Clinic PBC cohort (276 complete cases, 17 baseline variables). An XGBoost-AFT model was trained on 25 repeated 80/20 splits; performance was assessed using Harrell's C-index and the integrated Brier score, with Cox and Weibull AFT models as transparent baselines. TreeSHAP provided feature-level attributions, while Surv-TCAV probed model sensitivity along clinician-defined directions (cholestasis, coagulopathy, low albumin, older age and clinical complications). The boosted model achieved high discrimination (C-index ≈ 0.835) but weaker calibration (IBS ≈ 0.36), reflecting the usual flexibility–calibration trade-off. Both feature- and concept-level explanations aligned with hepatology practice: bilirubin, alkaline phosphatase, copper, low albumin, prolonged prothrombin time, age and clinical complications all reduced

predicted survival. These results show how boosted AFT models can be interpreted in familiar clinical terms. The PBC analysis is a methodological illustration; all code, figures and data are provided for reproducibility.

Evaluations

SERVICE	SUMMARY	VERSION	DATE	EVALUATED
Response to Reviewers		1	02/11/26	

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/tree/main

Paper

Clinically Interpretable Survival Prediction in Primary Biliary Cholangitis with TreeSHAP and GradientBoosted Models

By: Emmanuel Pio Pastore

1 Introduction

Machine-learning models offer increased flexibility and may capture interactions and nonlinearities, but their inner workings are often opaque, limiting their acceptance in clinical practice [1–3]. Post-hoc explanation techniques such as LIME and SHAP provide feature-level attributions that assign an importance value to each covariate for a given prediction [4,5]. For tree ensembles, TreeSHAP yields locally accurate and globally consistent attributions with an efficient algorithm that handles missing data [6]. However, feature-level explanations do not automatically recover the higher-order constructs used in hepatology. Primary biliary cholangitis (PBC) is a chronic autoimmune cholestatic disease that affects the small intrahepatic bile ducts [7–10]. Untreated, it progresses from biochemical disturbance to portal hypertension, fibrosis, cirrhosis and eventually liver failure. Timely risk stratification is therefore essential: clinicians must decide when to intensify surveillance, when to start or change therapy and when to refer a patient for transplantation [9]. Yet hepatologists rarely reason about a single number: they recognise cholestasis (high bilirubin and alkaline phosphatase), coagulopathy (prolonged prothrombin time), synthetic failure (low albumin) and overt clinical complications (ascites, oedema and vascular stigmata) as composite indicators of stage and prognosis. Recently, concept-based methods have been developed to test whether a model is sensitive to movements along user-defined directions in feature space rather than to individual variables alone [11–13]. The Testing with Concept Activation Vectors (TCAV) framework learns a direction representing a concept and estimates how moving along that direction affects the model output. This work extends that idea to survival modelling by proposing Surv-TCAV for gradient-boosted AFT models. In survival settings, time-to-event outcomes and censoring distributions are often summarised using Kaplan–Meier estimators [14]. Historically, risk scores for PBC have been derived from proportional hazards or parametric survival models that achieve good discrimination but are difficult to interpret and often treat each laboratory variable in isolation [15]. In line with the broader gradient boosting literature [16] and its efficient XGBoost implementation for survival with accelerated failure-time loss [17,18], Surv-TCAV provides a flexible framework for modelling non-linear effects while retaining clinically meaningful explanations. In the remainder of this work, model performance is assessed using Harrell's concordance index [19] and Brier-score–based measures of calibration [20–22]. The

empirical evaluation uses the well-known Mayo Clinic PBC dataset distributed with the *survival* package [23] and mirrored in the Rdatasets repository [24], and relies on standard Python tools such as scikit-learn [25] and lifelines [26] for data processing and survival analysis.

2 Methods

2.1 Cohort, variables, and experimental design

The analysis uses the Mayo Clinic PBC dataset distributed with the *survival* R package and mirrored via the Rdatasets repository [23,24]. The full cohort comprises 312 patients enrolled between 1974 and 1984 in a randomized trial; baseline variables were collected at study entry. Seventeen baseline covariates are considered: treatment assignment (trt), age, sex, ascites, hepatomegaly (hepato), spider angioma (spiders), peripheral oedema (edema), serum bilirubin (bili), serum cholesterol (chol), serum albumin (albumin), serum copper (copper), alkaline phosphatase (alk.phos), aspartate aminotransferase (ast), triglycerides (trig), platelet count (platelet), prothrombin time (protime) and histological stage (stage). Time zero is defined as the date of study entry; covariates were measured at or immediately before this time. The event of interest is death; liver transplantation is treated as rightcensoring in keeping with clinical practice. Of the 312 patients, 276 had complete data on all 17 variables and form the working dataset (*PBC276).*. Within this subset, there were approximately 125 deaths and 82 transplant censurings over a median followup of about nine years; the remaining patients were censored at last contact.

To obtain robust internal estimates without relying on a single split, twentyfive independent 80/20 train–validation partitions were generated. Stratification on the event indicator ensured that the proportion of deaths and censored observations was preserved across folds. Random seeds for splitting and model fitting are fixed in the accompanying code to enable exact reproduction. No observation appears in more than one validation set.

2.2 Modelling framework and conceptbased interpretability

Survival models were trained using gradient boosting with XGBoost’s accelerated failure time objective (*survival:aft),*, which directly models the logarithm of survival time while accommodating rightcensoring. Preliminary experiments compared Weibull, lognormal and extreme value (Gumbel) loss distributions; the Gumbel distribution with scale 0.4 offered numerically stable fits and performance comparable to Weibull on this dataset, so it was retained for all runs. Hyperparameters were fixed a priori: learning rate 0.02, maximum tree depth 10, minimum child weight 5.0, L2 regularisation λ 4.0, subsampling rate 0.6 for both rows and features, histogrambased tree builder with GPU acceleration when available, and 800 boosting rounds without early stopping. Fixing the number of rounds avoids peeking at the validation set and helps prevent information leakage. Full parameter values are listed in Table 1. Two transparent

comparators were fitted on the same training folds: a Cox proportional hazards model and a parametric Weibull AFT model [15,22]. These provide context for discrimination and calibration, as they represent wellunderstood baselines with closedform survival curves.

Interpretability was pursued at two complementary scales. At the **feature level**, TreeSHAP was used to compute local and global attributions for each tree ensemble. SHAP values quantify the contribution of each covariate to the predicted log survival time (AFT margin), taking into account nonlinearities and interactions. At the **concept level**, SurvTCAV adapts TCAV's directional sensitivity analysis to the AFT margin. Clinical concepts were defined a priori on the training data using simple, interpretable rules:

– *Cholestasis:* both bilirubin and alkaline phosphatase at or above the 75th percentile of their training distributions.

– *Coagulopathy:* prothrombin time at or above the 75th percentile.

– *Low albumin:* albumin at or below the 25th percentile.

– *Older age:* age at or above the 75th percentile.

– *Clinical complications:* presence of ascites, peripheral oedema or spider angioma.

For each concept, a balanced ℓ2regularised logistic regression was trained in the standardised numerical feature space to separate conceptpositive from conceptnegative samples; the normal vector of this classifier defines the concept direction. The direction is L2normalised and fixed. Directional sensitivity of the AFT location μ(x) was estimated for a one–standarddeviation step along each unitnorm concept direction using smoothed finite differences: each validation sample was perturbed by a small step of **h = 0.02** standard deviation units along the concept direction at 80 randomly selected points, the resulting changes in μ(x) were averaged and smoothed with a Gaussian kernel ($\sigma = 0.04$), and then normalised by the standard deviation of μ(x) across the validation set. Negative values indicate movement towards shorter predicted survival time on the log scale. This construction preserves clinically interpretable definitions while directly probing the model's internal representation.

2.3 Evaluation, uncertainty quantification, and reproducibility

Discrimination on each validation fold was assessed using **Harrell's concordance index (Cindex)** [19], which measures the ability to correctly rank pairs of survival times under censoring. Calibration was summarised by the **integrated Brier score (IBS)**[20,21], computed as the timeintegral of the Brier score up to an effective horizon τ_eff with inverseprobabilityofcensoring weights estimated from the Kaplan–Meier curve of the training fold's censoring distribution[14]. Integration was truncated when the censoring survival estimate fell below a small threshold to avoid unstable weights. Means and standard deviations of Cindex and IBS across the 25 validation

sets are reported; percentile bootstrap intervals (2.5th and 97.5th percentiles) based on 1 000 resamples of the splitwise metrics were computed to reflect uncertainty.

Timedependent Brier score curves were also calculated on a regular grid to visualise calibration over the followup period, and calibration curves at clinically relevant horizons were inspected. All analyses were carried out in Python 3 using the xgboost, pandas, numpy, scikitlearn and lifelines libraries. Code, configuration files and random seeds are available at *https://github.com/emmanuel6474/survtcavpbc* to enable exact reproduction of results.

Table 1. **Fixed XGBoostAFT hyperparameters used in all runs.**

Parameter	Value
objective	survival:aft
aft_loss_distribution	extreme (Gumbel)
aft_loss_distribution_scale	0.4
eta (learning rate)	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

Table 2. **Fair comparison on PBC276 (25×80/20 splits).** Mean ± standard deviation and bootstrap 95 % intervals for discrimination (Cindex) and calibration (IBS). Lower IBS indicates better calibration.

Model	Cindex (mean ± SD)	Cindex 95 % CI	IBS (mean ± SD)	IBS 95 % CI
XGBoostAFT	0.835 ± 0.039	[0.820, 0.850]	0.357 ± 0.018	[0.349, 0.363]
Cox proportional hazards	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]

Table 3. **SurvTCAV directional effects on the AFT location μ for a one–standarddeviation step along each concept direction (validation set).** Effects are reported as standardised shifts

Δμ / SD(μ); negative values indicate shorter predicted survival time. Positivity denotes 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

3 Results

Across the 25 train–validation splits, the XGBoostAFT model achieved a mean Cindex of 0.835 with standard deviation 0.039. Figure 1 (left) shows the distribution of Cindices; most splits fall within the mid0.8 range, indicating consistent discrimination across random partitions. Mean IBS was 0.357 with standard deviation 0.018, substantially higher (worse) than the IBS for Cox and Weibull models (Table 2). The right panel of Figure 1 plots IBS against Cindex for each split, highlighting that high discrimination does not necessarily imply good calibration; indeed, splits with similar Cindices exhibit IBS values ranging from 0.32 to 0.39.

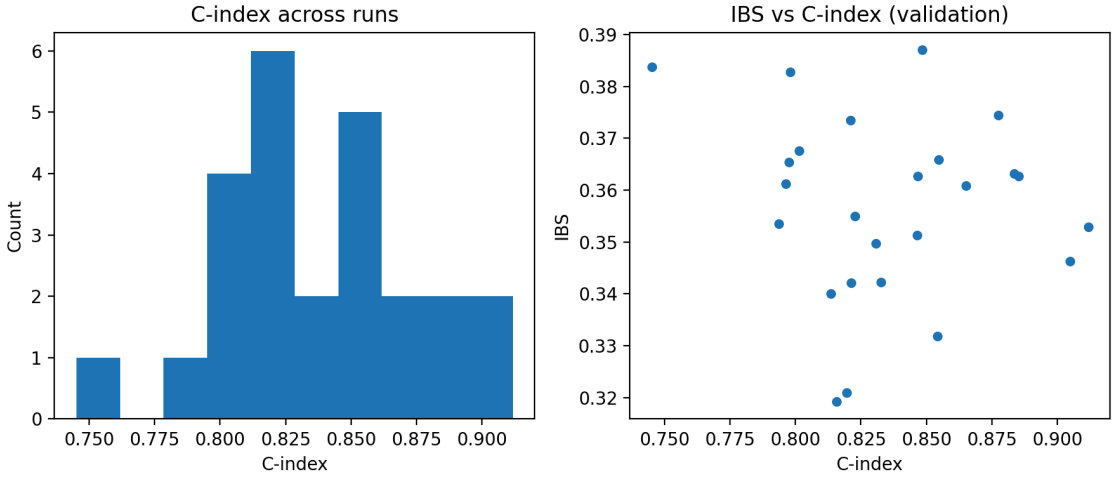

Figure 1: **Validation diagnostics across 25 splits.** Left: histogram of Harrell's C-index values; most splits fall in the mid–0.8 band. Right: IBS versus C-index; absence of a tight monotone relation reflects the different nature of discrimination and calibration.

Featurelevel explanations obtained with TreeSHAP (Figure 2) reveal clinically coherent patterns. Bilirubin, copper, albumin, alkaline phosphatase and prothrombin time are the largest contributors

to the AFT margin; high bilirubin and copper values push predictions towards shorter survival, while higher albumin prolongs survival. Age exerts a broadly monotone negative influence. Ascites and oedema introduce additional negative shifts even when laboratory values are moderate, capturing the step change associated with decompensation. Histological stage indicators align with disease severity, with stage 4 carrying the largest negative shift. SHAP values for sex_f are small but consistently positive, suggesting a residual protective association for females; however, the number of male patients in the cohort is small, so this observation is hypothesisgenerating rather than definitive.

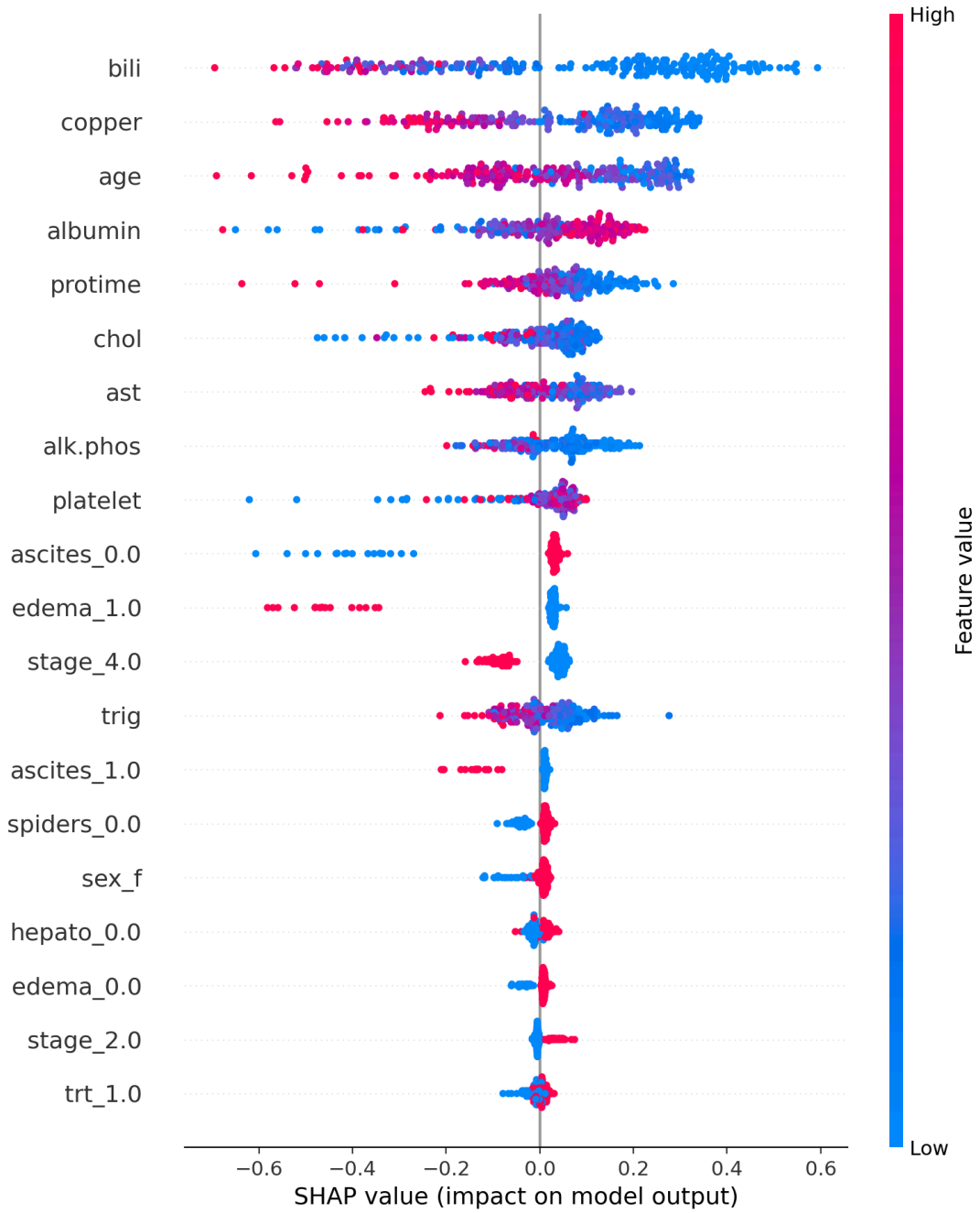

Figure 2: **TreeSHAP global summary for the XGBoostAFT model.** Each point represents a validation patient; colour encodes feature value (blue low, magenta high). Features are ordered by mean absolute SHAP value.

Conceptlevel sensitivities derived via SurvTCAV condense these patterns (Figure 3 and Table 3). Low albumin, older age and the clinical complications composite have moderate negative effects on the AFT location μ, on the order of −0.15 to −0.33 standard deviations. Cholestasis and coagulopathy

are less frequent at extreme thresholds and show smaller negative shifts. The qualitative agreement between TreeSHAP and SurvTCAV indicates that the boosted model aligns with hepatology practice: variables and concepts that clinicians associate with poor outcomes correspond to larger negative shifts in the model's logtime location parameter. To interpret these effect sizes in more concrete terms, recall that the AFT margin μ is defined on the natural logarithm of survival time. A change of Δμ in the logtime scale translates to a multiplicative factor of exp(Δμ) for the predicted survival time. For example, a concept effect of −0.33 implies exp(−0.33) ≈ 0.72, meaning that the predicted survival time decreases by about 28 %. Effect sizes in the range of −0.2 to −0.3 standard deviations therefore correspond to roughly one quarter to one third shorter expected survival. Such translations give clinicians an intuitive sense of the magnitude of concept effects while respecting the unitfree nature of the standardised effect.

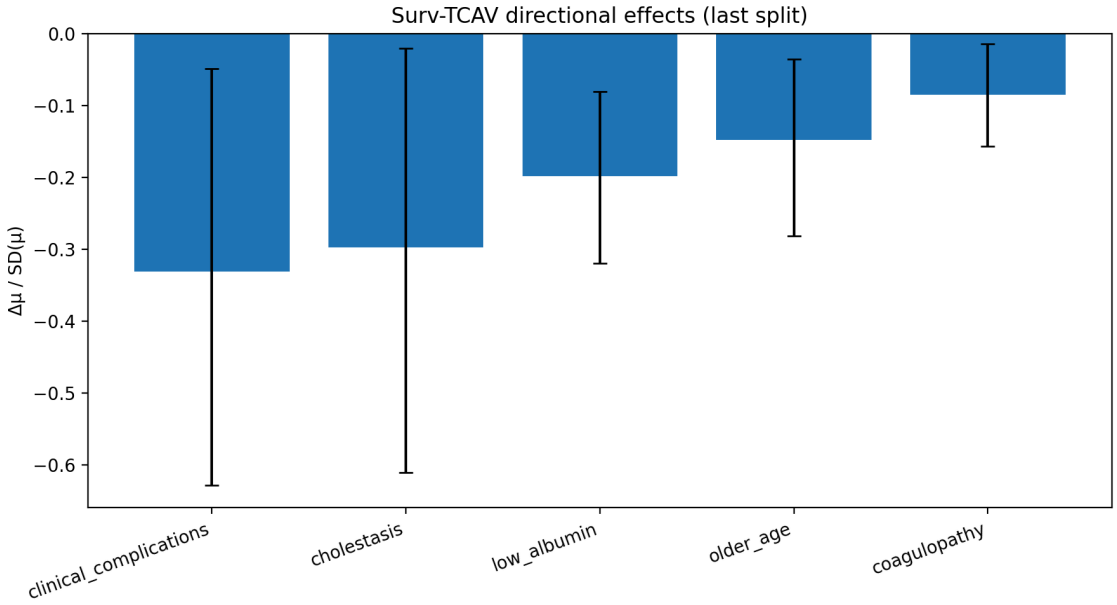

Figure 3: **SurvTCAV directional effects on the AFT location μ.** Bars show standardised effects with 95 % bootstrap confidence intervals; negative values correspond to shorter predicted survival.

Figure 4 displays timedependent Brier score curves for XGBoostAFT, Cox and Weibull models on a representative validation split. The boosted model consistently shows higher Brier scores across most of the followup, reflecting poorer calibration. Nevertheless, all three models maintain low Brier scores in the early years, which dominate the integrated Brier score due to the distribution of censoring. The calibration tradeoff observed in Table 2 is thus visually apparent over time.

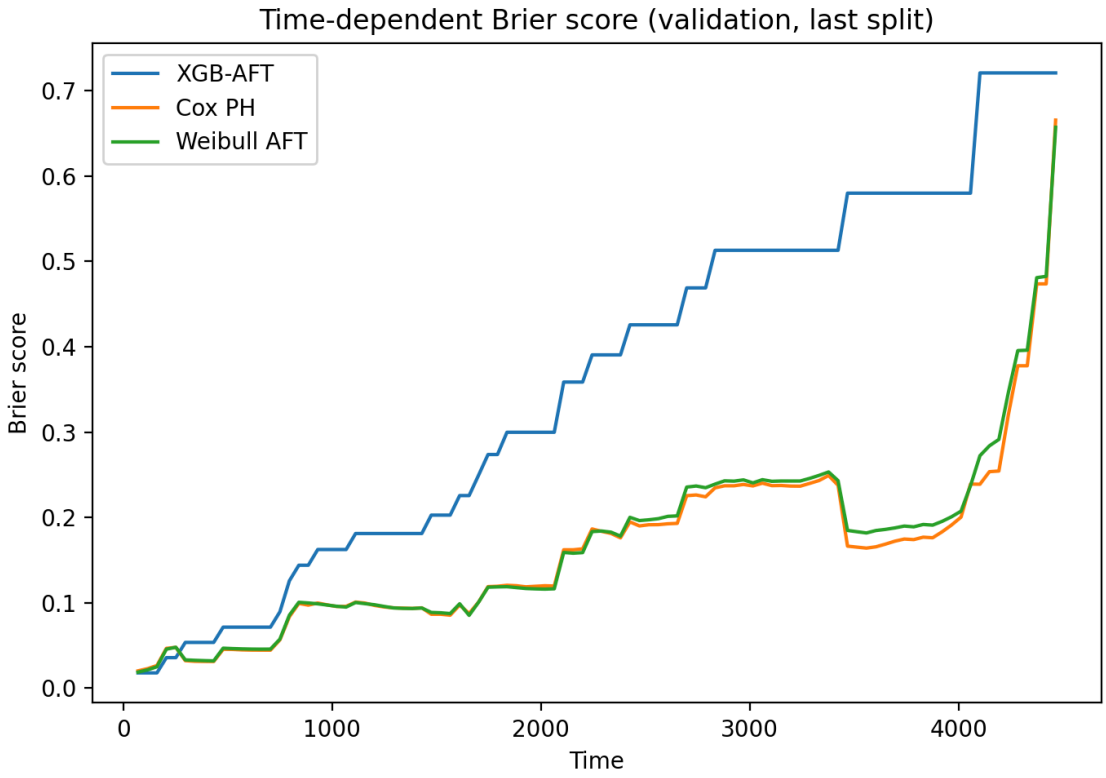

Figure 4: **Timedependent Brier score curves.** The XGBoostAFT model shows consistently higher Brier scores than Cox and Weibull comparators, illustrating the calibration tradeoff discussed in the text.

4 Discussion

4.1 Clinical and methodological interpretation

This study introduces SurvTCAV, a conceptbased interpretability framework for gradientboosted AFT models, and demonstrates its use in primary biliary cholangitis. On the historical PBC276 cohort the XGBoostAFT model achieved strong discrimination (Cindex ≈ 0.835), comparable with classical Cox and Weibull models, but showed markedly poorer calibration (IBS ≈ 0.36 vs. 0.15). This reflects a common tradeoff: boosting can capture nonlinearities and interactions that improve ranking, but without explicit probability alignment it yields survival curves that are too extreme. Recalibration techniques such as isotonic regression or parametric shrinkage applied to the AFT margin could reduce IBS while preserving ordering, but external validation would still be required before clinical use.

Feature and conceptlevel explanations reveal that the model has learned patterns consistent with hepatology. High bilirubin, copper and alkaline phosphatase and low albumin shorten predicted survival; ascites, oedema and spider angioma precipitate a step decline; older age reduces the logtime scale; and stage and prothrombin time follow intuitive gradients. SurvTCAV aggregates these signals into clinically named constructs, allowing auditors to ask whether “cholestasis” or “coagulopathy” matters to the model and by how much. In the PBC case, low albumin, older age and clinical complications exert the strongest effects, mirroring clinical practice where hypoalbuminaemia and decompensation herald poor outcomes. Cholestasis and coagulopathy have smaller negative effects, perhaps because their extreme thresholds are rare in this cohort or because their influence is partially absorbed by other variables.

4.2 Calibration, transportability, and limitations

The analysis highlights several limitations. The dataset derives from a single randomised trial conducted in the 1970s and early 1980s, with treatment arms and diagnostic practices that differ from current PBC management. Covariate distributions and event rates may not represent contemporary populations. Only complete cases were used, discarding patients with missing laboratory measurements; more sophisticated handling of missing data could enhance generalisability. Concept definitions relied on datadriven percentiles rather than guidelinebased thresholds (e.g., bilirubin > 2 mg/dL), so recalibration of concept boundaries may be needed in other cohorts. SurvTCAV measures directional sensitivity of the model, not causal effects; negative shifts along low albumin or coagulopathy directions do not imply that manipulating these variables in isolation will change survival. The small number of male patients means that any sexassociated effect is highly uncertain. Finally, internal validation via repeated random splits stabilises estimates in this moderatesized cohort, but independent datasets and prospective studies are essential to assess transportability and clinical utility. An alternative formulation would treat liver transplantation as a competing risk; this was not explored here but could be considered in future extensions. Future work could jointly model time to transplant and time to death rather than censor transplant events.

Despite these caveats, this case study demonstrates that modern boosting and conceptbased interpretability are not at odds with clinical reasoning. By expressing sensitivity along constructs that hepatologists already use, SurvTCAV complements granular feature attributions and offers a practical means to audit complex survival models. Extending this framework to larger, multicentre cohorts and incorporating clinical thresholds in concept definitions are promising directions. The open code base invites others to build upon these methods, explore recalibration strategies and test the approach in other diseases where timetoevent predictions and interpretability are critical.

5 Conclusions

SurvTCAV pairs gradientboosted AFT models with conceptlevel interpretability. Applied to primary biliary cholangitis, the framework delivers strong discrimination and transparent explanations at both feature and concept levels. While the boosted model's calibration is inferior to that of Cox and Weibull comparators, the interpretability tools reveal that it captures clinically meaningful patterns and can be audited in terms familiar to hepatologists. This work is intended as a methodological example rather than a readytouse prognostic score. Future efforts should explore recalibration, incorporate clinical thresholds for concept definitions, apply SurvTCAV to larger datasets and other diseases, and evaluate the clinical impact of combining feature and conceptbased explanations.

Acknowledgements

The author thanks Dr. Gabriel Antonio Videtta and Dr. Lucia Zaccato for their valuable suggestions.

References

1. F. DoshiVelez and B. Kim. Towards a rigorous science of interpretable machine learning. *arXiv preprint,* 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, M. Blasimme, V. Vayena, 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 *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining* (KDD 2016), 2016.

5. S. M. Lundberg and S. I. Lee. A unified approach to interpreting model predictions. In *Proceedings of the 31st International Conference on Neural Information Processing Systems* (NeurIPS 2017), 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: 2018 practice guidance from the American Association for the Study of Liver Diseases. *Hepatology,* 69(1):394–419, 2019.

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 and K. D. Lindor. Current pharmacotherapy for cholestatic liver disease. Expert Opinion on Pharmacotherapy, 13(18):2473–2484, 2012.

10. M. M. Kaplan and A.M. Gershwin. Primary biliary cirrhosis. *New England Journal of Medicine,* 353(12):1261–1273, 2005.

11. B. Kim, M. Wattenberg, J. Gilmer, et al. Interpretability beyond feature attribution: quantitative testing with concept activation vectors (TCAV). In *Proceedings of the 35th International Conference on Machine Learning* (ICML 2018), 2018.

12. A. Ghorbani, J. Wexler, J. Zou, and B. Kim. Towards automatic conceptbased explanations. In *Advances in Neural Information Processing Systems* (NeurIPS 2019), 2019.

13. P. W. Koh, J. Pierson, S. Sohoni, et al. Concept bottleneck models. In *Proceedings of the 37th International Conference on Machine Learning* (ICML 2020), 2020.

14. E. L. Kaplan and P. Meier. Nonparametric estimation from incomplete observations. *Journal of the American Statistical Association,* 53(282):457–481, 1958.

15. D. R. Cox. Regression models and lifetables. *Journal of the Royal Statistical Society: Series B (Methodological),* 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 *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining* (KDD 2016), 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 rightcensoring. *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.Rproject.org/package = survival* (accessed 2025).

24. V. ArelBundock. Rdatasets: datasets from R packages. *https://vincentarelbundock.github.io/Rdatasets* (accessed 2025).

25. F. Pedregosa, G. Varoquaux, A. Gramfort, et al. Scikitlearn: machine learning in Python. *Journal of Machine Learning Research,* 12:2825–2830, 2011.

26. C. DavidsonPilon, J. Kalbfleisch, S. Nelson, 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.

Competing Interests

The author declares no competing interests.

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.