

From Classification to Competing Risks: A Critical Re-Examination of a Widely-Used Cirrhosis Outcome Benchmark

Meseret Assefa Kerga¹

Axel Faes^{2,*}

¹ Data Science Institute (DSI), Hasselt University, Hasselt, Belgium.

² Biomedical Data Sciences, Data Science Institute (DSI) and Biomedical Research Institute (BIOMED), Hasselt University, Hasselt, Belgium.

Abstract

Analysing right-censored data with competing events as multi-class classification is a misspecification that recurs across clinical prediction benchmarks: it discards follow-up time and mismodels competing events as static labels. We study a widely-used instance, the Mayo primary biliary cirrhosis (PBC) cohort (418 patients), a machine-learning (ML) benchmark usually cast as three-class classification in which death and transplantation *compete*. We recast it as competing-risks survival and ask whether ML beats the five-variable Mayo score. Under a leakage-free cross-validation protocol we compared seven models (Mayo score, Cox, penalized Cox, random survival forests, gradient-boosted survival, DeepSurv, DeepHit), fitted cause-specific competing-risks models, and added longitudinal landmark and external-dataset analyses. Discrimination was high but *flat* across model families (Harrell’s C 0.81–0.83). No ML model, deep networks included, beat the Mayo score (C = 0.833, integrated Brier 0.143): a ceiling set by the static variables, not model capacity. The transplant outcome, poorly predicted by classifiers (F1 $\approx$ 0.21), reached cross-validated concordance 0.82 (25 events; 95% CI 0.80–0.84) as a competing event. Cause-specific models recovered *opposite* age effects: older age raised the death hazard (HR 1.36 per SD) but lowered the transplant hazard (HR 0.74 per SD), structure the three-class formulation cannot express. Longitudinally, a baseline measurement’s death-risk concordance decayed from 0.87 to 0.64 over six years but was restored to 0.78 by the most recent measurement (paired Δ C up to +0.14). Correcting the formulation, not adding complexity, yields calibrated, coherent risk for both competing outcomes; on small tabular clinical data, problem formulation matters more than model sophistication.

Keywords: liver cirrhosis; survival analysis; competing risks; machine learning; benchmark re-examination; reproducibility; clinical risk prediction

1 Introduction

A recurring methodological problem in clinical prediction is that right-censored time-to-event data with competing events are analysed as multi-class classification. This substitutes a wrong estimand for the prognostic one: it discards follow-up time and treats a competing event as a static outcome label, biasing absolute-risk estimates and obscuring cause-specific structure. The consequences are concrete and measurable in a widely-used cirrhosis outcome benchmark, which we use as the worked instance throughout this paper.

Liver cirrhosis is a leading cause of morbidity and mortality worldwide, responsible for over a million deaths annually and a substantial healthcare burden [1]. Accurate prognostication is central

*Corresponding author: axel.faes@uhasselt.be

to its management. It guides the timing of interventions, the prioritisation of candidates for liver transplantation, and the allocation of scarce resources. Decades of clinical research have produced validated prognostic scores for this purpose, of which the Mayo PBC natural-history model [2] and the Model for End-Stage Liver Disease (MELD) [3] are the most established.

The public “Cirrhosis Patient Survival Prediction” dataset [2, 4], comprising 418 patients from the Mayo Clinic PBC trial, has become a popular machine-learning (ML) benchmark. Most analyses, including numerous student and applied-ML studies, treat it as a *three-class classification* problem. They predict the recorded status label (alive, transplanted, or deceased) with classifiers such as random forests and gradient boosting, frequently combined with synthetic minority over-sampling (SMOTE) [5] to counter class imbalance [6, 7]. Reported macro-F1 scores hover around 0.6. The transplant class is invariably the worst predicted, with F1 scores near 0.21.

This framing answers the wrong question for prognosis. The dataset records, for each patient, a follow-up time and a terminal status. Patients still alive at last contact are *right-censored*, not negative examples, and a patient who receives a transplant is removed from the at-risk pool for death. Death and transplantation are therefore *competing risks* [8, 9]. Predicting the realised status under fixed administrative follow-up is itself a coherent task, but it is not the prognostic question: ordinary classification discards the follow-up time and mismodels transplantation as a static label rather than a competing event, which is precisely why the transplant class defeats classifiers. For the clinically relevant absolute-risk and cause-specific estimands, the correct formulation is time-to-event survival analysis with competing risks.

This paper is a critical re-examination of that practice. We quantify what the mis-specification costs, provide the principled competing-risks treatment, and release a reproducible framework so the benchmark can be analysed correctly. Our contributions are:

1. We recast the benchmark as competing-risks survival, modelling death and transplantation as competing events, and characterise the cumulative incidence and cause-specific predictors of each. Recent work has applied single-event survival models to this cohort [10], but, to our knowledge, none treat death and transplantation as competing events, and none benchmark against the established clinical score.
2. We benchmark seven survival models, from the historical Mayo clinical score to deep neural survival networks (DeepSurv, DeepHit), under a leakage-free protocol with survival-appropriate metrics.
3. We show that discrimination is *flat* across all model families and that no ML model beats the five-variable Mayo score. This is evidence of a ceiling set by the static feature set rather than by model capacity, consistent with the broader finding that simple models are hard to beat on small tabular clinical data.
4. We quantify what the reframing expresses that classification cannot. Modelled as a competing event, the transplant outcome is discriminated with cross-validated $C = 0.82$ (25 events; 95% CI 0.80–0.84); a classifier’s per-class F1 (near 0.21 in prior work) and a concordance index are not comparable, but only the survival formulation yields time-resolved competing-event risk. The treatment also recovers clinically-established structure that a three-class model cannot represent: the opposite directions in which age acts on death versus transplant, consistent with referral and listing practice [11].
5. We extend the analysis to the *longitudinal* cohort, showing with time-varying competing-risks models and landmarking that a baseline measurement’s prognostic value decays over follow-up but is recovered by updating biomarkers, a dynamic prediction gain no prior work on this cohort captures.

6. We show, across three further clinical datasets, that the benefit of model complexity scales with sample size and is negligible at the cirrhosis cohort’s scale, contextualising the null model-comparison result.
7. We release the complete, leakage-free pipeline as an openly reproducible framework for competing-risks survival analysis of this and comparable clinical cohorts.

2 Background and Related Work

Clinical prognostic scores. Cirrhosis prognosis has long relied on rule-based scores. The Child–Pugh score and MELD [3] are widely used, the latter combining serum creatinine, bilirubin and the international normalised ratio (INR) to predict short-term mortality. For primary biliary cirrhosis specifically, the Mayo natural-history model [2] predicts survival from five routinely available variables (bilirubin, albumin, age, prothrombin time and edema) and was derived on the very cohort distributed as this dataset. MELD cannot be computed here, because the dataset contains neither serum creatinine nor INR. Its `Prothrombin` field is prothrombin time in seconds, from which INR is not recoverable without the laboratory ISI and control time. The Mayo PBC score, by contrast, uses only variables present in the data and is therefore the appropriate clinical baseline. Two contemporary PBC-specific scores, the GLOBE score [12] and the UK-PBC risk score [13], are likewise non-computable here, as they require treatment-response biochemistry and markers absent from this dataset. More recent work has explored dynamic, longitudinally-updated models for cirrhosis survival [14].

ML on this dataset. Most ML studies on this dataset are classification-based. Representative examples predict the status label with tree ensembles and neural networks [6], and the dataset underpinned a multi-class prediction competition [7]. Such studies use `N_Days` only as a static feature, if at all, and treat transplantation as an ordinary label. Recently, Obite et al. [10] did apply single-event survival models to this PBC cohort, comparing Cox, random survival forests, DeepSurv, DeepHit and a proposed factor-enhanced DeepSurv by concordance and Brier score, and confirmed that the data are better treated as time-to-event. That work censors the competing transplant event rather than modelling it, and does not benchmark against the historical clinical score. Our contribution is distinct on both counts: we model death and transplantation as *competing* events, and we benchmark every model against the Mayo PBC score. That is the gap this paper addresses.

Survival machine learning. Beyond the semiparametric Cox proportional-hazards model [15], ML survival methods include random survival forests [16] and, more recently, neural approaches such as DeepSurv (a neural Cox model) [17] and DeepHit (a discrete-time model with native competing-risks support) [18]. Whether such models outperform Cox or simple clinical scores on small tabular datasets is contested. On general tabular data, a recurring finding is that tree ensembles and well-specified simple models are hard to beat with deep networks when sample sizes are limited [19, 20]; we test whether this pattern holds in the survival setting.

Competing risks. When several event types preclude one another, analysing one while censoring the others biases absolute-risk estimates [8]. Two complementary models are standard, each answering a different question [21]. Cause-specific Cox models (one hazard per event type) address aetiological questions, while the Fine–Gray subdistribution-hazard model [9] addresses absolute risk, with the Aalen–Johansen estimator [22] providing nonparametric cumulative incidence.

3 Materials and Methods

3.1 Dataset and cohort

We used the Mayo PBC cohort (UCI “Cirrhosis Patient Survival Prediction”) [2, 4]. The cohort is specifically primary biliary cirrhosis (now primary biliary cholangitis), an autoimmune cholestatic disease, so it is not representative of cirrhosis from viral, alcoholic or metabolic causes; we retain the historical name to match the dataset. It comprises 418 patients with a follow-up time (`N.Days`) and terminal status (“C” for censored/alive, $n = 232$; “D” for death, $n = 161$; “CL” for censored at liver transplant, $n = 25$), plus 17 demographic, clinical and laboratory variables. Median follow-up was 1730 days (range 41–4795). Several laboratory variables had substantial missingness, reflecting the well-known structure in which 106 patients enrolled without the full panel. Missing values were imputed inside each cross-validation fold (below).

3.2 Survival reformulation

We replaced the three-class target with time-to-event structure. For overall survival, the event was death and transplantation was treated as a competing event (cause-specific censoring). We additionally defined the transplant endpoint, with transplant as the event and death censored. For the competing-risks description we used the event code $\{0 : \text{censored}, 1 : \text{death}, 2 : \text{transplant}\}$.

3.3 Clinical baseline: the Mayo PBC score

As the clinical reference we computed the Mayo natural-history risk score [2],

$$R = 0.871 \ln(\text{bilirubin}) - 2.53 \ln(\text{albumin}) \\ + 0.039 \text{ age} + 2.38 \ln(\text{prothrombin}) + 0.859 \text{ edema},$$

with edema coded 0/0.5/1. We used the score as a single-covariate Cox model, fitting the baseline hazard within each training fold so that it produces calibrated absolute-risk curves comparable to the other models.

3.4 Survival models

We compared seven models: (i) the Mayo score; (ii) Cox proportional hazards [15] with ridge stabilisation; (iii) penalized Cox (elastic net); (iv) random survival forests [16]; (v) gradient-boosted survival (Cox loss); (vi) DeepSurv [17]; and (vii) DeepHit [18]. Tree and deep models used all 17 variables.

3.5 Preprocessing

All preprocessing was fit *inside* each training fold to prevent leakage, a flaw in prior analyses that preprocess before splitting. Numerical variables were median-imputed, log-transformed where skewed, and standardised. Categorical variables were mode-imputed and one-hot encoded. We applied no oversampling (e.g. SMOTE), because resampling is inappropriate for censored survival data.

3.6 Competing-risks analysis

We estimated cause-specific cumulative incidence functions for death and transplantation with the Aalen–Johansen estimator [22], and fit cause-specific Cox models for each event (a parsimonious three-covariate model for the rare transplant endpoint). Continuous covariates were standardised, so hazard ratios are per standard deviation.

3.7 Dynamic analysis on the longitudinal cohort

The static benchmark uses a single measurement per patient. To test whether updating biomarkers over follow-up improves prediction, we additionally analysed the sequential Mayo PBC data (`pbcseq`: 312 patients, 1945 serial measurements), which carries the same competing outcomes. We fit time-varying cause-specific Cox models (counting-process format) for death and transplantation using the most recent bilirubin, albumin, prothrombin time, edema and age. We then performed a landmark analysis at 2, 4 and 6 years. Among patients still at risk, we predicted 3-year death risk from either the baseline (day-0) measurement or the most recent measurement carried forward, comparing repeated 5-fold cross-validated concordance between the two covariate sets.

3.8 Generalisation across datasets

To test whether the limited benefit of model complexity is specific to the small cirrhosis cohort or reflects a broader pattern, we replicated the simple-versus-complex comparison on three widely-used public clinical survival datasets via `pycox`: METABRIC ($n = 1904$), GBSG ($n = 2232$) and SUPPORT ($n = 8873$). On each we compared Cox against random survival forests, gradient boosting and DeepSurv under the same standardised, repeated stratified cross-validation, reporting Harrell’s concordance.

3.9 Transportability to an independent cohort

We assessed whether a model fit on the Mayo PBC cohort transports to an independent cirrhosis cohort, using the Cirrhosis subgroup of the SUPPORT study ($n = 508$). Only markers common to both cohorts were available, namely bilirubin, albumin and age, because SUPPORT records neither prothrombin time nor edema. We therefore trained a 3-covariate cause-specific death model on the PBC cohort and applied it to SUPPORT. SUPPORT enrolled acutely hospitalised, severely ill patients, so case-mix and baseline risk differ markedly between the two cohorts. We accordingly evaluated *discrimination* (transportability of risk ordering, Harrell’s C) rather than absolute-risk calibration, for the death endpoint, since transplant is not ascertained as a competing event in SUPPORT.

3.10 Evaluation

Models were evaluated by repeated stratified 5-fold cross-validation (4 repeats, stratified on event type to balance the rare transplant cases). Confidence intervals were obtained by bootstrapping the per-fold metric values; they therefore reflect cross-validation variability rather than independent patient-level resampling, and on the rarest strata (the 25-event transplant endpoint and the later landmarks) they should be read as lower bounds on the true sampling uncertainty. The metrics were Harrell’s [23] and Uno’s [24] concordance indices, time-dependent AUC at 1/3/5 years, and the integrated Brier score [25]. Calibration was assessed at five years against Kaplan–Meier estimates. Reporting follows the TRIPOD guideline [26].

Table 1: Cohort characteristics by outcome. Continuous values are median [IQR].

	Overall	Alive	Death	Transpl.
<i>n</i>	418	232	161	25
Age, y	51 [43–58]	50 [41–57]	53.5 [47–61]	41 [37–47]
Bilirubin, mg/dL	1.4 [0.8–3.4]	0.9 [0.6–1.6]	3.2 [1.4–7.1]	3.1 [1.3–3.5]
Albumin, g/dL	3.5 [3.2–3.8]	3.6 [3.4–3.8]	3.4 [3.1–3.6]	3.5 [3.2–3.8]
Prothrombin, s	10.6 [10–11.1]	10.4 [9.9–10.8]	11.0 [10.5–11.8]	10.3 [10–10.7]
Mayo score	4.8 [4.1–5.9]	4.3 [3.8–4.8]	6.0 [5.0–6.8]	5.0 [4.4–5.2]
Ascites	6%	<1%	14%	0%
Stage 4	34%	22%	52%	40%

3.11 Implementation

Analyses used `scikit-survival` 0.27 [27], `lifelines` 0.30 and `pycox` 0.3 in Python 3.13, with a fixed random seed (42) throughout for reproducibility. The proportional-hazards assumption was checked by Schoenfeld residual tests. Code and the full pipeline are available (see Data Availability).

4 Results

4.1 Cohort

Table 1 summarises the cohort by outcome. Deceased patients were older, with higher bilirubin and Mayo scores. Transplanted patients were markedly *younger* (median age 41 vs. 53.5 years for deaths) at a similar median bilirubin (3.1 vs. 3.2 mg/dL, though over a narrower range), an early signal of the competing-risks structure.

4.2 Cumulative incidence of competing outcomes

The Aalen–Johansen cumulative incidence (Fig. 1) shows death accumulating steadily (7.2%, 19.7% and 29.2% at 1, 3 and 5 years) while transplantation remained rare (0%, 2.4%, 4.5%). Treating transplantation as just another class ignores both its rarity and its competing relationship with death.

4.3 Discrimination is flat across model families

All seven models achieved high and statistically indistinguishable discrimination for the death endpoint (Table 2; Supplementary Fig. S1). Harrell’s C ranged only from 0.812 (DeepHit) to 0.833 (Mayo). No ML model, whether linear, penalized, ensemble or deep, outperformed the five-variable Mayo clinical score, which also achieved the best integrated Brier score (0.143). The deep models (DeepSurv, DeepHit) were no better, and DeepHit calibrated worst (IBS 0.202), consistent with deep discrete-time models being data-hungry. This flatness indicates that the ~ 0.83 ceiling is set by the information in this static set of variables rather than by model capacity; the longitudinal analysis below shows that further signal becomes available when measurements are updated over time.

Two checks support this reading. First, paired per-fold comparisons against the Mayo score confirm the flatness statistically rather than by eye: Cox, penalized Cox, random survival forests and DeepSurv were indistinguishable from the score (paired ΔC 95% CIs included zero), while gradient boosting (ΔC -0.012 [-0.022 , -0.003]) and DeepHit (-0.021 [-0.033 , -0.011]) were significantly *worse*. Second, the ceiling is not an artefact of the Mayo score’s coefficients having

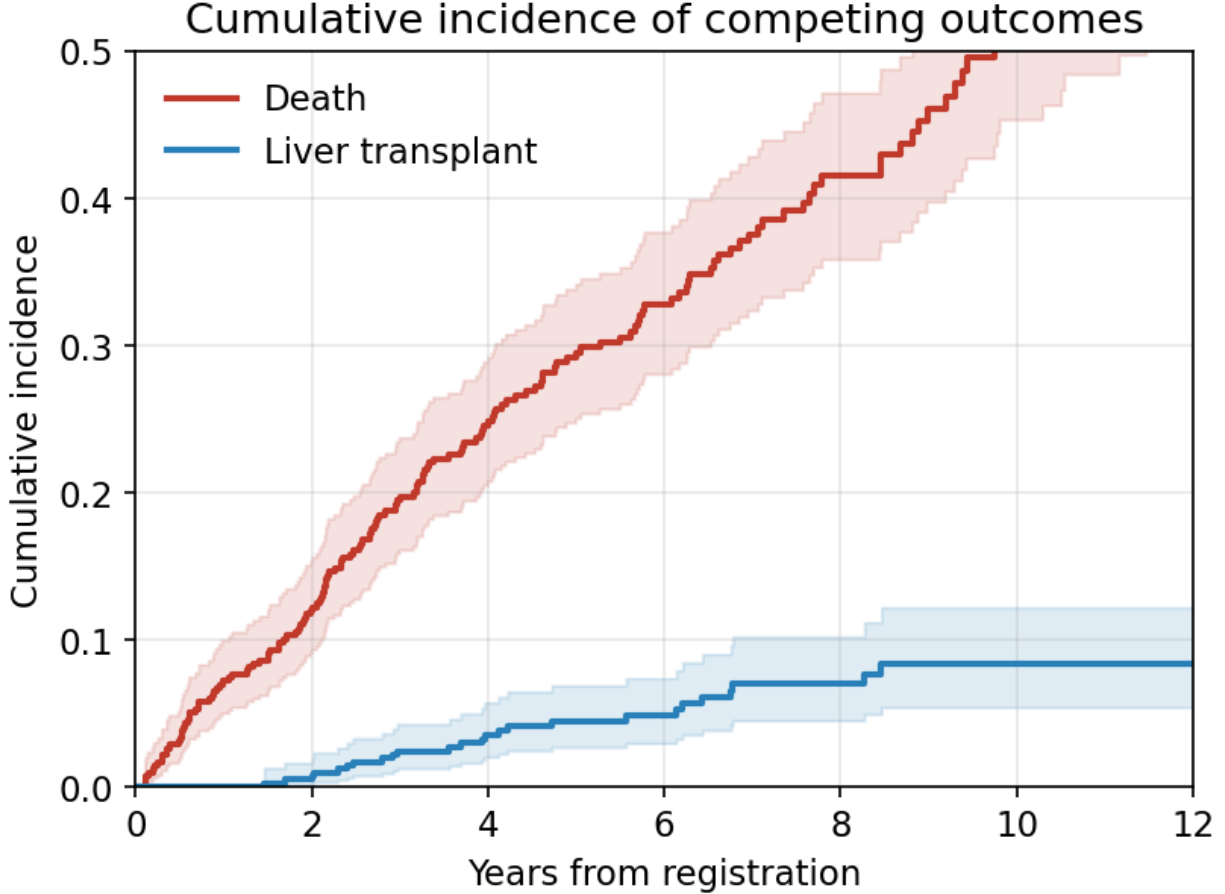


Figure 1: Cumulative incidence of death and liver transplantation (Aalen–Johansen, with 95% confidence bands).

been derived on this cohort: refitting a Cox model on the same five variables *inside* each fold gave identical discrimination ($C = 0.832$ vs. 0.833 for the fixed score; paired ΔC -0.001 $[-0.006, +0.004]$), so the externally-fixed score holds no in-sample advantage here.

4.4 Augmenting the clinical score does not help

Adding the biomarkers the Mayo score omits (copper, SGOT, platelets, cholesterol, alkaline phosphatase, fibrosis stage, ascites, and others) on top of the score yielded no improvement. The paired Δ Harrell’s C was $+0.0006$ (95% CI $-0.005, +0.006$) for a Cox extension and negative for a boosted extension. The established five-variable score is effectively at the information ceiling of this cohort.

4.5 Generalisation: complexity helps only as data grow

Replicating the comparison on three external clinical datasets (Table 3; Supplementary Fig. S2) indicated that the limited value of complexity is consistent with a sample-size-dependent pattern and not specific to the cirrhosis cohort. These datasets are single-event, so they bear only on the complexity-versus-Cox question, not on the competing-risks or clinical-score comparisons central to the cirrhosis analysis. On the small-to-moderate datasets, complex models barely exceeded

Table 2: Death-endpoint discrimination and calibration across model families (repeated stratified CV; 95% bootstrap CI). Best value per column in **bold**.

Model	Harrell’s C	Uno’s C	AUC @ 5y	IBS (lower better)
Mayo PBC score	0.833 [0.815, 0.852]	0.787 [0.764, 0.809]	0.907 [0.886, 0.926]	0.143 [0.130, 0.157]
Cox PH	0.830 [0.813, 0.847]	0.774 [0.755, 0.792]	0.904 [0.882, 0.922]	0.151 [0.138, 0.165]
Coxnet (elastic net)	0.833 [0.815, 0.849]	0.779 [0.761, 0.797]	0.906 [0.885, 0.925]	0.149 [0.136, 0.163]
Random survival forest	0.826 [0.810, 0.841]	0.787 [0.771, 0.805]	0.894 [0.874, 0.913]	0.153 [0.144, 0.162]
Gradient boosting	0.821 [0.805, 0.836]	0.772 [0.757, 0.787]	0.891 [0.871, 0.909]	0.153 [0.141, 0.165]
DeepSurv	0.828 [0.810, 0.845]	0.772 [0.751, 0.790]	0.896 [0.874, 0.916]	0.154 [0.141, 0.167]
DeepHit	0.812 [0.797, 0.827]	0.768 [0.752, 0.784]	0.891 [0.872, 0.910]	0.202 [0.184, 0.223]

Table 3: Harrell’s C across three external clinical survival datasets (repeated stratified CV).

Dataset	<i>n</i>	Cox	RSF	GBM	DeepSurv
METABRIC	1904	0.635	0.643	0.642	0.627
GBSG	2232	0.661	0.673	0.675	0.672
SUPPORT	8873	0.570	0.615	0.615	0.611

Cox (METABRIC, $n = 1904$: best gain $+0.008$; GBSG, $n = 2232$: $+0.014$), and DeepSurv was occasionally worse. Only on the large SUPPORT dataset ($n = 8873$) did they gain materially (up to $+0.046$). The cirrhosis cohort ($n = 418$) lies at the small-data end of this range, where little benefit from complexity should be expected, consistent with no model beating the clinical score.

4.6 Both competing endpoints are well predicted

Under cross-validation, both endpoints were discriminated well (Table 4). The death endpoint reached Harrell’s C = 0.826 (95% CI 0.816, 0.836). The *transplant* endpoint reached Harrell’s C = 0.822 (95% CI 0.799, 0.843) and Uno’s C = 0.813 (0.790, 0.836) when modelled as a competing event, encouraging discrimination, though necessarily imprecise given only 25 events. Classification approaches predict this outcome poorly (F1 ≈ 0.21). An F1 score and a concordance index are not directly comparable, but the qualitative point holds. A classifier cannot express time-resolved risk for a rare competing event. The survival formulation can.

4.7 Cause-specific predictors: opposite age effects

Cause-specific Cox models (Table 5; Supplementary Fig. S3) recovered structure that classification cannot represent. Death was driven by older age (HR 1.36 per SD, 95% CI 1.16–1.59), higher bilirubin (HR 1.89, 1.57–2.28), edema (HR 2.26, 1.32–3.87) and advanced fibrosis (stage 4 HR 1.56, 1.10–2.22), with albumin protective. The apparent (in-sample) cause-specific concordance for death was 0.85. For age, the transplant hazard moved in the *opposite* direction (HR 0.74 per SD, 0.58–0.95): younger patients with deteriorating liver function (higher bilirubin, HR 1.30, 1.01–1.67) were preferentially transplanted, while older patients with the same biochemistry died. This pattern is clinically established. It reflects referral and listing practice rather than biology; because the cohort was enrolled in 1974–1984, before MELD-based allocation, the mechanism is era-appropriate candidate selection, though its direction matches the modern competing-risks waitlist literature in which younger candidates are preferentially transplanted while older listed patients die [11]. It serves here as a validity check. The competing-risks treatment reproduces it directly, whereas collapsing the two events into one label makes the distinction structurally inexpressible.

Table 4: Cross-validated discrimination for each competing endpoint (cause-specific Cox; 95% bootstrap CI).

Endpoint	Events	Harrell’s C	Uno’s C
Death	161	0.826 [0.816, 0.836]	0.779 [0.764, 0.793]
Transplant	25	0.822 [0.799, 0.843]	0.813 [0.790, 0.836]

Table 5: Cause-specific hazard ratios (per SD for continuous covariates). Selected covariates; full model in the repository.

Covariate	Death HR [95% CI]	Transplant HR [95% CI]
Age	1.36 [1.16, 1.59]	0.74 [0.58, 0.95]
Bilirubin (log)	1.89 [1.57, 2.28]	1.30 [1.01, 1.67]
Albumin	0.80 [0.68, 0.95]	0.88 [0.68, 1.13]
Prothrombin (log)	1.20 [1.04, 1.38]	–
Edema	2.26 [1.32, 3.87]	–
Stage 4	1.56 [1.10, 2.22]	–

4.8 Dynamic prediction on the longitudinal cohort

Modelling the biomarker trajectory rather than a single snapshot both reproduced and sharpened the competing-risks picture. In time-varying cause-specific Cox models (Table 6), the most recent measurements were strongly prognostic for death (bilirubin HR 1.82 per SD, albumin 0.63, prothrombin 1.40, edema 2.47; all $p < 0.001$), and the *opposite* age effect persisted in the longitudinal setting (death HR 1.25, $p = 0.001$; transplant HR 0.83, $p = 0.03$). The landmark analysis (Fig. 2) shows why updating matters. A baseline measurement loses prognostic value as it ages: 3-year death-risk concordance fell from 0.87 at the 2-year landmark ($n = 278$, 55 deaths) to 0.64 at 6 years ($n = 166$, 26 deaths), whereas using the most recent measurement decayed far more slowly, from 0.89 to 0.78 over the same span. The cross-validated paired advantage grew with follow-up and excluded zero at every landmark ($\Delta C = +0.011$ [95% CI +0.004, +0.017] at 2y; +0.059 [+0.042, +0.077] at 4y ($n = 225$, 34 deaths); +0.139 [+0.116, +0.162] at 6y). The later landmarks rest on few events, and the intervals reflect cross-validation fold variability rather than independent resampling. The pattern suggests that, in these data, recent measurements carry information a stale baseline cannot.

4.9 Calibration and risk stratification

The survival models produced well-calibrated five-year predictions (Supplementary Fig. S4) and cleanly separated patients into low-, intermediate- and high-risk groups by out-of-fold predicted risk (Supplementary Fig. S5). Calibration of this kind is a prerequisite for absolute-risk communication rather than mere ranking, though clinical utility was not formally assessed. A decision-curve analysis of net benefit remains future work.

4.10 Transportability to an independent cirrhosis cohort

Discrimination did *not* transport to the SUPPORT cirrhosis subgroup ($n = 508$, 329 deaths). The 3-marker PBC model, which reached internal cross-validated $C = 0.823$, fell to $C = 0.558$ externally, barely above chance, and even refitting the same markers within SUPPORT reached only $C = 0.594$. The albumin coefficient reversed sign, from protective in PBC (HR 0.38) to

Table 6: Time-varying cause-specific hazard ratios on the longitudinal cohort (per SD for continuous covariates).

Covariate	Death HR [95% CI]	Transplant HR [95% CI]
Bilirubin (log)	1.82 [1.58, 2.09]	1.28 [1.07, 1.52]
Albumin	0.63 [0.55, 0.72]	0.87 [0.73, 1.04]
Prothrombin (log)	1.40 [1.27, 1.55]	1.16 [0.97, 1.37]
Age	1.25 [1.10, 1.42]	0.83 [0.69, 0.99]
Edema	2.47 [1.72, 3.56]	1.28 [0.74, 2.20]

mildly harmful in SUPPORT (HR 1.16). This is best read as population specificity rather than model failure. SUPPORT enrolled acutely hospitalised patients (65% died within a median 234 days), a fundamentally different phenotype from the chronic outpatient PBC trial. Chronic-PBC prognostic structure therefore does not carry to acute decompensated cirrhosis, a caution that prognostic models, like the formulation itself, are tied to clinical context. A fair external test would require an independent *chronic* cirrhosis cohort with the full marker set, which is not currently available in the public domain.

5 Discussion

Two results stand out. The first is the substantive contribution. Recasting this benchmark as competing-risks survival expresses risk that classification cannot. Modelled as a competing event, the transplant outcome reaches a cross-validated concordance of 0.82 (25 events; 95% CI 0.80–0.84), an encouraging if imprecise result, and cause-specific modelling recovers the clinically-established opposite age effect, in which younger patients with failing livers are transplanted while older ones die [11]. We present this not as a new discovery but as a validity check: the reframing reproduces known clinical structure directly, whereas a three-class formulation cannot express it at all.

The second is a cautionary result. Across seven model families spanning a five-variable clinical score, penalized linear models, tree ensembles and deep neural survival networks, discrimination was flat at $C \approx 0.83$, and no model beat the decades-old Mayo score. Augmenting the score with additional biomarkers did not help either, and a paired per-fold test confirmed that no model significantly beat the score while two complex models were significantly worse. The ceiling is therefore limited by the information in this static set of variables rather than by model capacity: refitting the five Mayo variables fresh within each fold matched the externally-derived score exactly, ruling out an in-sample-baseline artefact, and the longitudinal analysis shows the static snapshot, not the achievable signal, is the binding constraint. Our replication across three external (single-event) datasets is consistent with the complexity benefit being size-dependent. Complex models exceeded Cox only marginally on cohorts of one-to-two thousand patients and meaningfully only on the largest ($n \approx 9000$). These datasets speak only to the complexity-versus-Cox question, but they place $n = 418$ where little benefit should be expected. This replicates, in the survival setting, an established pattern on general tabular data, where simple and tree-based models are difficult to beat with deep networks at limited sample sizes [19, 20], and it tempers the common narrative that more elaborate ML yields better cirrhosis prognosis. Recent deep-survival work on this very cohort reports that an enhanced network outperforms simpler survival models [10], but compares deep models only against one another. Our results indicate that the more consequential benchmark, the five-variable clinical score, remains unbeaten, a comparison that prior work omitted.

The practical implication is that ML is far from useless here. Its value lies in *formulation*

Updating biomarkers restores discrimination as baseline goes stale

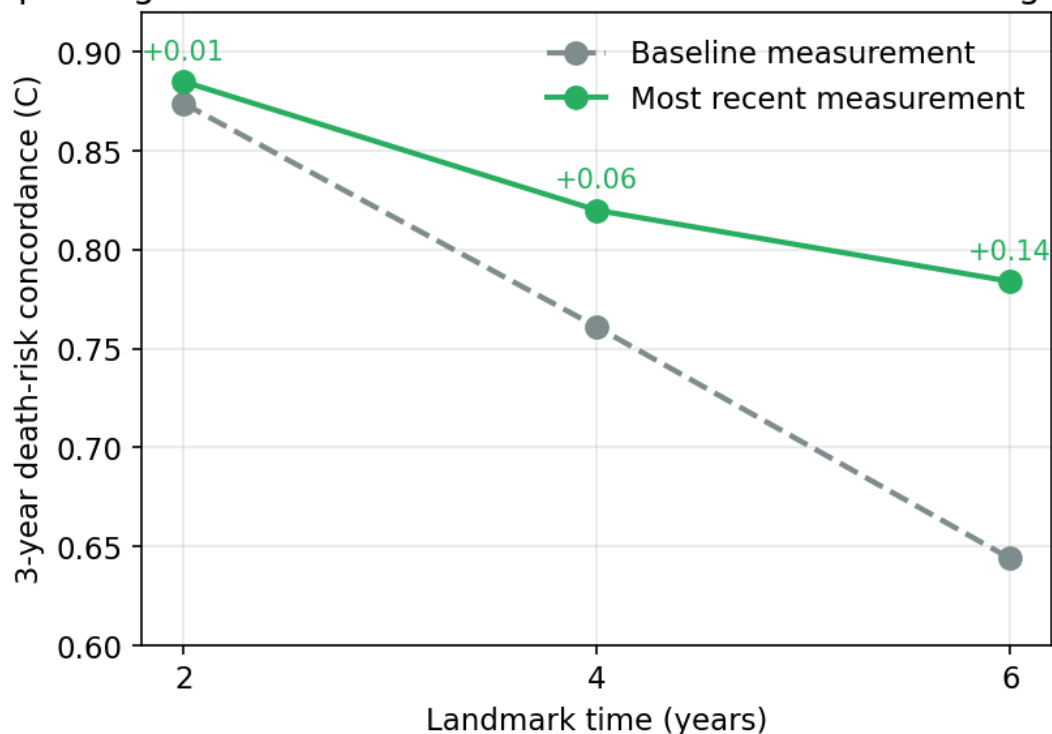


Figure 2: Landmark analysis on the longitudinal cohort. As the baseline measurement ages, its 3-year death-risk discrimination decays; the most recent measurement restores it, with a widening advantage.

and outputs rather than raw discrimination. A competing-risks survival model delivers calibrated, time-resolved survival estimates and cause-specific risk for both death and transplantation, handles censoring and missing data natively, and requires no manual scoring. These are advantages a classifier or a fixed score does not provide, and they come at no cost in discrimination.

A further result points forward. On the longitudinal cohort, the prognostic value of a measurement decayed markedly with time, with baseline death-risk concordance falling to 0.64 by six years, but was largely recovered by using the most recent measurement (0.78), the dynamic advantage widening throughout follow-up. This single-cohort, landmarking-based result is a proof of concept rather than a validated monitoring strategy. Still, it suggests that the prognostic value in these data may lie less in any single snapshot than in *serial monitoring*, motivating the dynamic, competing-risks prediction that updated laboratory trajectories enable [14].

Limitations. The cohort is single-centre, historical and modest ($n = 418$), and the transplant endpoint has only 25 events, roughly eight per covariate in the inferential model, so its estimates, though encouraging, are wide and at risk of optimism. Our primary analysis uses a single snapshot per patient. Although we analysed the longitudinal version of the cohort for the dynamic analyses, we used landmarking rather than full joint longitudinal-survival modelling, which may extract further signal. Our attempt at external validation was limited by data availability. The only freely-available independent cirrhosis cohort (SUPPORT) is an acutely hospitalised population too dissimilar from the chronic outpatient PBC trial for a fair test, and the model did not transport to it. A fair external validation awaits an independent chronic cirrhosis cohort with the full marker

set. The data are also PBC-specific, so generalisation to other cirrhosis aetiologies is unverified. The proportional-hazards assumption held for most covariates but was violated for prothrombin time (Schoenfeld $p = 0.002$); because the analysis is discrimination-focused this has limited impact, but it qualifies the cause-specific hazard ratio for that covariate. Finally, MELD could not be benchmarked because its inputs are absent from the dataset.

Future work. Natural extensions are external validation on contemporary multi-centre cohorts, joint longitudinal-survival modelling that goes beyond the landmarking used here, Fine-Gray absolute-risk modelling alongside the cause-specific analyses presented here, and assessment of clinical utility through decision-curve analysis [28].

6 Conclusion

A widely-used cirrhosis ML benchmark is routinely treated as three-class classification, a framing that answers a weaker question than prognosis requires. Correcting it to competing-risks survival is the consequential step. It turns the poorly-predicted transplant label into an interpretable competing-risks outcome with encouraging, if imprecise, discrimination ($C = 0.82$; 25 events), and it reproduces clinically-established, opposite-direction predictors of death and transplantation that classification cannot express. At the same time, no survival model, including deep networks, beats the simple, validated Mayo clinical score; this is a ceiling set by the available static variables, not an in-sample artefact (refitting the score within folds reproduced it) and consistent in its size-dependence with external datasets. Longitudinally, updating biomarkers recovers discrimination that a stale baseline loses, suggesting value in serial monitoring. On this benchmark, and consistent with broader tabular-ML evidence, correct problem formulation, dynamic prediction and calibrated, cause-specific outputs matter more than model complexity. The lesson transfers beyond cirrhosis: censored, competing-event prediction benchmarks should be posed as competing-risks survival rather than classification, and on small tabular clinical data the returns to getting the formulation right exceed the returns to model sophistication.

Data availability

All datasets are publicly available: the primary cohort from the UCI Machine Learning Repository (“Cirrhosis Patient Survival Prediction”), the longitudinal version (`pbcseq`) from the R `survival` package, and the external datasets via `pycox` and the SUPPORT study. The complete, leakage-free analysis pipeline, covering the competing-risks formulation, the seven-model benchmark, dynamic landmarking, evaluation and figures, is released as a reproducible framework intended for reuse on this and comparable clinical cohorts. It is available at <https://github.com/TheAxeC/cirrhosis-competing-risks>, and will be archived with a citable Zenodo DOI on acceptance.

Ethics

This study used a fully anonymised, publicly available dataset and did not involve new human-subjects research. No ethics approval was required.

CRediT author contributions

Meseret Assefa Kerga: conceptualization, methodology, software, formal analysis, investigation, writing – original draft. **Axel Faes:** conceptualization, methodology, supervision, writing – review & editing.

Declaration of competing interest

The authors declare no competing interests.

Declaration of generative AI and AI-assisted technologies in the writing process

During the preparation of this work the authors used a generative AI assistant (Claude) to support software implementation and to edit and improve the language of the manuscript. After using this tool, the authors reviewed and edited the content as needed and take full responsibility for the content of the publication.

References

- [1] Sumeet K Asrani, Harshad Devarbhavi, John Eaton, and Patrick S Kamath. Burden of liver diseases in the world. *Journal of Hepatology*, 70(1):151–171, 2019.
- [2] E. R. Dickson, P. M. Grambsch, T. R. Fleming, L. D. Fisher, and A. Langworthy. Prognosis in primary biliary cirrhosis: model for decision making. *Hepatology*, 10(1):1–7, 1989. doi: 10.1002/hep.1840100102.
- [3] Patrick S Kamath, Russell H Wiesner, Michael Malinchoc, Walter Kremers, Terry M Therneau, Catherine L Kosberg, Gennaro D’Amico, E Rolland Dickson, and W Ray Kim. A model to predict survival in patients with end-stage liver disease. *Hepatology*, 33(2):464–470, 2001.
- [4] UCI Machine Learning Repository. Cirrhosis patient survival prediction dataset, 2023. <https://archive.ics.uci.edu/dataset/878>.
- [5] Nitesh V Chawla, Kevin W Bowyer, Lawrence O Hall, and W Philip Kegelmeyer. Smote: synthetic minority over-sampling technique. *Journal of Artificial Intelligence Research*, 16: 321–357, 2002.
- [6] E. Thirumagal, B. L. Ananya, V. Nikhitha, and S. Arjun. Liver cirrhosis prediction using machine learning classification techniques. In *Proceedings of the 4th International Conference on Recent Trends in Machine Learning, IoT, Smart Cities and Applications*, volume 873 of *Lecture Notes in Networks and Systems*, pages 45–57. Springer Nature Singapore, 2024. https://doi.org/10.1007/978-981-99-9442-7_5.
- [7] Walter Reade and Ashley Chow. Multi-class prediction of cirrhosis outcomes. Kaggle Playground Series, Season 3, Episode 26, 2023. <https://kaggle.com/competitions/playground-series-s3e26>.

- [8] Hein Putter, M Fiocco, and Ronald B Geskus. Tutorial in biostatistics: competing risks and multi-state models. *Statistics in Medicine*, 26(11):2389–2430, 2007.
- [9] Jason P Fine and Robert J Gray. A proportional hazards model for the subdistribution of a competing risk. *Journal of the American Statistical Association*, 94(446):496–509, 1999.
- [10] Chukwudi Paul Obite, Emmanuella Onyinyechi Chukwudi, Merit Uchechukwu, and Ugochinyere Ihuoma Nwosu. Factor enhanced DeepSurv: a deep learning approach for predicting survival probabilities in cirrhosis data. *Computers in Biology and Medicine*, 189:109963, 2025. doi: 10.1016/j.compbimed.2025.109963.
- [11] Ashwani K Singal, Xiao Fang, Mohamed Kaif, Mohsen Hasanin, Brendan M McGuire, Yong-Fang Kuo, and Russell H Wiesner. Primary biliary cirrhosis has high wait-list mortality among patients listed for liver transplantation. *Transplant International*, 30(5):454–462, 2017. doi: 10.1111/tri.12877.
- [12] Willem J Lammers et al. Development and validation of a scoring system to predict outcomes of patients with primary biliary cirrhosis receiving ursodeoxycholic acid therapy. *Gastroenterology*, 149(7):1804–1812, 2015. doi: 10.1053/j.gastro.2015.07.061.
- [13] Marco Carbone et al. The UK-PBC risk scores: derivation and validation of a scoring system for long-term prediction of end-stage liver disease in primary biliary cholangitis. *Hepatology*, 63(3):930–950, 2016. doi: 10.1002/hep.28017.
- [14] David S Goldberg and Afsoon Zarnegarnia. Longitudinal prediction of survival in patients with cirrhosis: a dynamic artificial intelligence model. *Hepatology Communications*, 7(11):e0270, 2023.
- [15] David R Cox. Regression models and life-tables. *Journal of the Royal Statistical Society: Series B*, 34(2):187–202, 1972.
- [16] Hemant Ishwaran, Udaya B Kogalur, Eugene H Blackstone, and Michael S Lauer. Random survival forests. *The Annals of Applied Statistics*, 2(3):841–860, 2008.
- [17] Jared L Katzman, Uri Shaham, Alexander Cloninger, Jonathan Bates, Tingting Jiang, and Yuval Kluger. DeepSurv: personalized treatment recommender system using a cox proportional hazards deep neural network. *BMC Medical Research Methodology*, 18(1):24, 2018.
- [18] Changhee Lee, William R Zame, Jinsung Yoon, and Mihaela van der Schaar. DeepHit: a deep learning approach to survival analysis with competing risks. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 32, 2018.
- [19] Léo Grinsztajn, Edouard Oyallon, and Gaël Varoquaux. Why do tree-based models still outperform deep learning on typical tabular data? In *Advances in Neural Information Processing Systems 35 (NeurIPS), Datasets and Benchmarks Track*, pages 507–520, 2022.
- [20] Ravid Shwartz-Ziv and Amitai Armon. Tabular data: deep learning is not all you need. *Information Fusion*, 81:84–90, 2022. doi: 10.1016/j.inffus.2021.11.011.
- [21] Peter C Austin, Douglas S Lee, and Jason P Fine. Introduction to the analysis of survival data in the presence of competing risks. *Circulation*, 133(6):601–609, 2016. doi: 10.1161/CIRCULATIONAHA.115.017719.

- 459 [22] Odd O Aalen and Søren Johansen. An empirical transition matrix for non-homogeneous
460 markov chains based on censored observations. *Scandinavian Journal of Statistics*, 5(3):141–
461 150, 1978.
- 462 [23] Frank E Harrell, Kerry L Lee, and Daniel B Mark. Multivariable prognostic models: issues in
463 developing models, evaluating assumptions and adequacy, and measuring and reducing errors.
464 *Statistics in Medicine*, 15(4):361–387, 1996.
- 465 [24] Hajime Uno, Tianxi Cai, Michael J Pencina, Ralph B D’Agostino, and Lee-Jen Wei. On the
466 c-statistics for evaluating overall adequacy of risk prediction procedures with censored survival
467 data. *Statistics in Medicine*, 30(10):1105–1117, 2011.
- 468 [25] Erika Graf, Claudia Schmoor, Willi Sauerbrei, and Martin Schumacher. Assessment and com-
469 parison of prognostic classification schemes for survival data. *Statistics in Medicine*, 18(17-18):
470 2529–2545, 1999.
- 471 [26] Gary S Collins, Johannes B Reitsma, Douglas G Altman, and Karel GM Moons. Transparent
472 reporting of a multivariable prediction model for individual prognosis or diagnosis (tripod).
473 *BMJ*, 350:g7594, 2015.
- 474 [27] Sebastian Pölsterl. scikit-survival: a library for time-to-event analysis built on top of scikit-
475 learn. *Journal of Machine Learning Research*, 21(212):1–6, 2020.
- 476 [28] Andrew J Vickers and Elena B Elkin. Decision curve analysis: a novel method for evaluating
477 prediction models. *Medical Decision Making*, 26(6):565–574, 2006.