Beyond Binary Classification: A Comprehensive Survival Analysis and Calibration Assessment of the Heart Failure Clinical Records Datase

Ipek Balikci Cicek [1*], Zeynep Kucukakcali [2]
[1-2] Inonu University Faculty of Medicine, Department of Biostatistics and Medical Informatics, 44280, Malatya, Turkey

* Corresponding Author: **Ipek Balikci Cicek**

Article Info

ISSN (online): 2582-8940
Volume: 07
Issue: 03
Received: 11-06-2026
Accepted: 10-07-2026
Published: 09-08-2026
Page No: 183-190

Abstract
Background: The Heart Failure Clinical Records dataset (n = 299) is among the most extensively reused benchmark datasets in clinical machine learning, yet the overwhelming majority of published analyses treat patient mortality as a static binary classification target and rarely test the statistical assumptions underlying their models.
Methods: We instead modeled mortality as a time-to-event outcome using Kaplan-Meier estimation, multivariable Cox proportional hazards (PH) regression, and Random Survival Forest (RSF). We explicitly tested the proportional hazards assumption, evaluated discrimination using time-dependent AUC, assessed calibration via the integrated Brier score, and examined model performance across sex, age, and diabetes subgroups.
Results: Age, ejection fraction, and serum creatinine emerged as the dominant predictors of mortality across both Cox PH and RSF models. Ejection fraction violated the proportional hazards assumption (p = 0.013); after stratifying by ejection fraction category, RSF outperformed the corrected Cox model on an identical hold-out set (C-index 0.775 vs. 0.716). RSF achieved a mean time-dependent AUC of 0.820, but calibration deteriorated sharply in late follow-up (Brier score rising to 0.53 near day 270), despite high discrimination in that same window. No sex-based survival difference was observed (log-rank p = 0.950), and subgroup C-index values were broadly comparable across sex, age, and diabetes status, though small subgroup sizes limit precision.
Conclusions: Framing this dataset as a survival problem reveals assumption violations, calibration failures, and confounding effects that are invisible to standard binary classification pipelines, even though the same core clinical predictors are recovered. We recommend that future studies on this dataset report proportional hazards diagnostics and calibration alongside discrimination metrics.

DOI: https://doi.org/10.54660/IJMBHR.2026.7.3.183-190

Keywords: heart failure, survival analysis, Cox proportional hazards, random survival forest, calibration, fairness, machine learning

1. Introduction

Cardiovascular diseases (CVDs) remain the leading cause of death worldwide, accounting for an estimated 19.8 million deaths in 2022, approximately 32% of all global deaths[1]. Heart failure (HF), a major contributor to this burden, is a complex clinical syndrome in which the heart cannot pump sufficient blood to meet the body's metabolic demands, and predicting patient survival from routinely collected clinical variables remains a central goal of clinical decision support research. The Heart Failure Clinical Records dataset, originally collected from 299 patients treated at the Institute of Cardiology and Allied Hospital in Faisalabad, Pakistan, in 2015, was first analyzed using Cox regression and a nomogram [2], and was later popularized as a machine learning

benchmark, with serum creatinine and ejection fraction alone reported as sufficient to predict survival with reasonable accuracy [3].

Since then, the dataset has been reanalyzed in a very large number of studies, the majority of which frame the problem as static binary classification: the time variable, representing each patient's follow-up duration, is either discarded or used as an ordinary numerical feature, and model performance is reported using accuracy, F1-score, or a single receiver operating characteristic AUC value. Representative examples include SMOTE-based class balancing applied to improve classification metrics [4, 5], a transfer-learning-based feature engineering approach proposed for the same classification task [6], and post hoc explainability techniques applied to both a classification model and, more unusually, a survival gradient boosting model, with the latter reporting a concordance index of 0.714, comparable to the unstratified Cox result obtained in the present study [7]. None of these studies, however, formally tested the proportional hazards assumption or jointly evaluated discrimination and calibration over the follow-up period.

A smaller body of work has revisited the dataset using genuine survival-analytic methods. Most recently, one study examined how follow-up adequacy, quantified via the Person-Time Follow-up Rate, affects the comparative performance of Cox proportional hazards regression and Random Survival Forest (RSF) on this same cohort [8]. However, systematic diagnostic testing of the proportional hazards assumption, joint evaluation of discrimination and calibration over time, and subgroup-level (fairness) performance auditing remain largely absent from the literature on this dataset.

The present study addresses these gaps. We (i) model mortality explicitly as a time-to-event outcome using Kaplan-Meier estimation, Cox PH regression, and RSF; (ii) formally test the proportional hazards assumption and quantify the consequence of its violation; (iii) evaluate model performance jointly through time-dependent discrimination (AUC) and calibration (Brier score); and (iv) assess whether predictive performance is consistent across clinically relevant patient subgroups (sex, age, diabetes status).

2. Materials and Methods

2.1. Dataset

We used the publicly available Heart Failure Clinical Records dataset [2, 3], comprising 299 patients aged 40 years or older with left ventricular systolic dysfunction (NYHA class III–IV), followed for a mean of 130.3 days. The dataset contains 11 clinical/demographic covariates (age, anaemia, creatinine phosphokinase, diabetes, ejection fraction, high blood pressure, platelets, serum creatinine, serum sodium, sex, and smoking status) together with the follow-up duration (time, in days) and a binary event indicator (DEATH_EVENT). Of the 299 patients, 96 (32.1%) died during follow-up and 203 (67.9%) were right-censored at their last observation. As detailed in Section 2.2, ejection fraction was subsequently categorized into three groups (≤30%, 30–45%, and >45%) based on the dataset's observed distribution [9].

2.2. Survival Analysis Framework

Unlike the majority of prior work on this dataset, we treated time and DEATH_EVENT jointly as a censored time-to-event outcome rather than discarding or dichotomizing the follow-up duration. Kaplan-Meier estimation [10] was used to characterize overall and subgroup-specific survival curves, with between-group differences assessed via the log-rank test (and the multivariate log-rank test for the three-level ejection-fraction stratum).

A multivariable Cox proportional hazards model [11] was fitted using all 11 covariates. The proportional hazards assumption was formally tested for each covariate using both Kaplan-Meier-based and rank-based statistics derived from scaled Schoenfeld residuals[12], implemented via lifelines' check_assumptions routine. Where the assumption was violated, the offending covariate was re-expressed as a categorical stratification variable and the model was refit.

A Random Survival Forest [13], building on the random forest algorithm[14] (500 trees, minimum split size 10, minimum leaf size 5, sqrt feature subsampling), was trained on a 75/25 train–test split (stratified by event status, random_state = 42) using the same 11 covariates. Model discrimination was quantified using Harrell's concordance index (C-index)[15] on the held-out test set, computed identically for both the Cox and RSF models to ensure a fair comparison. Permutation importance (15 repeats) was used to rank covariate contributions in the RSF model.

2.3. Time-dependent Discrimination and Calibration

Beyond a single aggregate C-index, we evaluated model performance across the follow-up period using cumulative/dynamic time-dependent AUC at fixed time points (30–270 days, 20-day increments), and assessed calibration using the time-dependent Brier score and its integrated summary (Integrated Brier Score, IBS) [16], computed on RSF-predicted survival probabilities at the same time points. Calibration was evaluated as a distinct property from discrimination, following the recommendation that predictive models be assessed on both dimensions rather than discrimination alone [17].

2.4. Subgroup (Fairness) Evaluation

To assess whether model performance was consistent across clinically relevant patient groups, we computed the C-index separately within the held-out test set for sex (female vs. male), age (below vs. at/above the sample median), and diabetes status (absent vs. present).

2.5. Software

All analyses were performed in Python 3 using lifelines [18] for Kaplan-Meier estimation, Cox PH regression, and proportional hazards diagnostics; scikit-survival [18] for Random Survival Forest, the concordance index, time-dependent AUC, and Brier score; and scikit-learn [19] for train/test splitting and permutation importance. Data handling, numerical computation, and visualization relied on pandas [20], NumPy [21], matplotlib [22], and seaborn [23].

3. Results

3.1. Cohort Characteristics

The cohort comprised 299 patients, of whom 96 (32.1%) experienced the primary endpoint (death) during a mean follow-up of 130.3 days, while 203 (67.9%) were right-censored (Figure 1). Deaths were concentrated in the earlier part of the follow-up period, consistent with the well-documented high early-mortality risk in heart failure. Univariate comparisons showed clearly separated distributions for ejection fraction (lower among non-

survivors) and serum creatinine (higher among non-survivors), with weaker or negligible separation for platelets and creatinine phosphokinase (Figure 2).

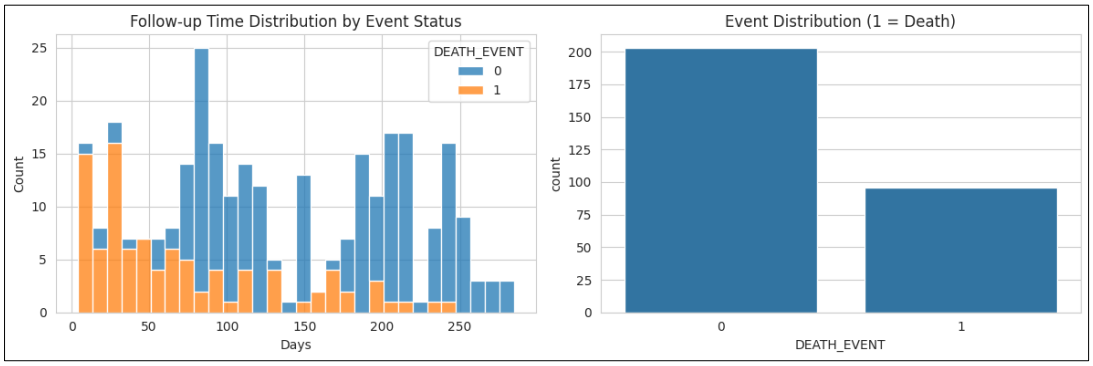

Fig 1: Follow-up time distribution by event status (left) and overall event distribution (right).

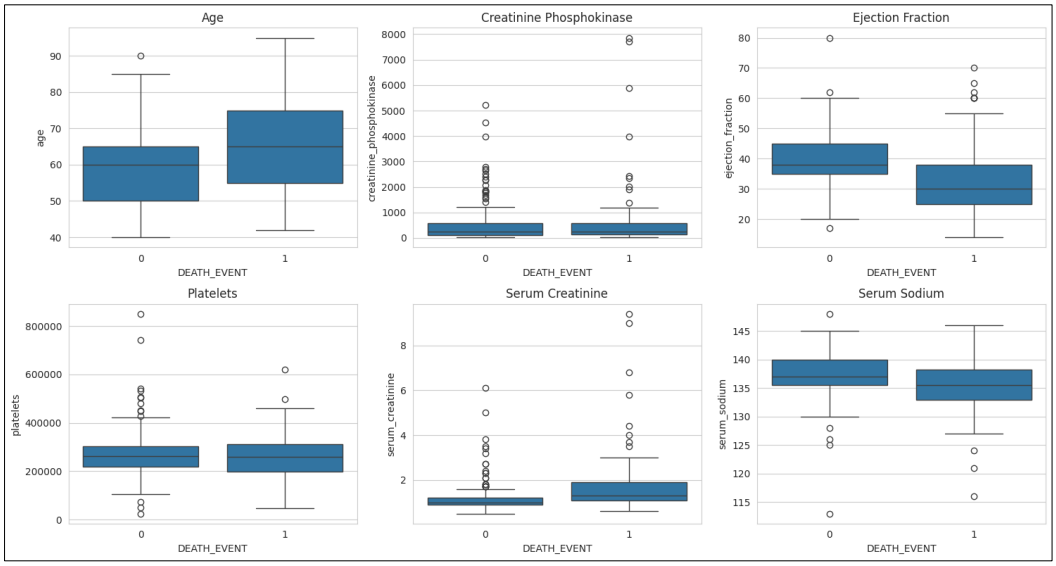

Fig 2: Distribution of six continuous clinical variables stratified by death event (0 = survived, 1 = died).

3.2. Kaplan-Meier Survival Analysis

The overall Kaplan-Meier curve showed a steep early decline in survival probability, falling from 100% to approximately 83% within the first 50 days, followed by a more gradual decline and a plateau near 58% after day 250 (Figure 3). Survival did not differ significantly by sex (log-rank p = 0.950; Figure 4), but was significantly lower in patients with hypertension (log-rank p = 0.036; Figure 5) and in older patients (age ≥ median; log-rank p = 0.003).

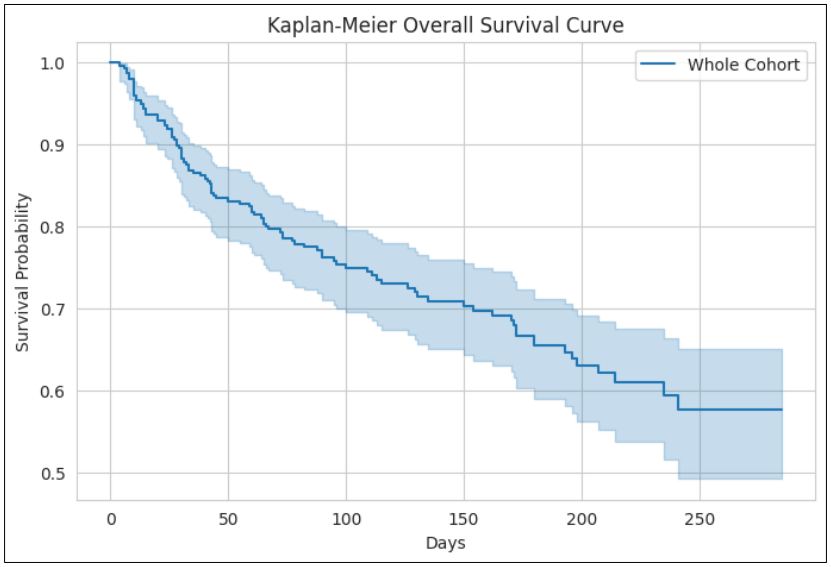

Fig 3: Kaplan-Meier overall survival curve (shaded region: 95% confidence interval).

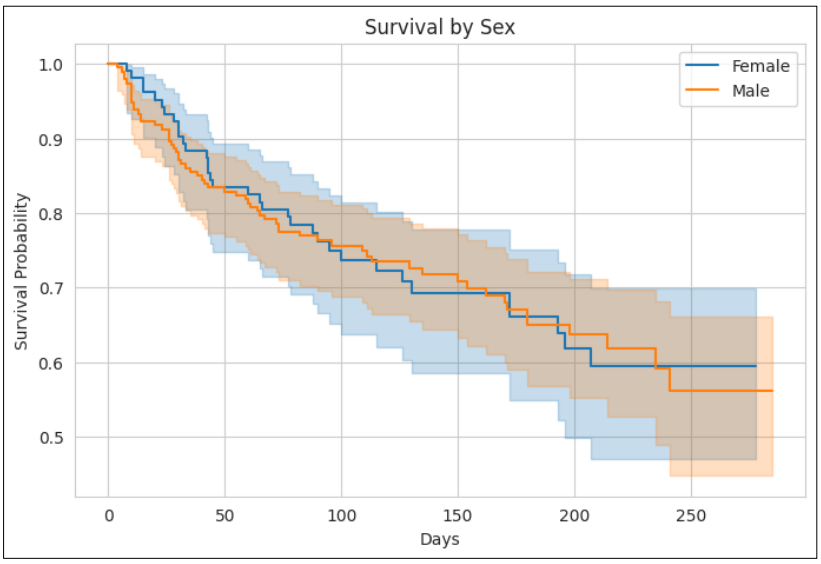

Fig 4: Kaplan-Meier survival curves by sex (log-rank p = 0.950).

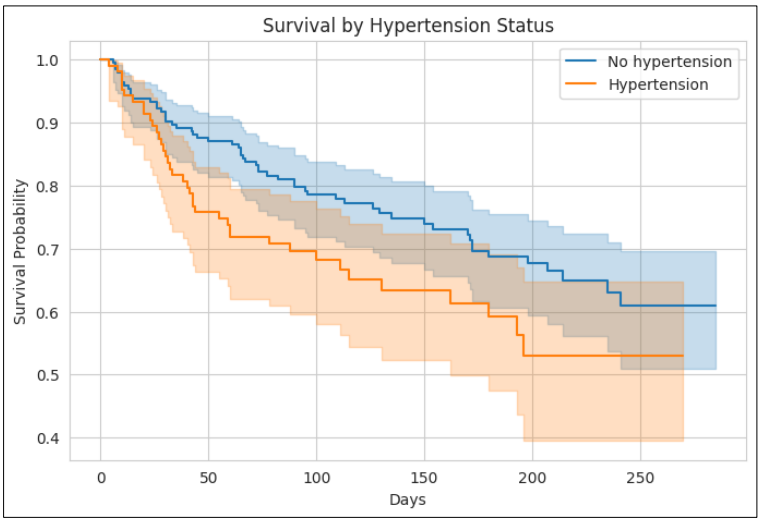

Fig 5: Kaplan-Meier survival curves by hypertension status (log-rank p = 0.036).

3.3. Cox Proportional Hazards Model and Assumption Testing

In the initial (unstratified) multivariable Cox model fitted on the full dataset, age (HR = 1.05, p < 0.005), ejection fraction (HR = 0.95, p < 0.005), serum creatinine (HR = 1.38, p < 0.005), anaemia (HR = 1.58, p = 0.03), and high blood pressure (HR = 1.61, p = 0.03) were independently associated with mortality (overall concordance = 0.74). Formal proportional hazards testing revealed that ejection fraction violated the assumption (p = 0.013 for both the Kaplan-Meier-based and rank-based tests), while serum creatinine approached the significance threshold (p = 0.07 for the Kaplan-Meier-based test and p = 0.06 for the rank-based test. After stratifying the model by ejection fraction category (≤ 30%, 30–45%, > 45%) to resolve this violation, overall concordance decreased slightly to 0.71 and the covariate significance profile shifted: anaemia was no longer significant (p = 0.13), while high blood pressure (HR = 1.73, p = 0.01) and creatinine phosphokinase (p = 0.01) strengthened. Age and serum creatinine remained robust, significant predictors in both model specifications (Table 1, Figure 6).

Table 1: Stratified Cox proportional hazards model (stratified by ejection fraction category). HR = hazard ratio; CI = confidence interval.

Variable	HR (exp(coef))	95% CI	p-value
Age	1.06	1.04–1.08	<0.005
Anaemia	1.39	0.90–2.14	0.13
Creatinine phosphokinase	1.00	1.00–1.00	0.01
Diabetes	1.16	0.75–1.80	0.50
High blood pressure	1.73	1.12–2.66	0.01
Platelets	1.00	1.00–1.00	0.57
Serum creatinine	1.29	1.13–1.48	<0.005
Serum sodium	0.96	0.91–1.00	0.05
Sex	0.87	0.53–1.43	0.59
Smoking	1.04	0.63–1.71	0.88

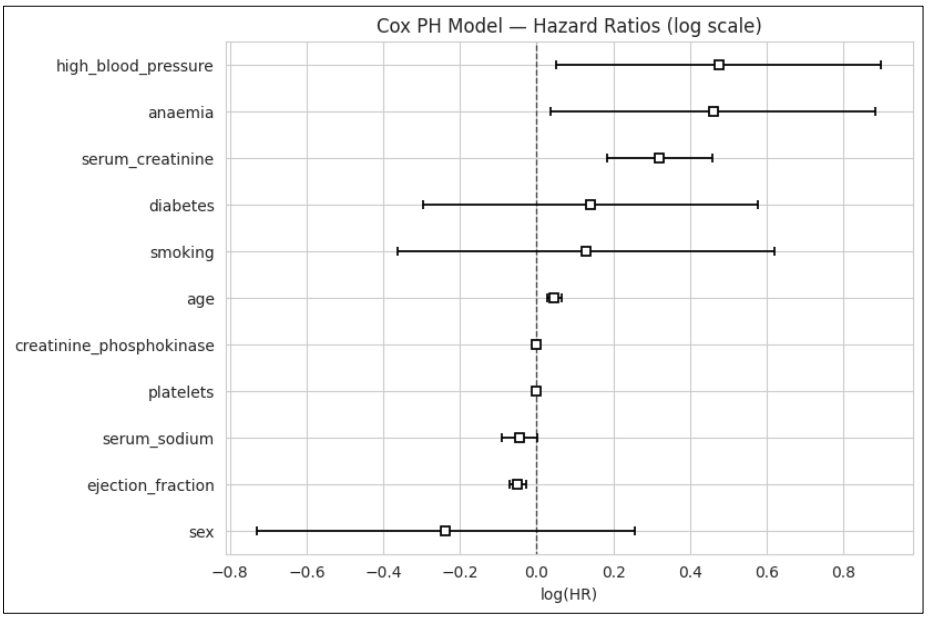

Fig 6: Forest plot of log hazard ratios (unstratified Cox model) with 95% confidence intervals.

3.4. Cox PH vs. Random Survival Forest

On an identical hold-out test set, RSF achieved higher discrimination than the (ejection-fraction-stratified) Cox model (C-index 0.775 vs. 0.716; Table 2), consistent with RSF's ability to capture the non-linear, time-varying effect of ejection fraction that violated the Cox model's core assumption. Permutation importance identified age, ejection fraction, and serum creatinine as the three dominant predictors in the RSF model, mirroring the Cox model's findings and results reported previously for this dataset [24] (Figure 7). Notably, high blood pressure, a significant predictor in the Cox model, ranked low in RSF importance, a discrepancy plausibly attributable to correlation with stronger predictors that reduces its marginal contribution within a tree-based ensemble.

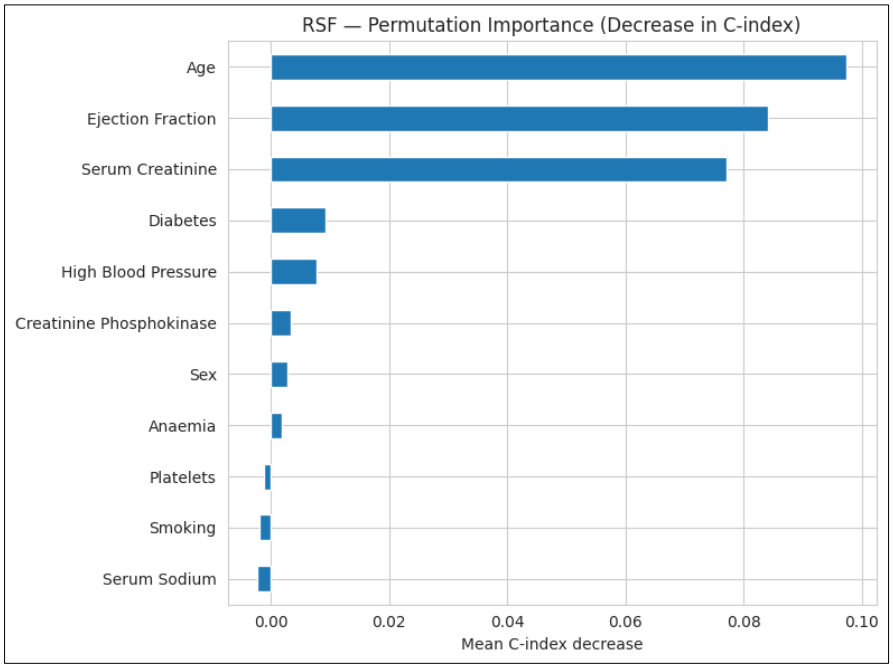

Fig 7: Random Survival Forest permutation importance (mean decrease in C-index over 15 repeats).

3.5. Time-dependent Discrimination and Calibration

RSF achieved a mean time-dependent AUC of 0.820 across the follow-up period, with discrimination weakest around day 90 (AUC = 0.735) and strongest beyond day 250 (AUC > 0.92; Figure 8). However, the integrated Brier score (IBS = 0.185) indicated only moderate overall calibration, and calibration deteriorated markedly at later time points: the Brier score rose from approximately 0.09 at day 30 to 0.53 at day 270 (Figure 9), the same interval in which time-dependent AUC was highest. This divergence indicates that strong rank-based discrimination does not guarantee reliable probability calibration, particularly in regions with a shrinking at-risk population.

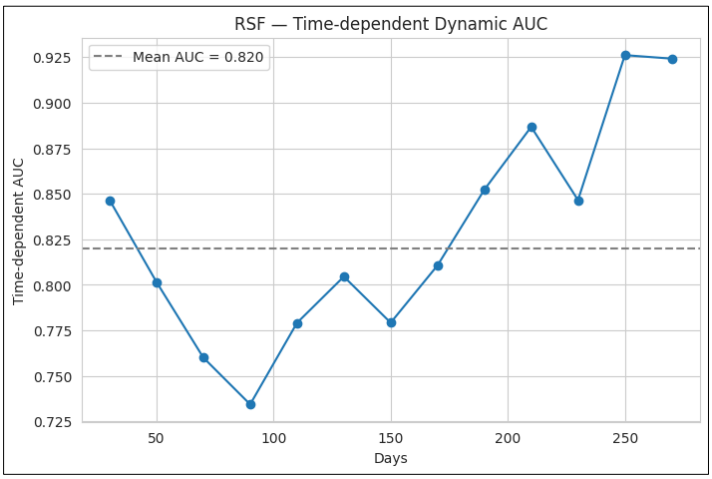

Fig 8: Random Survival Forest time-dependent (cumulative/dynamic) AUC.

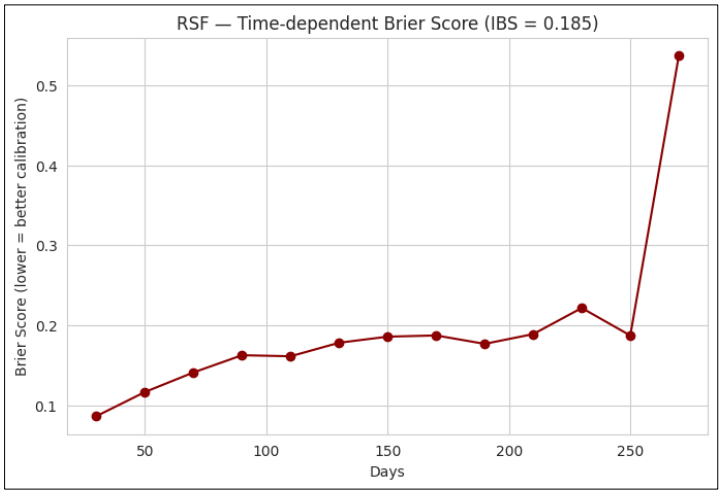

Fig 9: Random Survival Forest time-dependent Brier score (IBS = 0.185; lower values indicate better calibration).

Table 2: Summary of model discrimination and calibration metrics on the held-out test set.

Model / Metric	Value
Cox PH: C-index (stratified by EF category)	0.716 (test set)
Random Survival Forest: C-index	0.775 (test set)
RSF: Mean time-dependent AUC	0.820
RSF: Integrated Brier Score (IBS)	0.185

3.6 Subgroup (Fairness) Performance

RSF performance was broadly consistent across sex, age, and diabetes subgroups within the test set (Table 3), with no subgroup showing a marked performance collapse. Combined with the absence of a significant sex-based survival difference (log-rank p = 0.950) and the low permutation importance of sex in the RSF model, these results suggest no strong evidence of systematic sex-based bias in this cohort. Given the small subgroup sample sizes (ranging from n = 24 in the diabetes subgroup to n = 51 in the no-diabetes subgroup; see Table 3), however, these point estimates should be interpreted cautiously, as their precision is limited and confidence intervals were not computed.

Table 3: Random Survival Forest concordance index by patient subgroup, held-out test set.

Subgroup	n (test set)	C-index
Female	25	0.801
Male	50	0.765
Younger (< median age)	37	0.727
Older (≥ median age)	38	0.688
No diabetes	51	0.790
Diabetes	24	0.761

4. Discussion

4.1. Principal Findings

This study reframed a dataset that has almost exclusively been analyzed as a binary classification problem as a time-to-event survival problem. Despite this different framing, both Cox PH and RSF converged on the same three dominant predictors (age, ejection fraction, and serum creatinine), corroborating earlier findings[2, 3] through an independent methodological lens. This convergence across three separate analytical approaches (Kaplan-Meier stratification, Cox

regression, and RSF) strengthens confidence in these variables as genuine, robust prognostic markers in this cohort, rather than artifacts of a particular modeling choice. The additional risk factors identified (anaemia, hypertension, and impaired renal function as reflected by serum creatinine) are each independently well established in the broader heart failure literature: anaemia has been associated with increased mortality in a meta-analysis of over 150,000 patients [25], hypertension is a recognized precursor in the progression to heart failure [26], and impaired renal function, commonly assessed via serum creatinine, is a well-documented determinant of prognosis in heart failure through the cardiorenal syndrome [27]. The consistency between our data-driven findings and this established clinical literature lends external plausibility to the models beyond their internal statistical performance.

4.2. The Cost of Ignoring Proportional Hazards Violations

A central finding of this study is that ejection fraction violates the proportional hazards assumption, a diagnostic step that is essentially absent from the wider literature on this dataset. Correcting for this violation via stratification altered the model's covariate significance profile, most notably rendering anaemia non-significant after adjustment. This suggests that anaemia's apparent prognostic value in unadjusted Cox analyses may be, at least in part, confounded by the time-varying effect of ejection fraction. More broadly, it demonstrates that an acceptable aggregate concordance index does not guarantee that a Cox model's individual covariate effects are being correctly estimated.

4.3. Discrimination Is Not Calibration

A second key contribution is the demonstration that time-dependent AUC and the Brier score can diverge sharply within the same model. RSF's discrimination improved markedly beyond day 250 (AUC > 0.92), yet its calibration deteriorated in the same window (Brier score rising to 0.53), most plausibly reflecting the shrinking number of patients remaining at risk late in follow-up. This pattern is a concrete illustration of a broader, well-documented concern: calibration has been described as the “Achilles heel” of predictive analytics[28], since models with strong discriminative ability are routinely deployed and reported without any accompanying calibration assessment, even though poor calibration can directly translate into misleading risk estimates for individual patients. Because the vast majority of prior studies on this dataset report only accuracy, F1-score, or a single AUC value, this discrimination-calibration divergence would go entirely undetected under standard reporting practices. For models intended to inform clinical risk communication (where a calibrated probability of survival is arguably more actionable than a rank-ordering of risk), this finding argues for routine joint reporting of discrimination and calibration metrics, evaluated across the follow-up period rather than at a single time point.

4.4. Fairness Considerations

Subgroup analysis found no evidence of a systematic sex-based disparity in either survival (log-rank p = 0.950) or model discrimination, and sex ranked lowest among covariates in RSF permutation importance. The modest differences observed in subgroup C-index (e.g., female 0.801 vs. male 0.765) are more plausibly attributable to sampling variance in small subgroups (n = 24–51 in the test set) than to genuine model bias, though this cannot be confirmed without bootstrapped confidence intervals, an important direction for future fairness auditing. Such auditing is not a purely theoretical concern: a widely deployed healthcare risk-prediction algorithm has been shown to exhibit substantial racial bias that was invisible under its original, aggregate evaluation and only became apparent through explicit subgroup analysis [29]. This precedent underscores the importance of routine subgroup evaluation for any clinical prediction model, particularly as models trained on small, single-center datasets such as this one are increasingly proposed for broader deployment.

4.5. Limitations

This study inherits the well-documented limitations of the underlying dataset: a single-center cohort of 299 patients collected in 2015 at one hospital in Faisalabad, Pakistan, with a documented Person-Time Follow-up Rate well below recommended thresholds[8] [24], which has itself been shown to affect the relative performance of Cox and RSF models. A sample of this size falls well short of the minimum recommended for developing a stable multivariable prediction model with 11 candidate predictors[30], which likely contributes to the wide confidence intervals observed for several covariates and to the instability of subgroup estimates. The relatively small sample size (and the even smaller size of the test-set subgroups analyzed for fairness) limits the precision and generalizability of the time-dependent and subgroup-level estimates reported here. External validation on an independent, multi-center heart failure cohort is required before any of these findings could inform clinical practice, and future reporting on this dataset would benefit from adherence to established reporting guidelines for prediction model studies [31]. Future work could also extend this comparison to deep-learning-based survival models such as DeepSurv [32], which were outside the scope of the present study but may further improve on the non-linear effects captured here by RSF.

5. Conclusion

Beyond confirming ejection fraction, serum creatinine, and age as central predictors of mortality in this frequently reused benchmark dataset, this study shows that standard binary classification pipelines can obscure assumption violations, calibration failures, and covariate confounding that only become visible under an explicit survival-analytic framework. We recommend that future work on the Heart Failure Clinical Records dataset, and on similarly small, single-center clinical datasets more broadly, report proportional hazards diagnostics, time-dependent calibration, and subgroup performance alongside conventional discrimination metrics.

References

1. World Health Organization. Cardiovascular diseases (CVDs) [Internet]. Geneva: World Health Organization; 2021 [cited 2024 Jan [1]. Available from: https://www.who.int/news-room/fact-sheets/detail/cardiovascular-diseases-(cvds)
2. Ahmad T, Munir A, Bhatti SH, Aftab M, Raza MA. Survival analysis of heart failure patients: a case study. PLoS One. 2017;12:e0181001.
3. Chicco D, Jurman G. Machine learning can predict

survival of patients with heart failure from serum creatinine and ejection fraction alone. BMC Med Inform Decis Mak. 2020;20:16.
4. Chawla NV, Bowyer KW, Hall LO, Kegelmeyer WP. SMOTE: synthetic minority over-sampling technique. J Artif Intell Res. 2002;16:321-357.
5. Rupapara V. Improving the prediction of heart failure patients' survival using SMOTE and effective data mining techniques. IEEE Access. [Year unknown].
6. Qadri AM, Hashmi MSA, Raza A, Zaidi SAJ, Rehman AU. Heart failure survival prediction using novel transfer learning based probabilistic features. PeerJ Comput Sci. 2024;10:e1894.
7. Moreno-Sánchez PA. Improvement of a prediction model for heart failure survival through explainable artificial intelligence. Front Cardiovasc Med. 2023;10:1219586.
8. Özgür EG, Bekiroğlu GN. Why follow-up matters in survival analysis: comparing Cox proportional hazard regression and random survival forest for predicting heart failure outcomes. BMC Cardiovasc Disord. 2025;25:673.
9. Bozkurt B, Coats AJ, Tsutsui H, Abdelhamid M, Adamopoulos S, Albert N, *et al*. Universal definition and classification of heart failure: a report of the Heart Failure Society of America, Heart Failure Association of the European Society of Cardiology, Japanese Heart Failure Society and Writing Committee of the Universal Definition of Heart Failure. J Card Fail. 2021;27:387-413.
10. Kaplan EL, Meier P. Nonparametric estimation from incomplete observations. J Am Stat Assoc. 1958;53:457-481.
11. Cox DR. Regression models and life-tables. J R Stat Soc Series B Methodol. 1972;34:187-202.
12. Grambsch PM, Therneau TM. Proportional hazards tests and diagnostics based on weighted residuals. Biometrika. 1994;81:515-526.
13. Ishwaran H, Kogalur UB, Blackstone EH, Lauer MS. Random survival forests. Ann Appl Stat. 2008;2:841-860.
14. Breiman L. Random forests. Mach Learn. 2001;45:5-32.
15. Harrell FE Jr, Lee KL, Mark DB. Multivariable prognostic models: issues in developing models, evaluating assumptions and adequacy, and measuring and reducing errors. Stat Med. 1996;15:361-387.
16. Graf E, Schmoor C, Sauerbrei W, Schumacher M. Assessment and comparison of prognostic classification schemes for survival data. Stat Med. 1999;18:2529-2545.
17. Van Calster B, McLernon DJ, Van Smeden M, Wynants L, Steyerberg EW. Calibration: the Achilles heel of predictive analytics. BMC Med. 2019;17:230.
18. Davidson-Pilon C. lifelines: survival analysis in Python. J Open Source Softw. 2019;4:1317.
19. Pedregosa F, Varoquaux G, Gramfort A, Michel V, Thirion B, Grisel O, *et al*. Scikit-learn: machine learning in Python. J Mach Learn Res. 2011;12:2825-2830.
20. McKinney W. Data structures for statistical computing in Python. In: Proceedings of the 9th Python in Science Conference; 2010. p. 51-56.
21. Harris CR, Millman KJ, van der Walt SJ, Gommers R, Virtanen P, Cournapeau D, *et al*. Array programming with NumPy. Nature. 2020;585:357-362.
22. Hunter JD. Matplotlib: a 2D graphics environment. Comput Sci Eng. 2007;9:90-95.
23. Waskom ML. Seaborn: statistical data visualization. J Open Source Softw. 2021;6:3021.
24. Chicco D, Jurman G. Machine learning can predict survival of patients with heart failure from serum creatinine and ejection fraction alone. BMC Med Inform Decis Mak. 2020;20:16.
25. Groenveld HF, Januzzi JL, Damman K, van Wijngaarden J, Hillege HL, van Veldhuisen DJ, *et al*. Anemia and mortality in heart failure patients: a systematic review and meta-analysis. J Am Coll Cardiol. 2008;52:818-827.
26. Levy D, Larson MG, Vasan RS, Kannel WB, Ho KKL. The progression from hypertension to congestive heart failure. JAMA. 1996;275:1557-1562.
27. Damman K, Valente MA, Voors AA, O'Connor CM, van Veldhuisen DJ, Hillege HL. Renal impairment, worsening renal function, and outcome in patients with heart failure: an updated meta-analysis. Eur Heart J. 2014;35:455-469.
28. Van Calster B, McLernon DJ, Van Smeden M, Wynants L, Steyerberg EW, *et al*. Calibration: the Achilles heel of predictive analytics. BMC Med. 2019;17:230.
29. Obermeyer Z, Powers B, Vogeli C, Mullainathan S. Dissecting racial bias in an algorithm used to manage the health of populations. Science. 2019;366:447-453.
30. Riley RD, Ensor J, Snell KIE, Harrell FE Jr, Martin GP, Reitsma JB, *et al*. Calculating the sample size required for developing a clinical prediction model. BMJ. 2020;368:m441.
31. Collins GS, Reitsma JB, Altman DG, Moons KGM. Transparent reporting of a multivariable prediction model for individual prognosis or diagnosis (TRIPOD): the TRIPOD statement. Ann Intern Med. 2015;162:55-63.
32. Katzman JL, Shaham U, Cloninger A, Bates J, Jiang T, Kluger Y. DeepSurv: personalized treatment recommender system using a Cox proportional hazards deep neural network. BMC Med Res Methodol. 2018;18:24.

How to Cite This Article
Balikci Cicek I, Kucukakcali Z. Beyond binary classification: a comprehensive survival analysis and calibration assessment of the Heart Failure Clinical Records dataset. Int J Med All Body Health Res. 2026;7(3):183-190.
doi:10.54660/IJMBHR.2026.7.3.183-190.