Research Square

Preprints are preliminary reports that have not undergone peer review.
They should not be considered conclusive, used to inform clinical practice,
or referenced by the media as validated information.

Survival Analysis Outperforms Regression for Field Prognostics of Valve-Regulated Lead-Acid Batteries: A Reproducible Benchmark on 1,027 Solar Off-Grid Systems

Dikshant Dikshant

dikshant1103@gmail.com

Motilal Nehru National Institute of Technology https://orcid.org/0009-0006-9434-610X

Research Article

Keywords: lead-acid battery, VRLA, remaining useful life, survival analysis, right censoring, field data, solar home systems, predictive maintenance, benchmark

Posted Date: August 7th, 2026

DOI: https://doi.org/10.21203/rs.3.rs-10606266/v1

License: This work is licensed under a Creative Commons Attribution 4.0 International License.
Read Full License

Additional Declarations: The authors declare no competing interests.

Abstract

Valve-regulated lead-acid (VRLA) batteries remain the dominant stationary storage technology across much of the developing world — telecom towers, inverter backup and solar home systems — yet prognostics research is concentrated almost entirely on lithium-ion. A 2023 systematic mapping study of lead-acid prognostics reported that no public dataset could be found, and errors reported on private cohorts of 3–70 cells (0.0005–7.6%) are mutually incomparable. Using the Oxford/BBOXX release of 1,027 field-deployed 12 V/20 Ah VRLA batteries on solar home systems in sub-Saharan Africa (approximately 620 million rows; 400–760 days per unit; 491 observed failures; 536 right-censored units), this work establishes, to our knowledge, the first openly specified and reproducible benchmark for lead-acid field prognostics. The central methodological finding is that, with 52% right censoring, the task is substantially better posed as survival analysis than as regression: a bidirectional Mamba regressor for remaining useful life attains 83.3 d mean absolute error — worse than the one-line heuristic mean lifetime minus age (63.8 d) and worse than a ridge regression (57.2 d) — whereas landmark survival analysis over 6,003 battery-landmark observations with 97 level, trend and volatility features attains AUC 0.848 (95% CI 0.819–0.877) at a four-week horizon and 0.832 (0.808–0.855) at twelve weeks on the at-risk population, with Uno's IPCW concordance of 0.737–0.756 across five estimators (best 0.756, Coxnet). Replicating the protocol and metric of Aitio and Howey (Joule, 2021) on the same dataset, balanced accuracy reaches 84.1% eight weeks before failure (published: 73%) and 89.7% at the point of failure (published: 82%), using only daily aggregates and no state-of-health estimate. A five-model ablation indicates that, within the at-risk population, battery age alone is uninformative (balanced accuracy 52.5–57.5%) while telemetry alone recovers the full model (77–89%). Three negative results are reported in full, and code, features and deterministic splits are released.

1. Introduction

1.1 Motivation

Lead-acid is the oldest rechargeable chemistry in commercial use and remains, by installed energy capacity, the dominant stationary storage technology in South Asia and sub-Saharan Africa. Telecom tower backup, residential inverter systems, uninterruptible power supplies and solar home systems are overwhelmingly served by flooded tubular and valve-regulated (VRLA) lead-acid banks, chosen for capital cost, established recycling infrastructure and tolerance of abuse. The economics of these fleets is dominated by replacement: a battery bank is a recurring operating expense, and knowing which unit will fail next determines maintenance routing, warranty exposure and, for pay-as-you-go solar operators, customer default risk.

Despite this, prognostics research has followed the lithium-ion market. The consequences are concrete, and they are documented by the field's own review literature.

1.2 The gap this paper addresses

A 2023 systematic mapping study of machine-learning methods for lead-acid state-of-health (SoH) and remaining-useful-life (RUL) estimation [3] surveyed 17 papers published between 2013 and 2023 and reached four conclusions.

(1) *No shared benchmark.* The review states that "a publicly available dataset was not found during the present research"; every surveyed study reports on private data. (2) *Cohorts too small to generalise.* Reported cohorts range from three cells to seventy retired batteries; only one study approaches fleet scale. (3) *Incomparable metrics.* Reported errors span 0.0005% to 7.6%, but the majority address SoH estimation — interpolating present capacity, for which a laboratory label exists — rather than RUL prediction, which extrapolates to an unobserved future event. The two tasks are distinct and their error figures cannot be placed side by side. (4) *Narrow algorithm coverage.* Multilayer perceptrons appear in 13 of the 17 studies; the review lists support vector machines, tree ensembles and modern deep architectures as under-explored and notes an absence of comparisons on identical data.

The review's conclusion that no public dataset exists is no longer accurate: the Oxford/BBOXX field release accompanying Aitio and Howey [1, 2] was published in December 2021 and contains 1,027 instrumented batteries. The field appears not to have adopted it. That dataset is the basis of this paper.

1.3 Why field data changes the problem

Laboratory lead-acid studies rely on two conditions that field deployment removes, and both are consequential.

Capacity ground truth. Laboratory protocols interleave reference capacity tests, providing a direct state-of-health label. The Oxford/BBOXX release contains only current, voltage and temperature at nominal 60-second resolution, plus a failed/alive flag. State of health is not a label that can be read; it must be inferred, and § 7.3 shows that the natural inference procedure is biased.

Complete observation. Laboratory cells are cycled to end of life. In this field cohort, 52% of units were still alive when observation ended. Their remaining life is not unknown-and-therefore-noisy; it is bounded below and otherwise unobserved. Discarding these units halves the cohort and biases the survivors out of the training distribution; treating the last observation as a failure teaches the model that every healthy battery was about to die. Neither treatment is acceptable, and both are common.

Right censoring at this rate is the defining statistical feature of field battery prognostics, and it is the reason we argue the task is misposed as regression.

1.4 Contributions

C1 — An openly specified, reproducible benchmark for lead-acid field prognostics. We release a complete pipeline from raw 60-second telemetry to evaluated predictions on 1,027 batteries, with deterministic battery-level splits, four leakage guards asserted in code, and results for five survival models under Uno's IPCW concordance, fixed-horizon AUC, Brier score and balanced accuracy. To our

knowledge this is the first such openly specified protocol for this chemistry, and it directly addresses the “no public dataset, no shared comparison” gap identified by the review literature [3].

C2 — The survival framing, quantified against regression. On the same data and largely the same features, a bidirectional Mamba RUL regressor attains 83.3 d MAE — worse than a one-line age heuristic (63.8 d) and a ridge regression (57.2 d) — while a landmark survival model attains AUC 0.848 at four weeks and 0.832 at twelve. The two task framings are evaluated on different metrics and are therefore not directly commensurable; the contrast we draw is that the regression framing cannot exceed trivial baselines here, while the survival framing, which consumes the censoring bound as information, yields strong discrimination. This pattern is consistent with censoring handling, rather than feature content, being the dominant factor.

C3 — A head-to-head improvement on the published result for this dataset. Replicating the protocol and metric of Aitio and Howey [1], we obtain 84.1% balanced accuracy eight weeks before failure against their published 73%, and 89.7% at the point of failure against 82%, using daily aggregates and no state-of-health estimate — strictly less input information than their Gaussian-process classifier consumed.

C4 — An ablation indicating that telemetry, not age, carries the signal. Across all five survival models, on the at-risk population, age alone is close to random (balanced accuracy 52.5–57.5%; AUC 0.542–0.546) while telemetry alone recovers the full model. This is the result that justifies instrumenting a battery at all, and it holds independently of the estimator.

C5 — Three negative results, reported in full. (i) Linear models matched or beat non-linear ones in four independent comparisons on this dataset; (ii) isotonic and Platt recalibration worsen the Brier score under a correctly nested protocol, although isotonic regression reduces worst-case calibration error by a factor of 3.3; (iii) OCV-anchored capacity estimation is biased 22% low because VRLA relaxation is incomplete even after three hours of rest. Each finding contradicts a reasonable prior expectation and would otherwise be rediscovered at cost.

2. Related work

2.1 Lead-acid state-of-health and RUL estimation

The mapping study [3] identifies 13 distinct algorithms across 17 papers. Multilayer perceptrons dominate (13 studies; 0.49–7.6% MAPE); LSTM and BiLSTM appear in four (approximately 2.8–3.0% error); radial-basis networks, fuzzy logic, k-nearest neighbours, particle-swarm-optimised models and a single random survival forest account for the remainder. The best reported figure, 0.0005% [3], is strongly temperature-dependent by the original authors' own account and derives from a controlled setting. The majority of these studies perform SoH estimation on laboratory data with capacity labels; they are therefore not baselines for the present task.

2.2 Fleet-scale survival approaches

The closest methodological precedent is the Linköping group's work on Scania heavy-truck starter batteries. Frisk, Krysander and Larsson [6] applied random survival forests to a fleet of 33,603 vehicles across five European markets using 291 fleet-management variables; Voronov, Frisk and Krysander [4] extended the approach to a cohort of more than 50,000 trucks with confidence estimation, and Voronov, Krysander and Frisk [5] addressed the sparse, non-equidistant sampling of the same fleet data; accumulative variables in this setting are treated in [7]. Two properties make this line of work a genuine precedent: the data are real, and they are right-censored. Two properties make it structurally different from ours: the fleet data contain no measurements related directly to battery health, and each vehicle contributes a single readout rather than a time series. Because those studies report error rate and concordance rather than fixed-horizon AUC, their headline numbers are not directly comparable to ours, and we do not compare numerically; we treat this work as methodological precedent for the survival framing.

2.3 The dataset and its original analysis

Aitio and Howey [1] released the dataset used here and reported a Gaussian-process classifier trained on estimated state of health, its gradient, and usage stress factors, achieving 82% balanced accuracy at the point of failure and 73% eight weeks in advance, against reported benchmarks of 66% and 51% respectively, under 5-fold stratified cross-validation. § 6.5 replicates this protocol.

2.4 Deep survival analysis

Deep survival methods — DeepSurv, DeepHit, Deep Survival Machines and Deep Recurrent Survival Analysis [13–16] — model censored event times directly, either through a Cox partial likelihood with a learned risk score or by parameterising a discrete-time event distribution. They are absent from the lead-acid literature. We do not evaluate them here (§ 9), but our results suggest that the framing, rather than the capacity of the model, is the binding constraint on this dataset.

3. Dataset

3.1 Source and composition

We use the Oxford University Research Archive release accompanying Aitio and Howey [1] (dataset [2], DOI 10.5287/bodleian:aVR4oDV4N), summarised in Table 1a below.

Table 1
a. Composition of the Oxford/BBOXX release [2].

Property	Value
Units	1,027 VRLA batteries
Configuration	12 V nominal, 20 Ah, six 2 V cells in series
Application	Solar home systems, 50 Wp PV panel, sub-Saharan Africa
Observation	400–760 days per unit [2]
Telemetry	UNIX timestamp, current, voltage, temperature
Sampling	Nominally 60 s, non-uniform
Volume	≈ 620 million rows, 2.9 GB compressed
Observed failures	**491 (47.8%)**
Right-censored	**536 (52.2%)**
Licence	CC BY-NC 4.0

An auxiliary laboratory file provides a galvanostatic intermittent titration (GITT) sweep on the same cell type, measured on a Biologic SP-150 (239,547 rows, 180 h).

3.2 Measured operating envelope

Aggregate statistics over the fleet, computed in this work, are given in Table 1b. The computed lifetime range of failed units (401–761 d) differs from the released observation range (400–760 d) by one day at each end; the day-counting convention should be reconciled against the release documentation, and the shorter bound (401 d) is the one used in § 5.3.

Table 1
b. Fleet operating envelope, computed in this work.

Quantity	Value
Terminal voltage, per cell	p1 2.071 V, mean 2.185 V, p99 2.413 V
Mean cell temperature	28.6°C (p95 33.5°C)
Equivalent full cycles per day	0.15–0.18
Observed lifetime, failed units	515 d mean (computed range 401–761 d)
Observed duration, censored units	546 d mean

The p99 per-cell voltage of 2.413 V is within 8 mV of the 2.42 V upper cut-off of the PyBaMM Sulzer2019 VRLA parameter set [20, 21], and the duty cycle of C/20–C/5 is consistent with stationary solar service. Both indicate that the fleet operates in the regime its specification implies.

3.3 Measured open-circuit-voltage curve

We derive an OCV–SOC curve from the GITT file rather than adopting a literature table. Rest periods are located where |I| < 1 mA for longer than 600 s; the final sample of each rest is taken as the relaxed voltage, charge and discharge branches are averaged to cancel hysteresis, and monotonicity is enforced. Forty usable rest periods were found (median duration 3.17 h), yielding a curve spanning 1.913–2.199 V per cell.

The measured curve differs materially from a standard textbook table — by up to 88 mV at 90% SOC — and maps a rested 12.6 V pack to 61% SOC rather than 90%. All state-of-charge quantities in this paper use the measured curve.

A caveat that propagates. Even after 3.17 h of rest the GITT voltages are still drifting, at approximately 28 mV per 10 min at the end of the rest. These are three-hour-relaxed values, not equilibrium values, and every SOC derived from them inherits that bias; § 7.3 quantifies the consequence.

4. Method

4.1 Stage 1 — daily aggregation

The diurnal cycle of a solar home system makes the day the natural unit of aggregation: one charge from the panel, one discharge overnight. We therefore aggregate the 60-second telemetry into per-battery-per-day records, the analogue of the per-cycle features a lithium-ion model consumes. Three conventions are fixed once, at ingest: (1) the source encodes negative current as charging, which is inverted so that positive denotes charge; (2) voltage collapses to approximately 0.01 V during sensor dropouts and disconnections, and these samples are removed rather than clipped; (3) every integral clamps its timestep to 900 s so that a multi-day communication gap is not integrated as continuous current. Days with fewer than 200 samples are discarded. The result is 547,037 battery-days across 1,027 batteries (median 523 days per battery), with 24 daily quantities: voltage statistics, temperature statistics, charge and discharge throughput, SOC statistics from the measured OCV curve, rest and charge duty fractions, coulombic efficiency, and derived swings.

4.2 Stage 2 — landmark survival construction

Each battery is observed for 400–760 days, and a prediction may be required at any point in that life. We therefore adopt landmarking [22], the standard construction for repeated-measures survival data. At each landmark day $L \in \{180, 240, \ldots, 540\}$: (i) batteries still at risk at L are retained; (ii) features are computed from the window [L − 90, L]; (iii) the outcome is recorded as (duration = lifetime − L, event = observed failure). This yields 6,003 battery-landmark observations from 1,026 of the 1,027 batteries — one unit contributes no valid landmark window under the data-sufficiency rule of § 4.1 — with 2,783 events (46.4 %).

4.3 Features

For each of the 24 daily quantities we compute four statistics over the 90-day window, giving 96 features plus battery age (97 in total): _mean (the level); _last (the mean of the final seven days, the most recent state); _slope (an ordinary-least-squares trend, annualised); and _std (volatility). The slope terms matter: a single snapshot cannot express degradation, whereas a trend can. Battery age is deliberately retained as a feature. It is not leakage — at inference the age of a battery is always known, and only its total lifetime is not — and § 7.2 records that excluding it was a material error in our own earlier experiments.

4.4 Models

We evaluate five survival estimators, all from scikit-survival 0.28 [17], with configurations listed in Table 1c; underlying methods are the Cox proportional hazards model [8], its elastic-net variant, random survival forests [10], extremely randomised survival trees and gradient-boosted survival analysis. Ensemble sizes were selected per fold from the stated range by inner validation; exact per-fold values are recorded in the released configuration files.

Table 1
c. Survival estimators and configurations (scikit-survival 0.28 [17]).

Model	Configuration
Cox proportional hazards	ridge penalty α = 1.0, standardised inputs
Coxnet	elastic net, ℓ_1 ratio 0.9
Random survival forest	200–300 trees, min leaf 15, √p features
Extra survival trees	200–300 trees, min leaf 15
Gradient-boosted survival	200–300 stages, learning rate 0.05, depth 3, subsample 0.8

For the head-to-head comparison of § 6.5 we additionally fit a class-weighted logistic regression and a histogram gradient-boosting classifier [18], matching the binary framing of the original study [1].

5. Experimental protocol

5.1 Leakage control

Four guards are asserted in code before any model is fitted, and each fails loudly rather than degrading silently: (1) *grouping by battery* — a battery contributes all of its landmark rows to exactly one fold, since splitting by row would place day-180 and day-420 of the same unit on both sides of a split; (2) *causal windows* — features come only from [L − 90, L], and the outcome is measured strictly after L; (3) *statistics fitted on training data only* — normalisation constants are computed on training folds; (4) *no target-derived features* — nothing in the feature set is computed from lifetime or the failure flag, and an assertion verifies that RUL = lifetime − landmark holds numerically for every row.

5.2 Metrics

Discrimination. Uno's IPCW concordance [11] is the headline discrimination measure; Harrell's C is computed alongside but is optimistically biased under the 52% censoring present here. We additionally report AUC at fixed horizons, computed as a binary AUC in which rows censored before the horizon — whose status at that horizon is unknown — are excluded; this estimator is unbiased when censoring is uninformative at the horizon, an assumption we adopt and state explicitly. *Calibration.* Brier score [12], expected calibration error (ECE) and maximum calibration error (MCE). *Comparability.* Balanced accuracy, (sensitivity + specificity)/2, for comparability with the published result on this dataset [1]; on a 9% base rate, raw accuracy reaches 91% trivially and is not reported. Confidence intervals are bootstrapped over batteries, not rows, because rows from one battery are not independent.

5.3 The at-risk population

The shortest observed lifetime in the fleet is 401 days. At landmark day 180, failure within 28 days is therefore arithmetically impossible, and every such row is a guaranteed negative that battery age identifies perfectly. Including these landmarks inflates AUC without any health information being used. We therefore report two populations throughout: the full landmark set and the at-risk subset ($L \geq 420$ d; 1,961 rows, 868 events, 927 batteries). The at-risk figures are the ones we consider quotable, and we recommend the same practice for future work on this dataset.

6. Results

6.1 Survival model comparison

Table 1
reports the five survival models under grouped 5-fold cross-validation, on the full landmark set and on the at-risk subset; Fig. 1 visualises the four-week AUC for both populations.

Model	Uno C (full)	AUC@4wk (full)	AUC@12wk (full)	Uno C (at-risk)	AUC@4wk (at-risk)	AUC@12wk (at-risk)	Brier@4wk
Cox PH	0.783	0.941	0.916	0.754	0.845	0.827	0.0750
Coxnet	0.779	0.938	0.916	**0.756**	0.847	0.828	0.0736
Random survival forest	0.759	0.912	0.880	0.747	0.843	0.829	**0.0695**
Extra survival trees	0.751	0.882	0.856	0.737	0.819	0.812	0.0735
Gradient-boosted survival	0.777	0.935	0.912	0.749	**0.848**	**0.832**	0.0697

Table 1. Five survival models, grouped 5-fold cross-validation: full landmark set versus at-risk subset. Bold marks the best value in each at-risk column.

Four of the five models fall within 0.843–0.848 at the four-week horizon on the at-risk population — differences smaller than the width of the bootstrap confidence intervals. Algorithm choice is not the binding constraint on this dataset; the framing is.

Notably, the linear Cox proportional hazards model matches or exceeds the random survival forest on this dataset. We attribute this to feature structure: the covariates are dense continuous trends over which the log-hazard is approximately linear, a regime in which a linear risk score has little to gain from tree ensembles.

The headline result, bootstrapped over batteries, is: gradient-boosted survival, AUC 0.848 (95% CI 0.819–0.877) at four weeks and 0.832 (0.808–0.855) at twelve weeks. Both intervals exclude the no-discrimination value of 0.5 by a wide margin, and their lower bounds additionally exceed 0.80. We do not compare numerically against the heavy-truck lead-acid literature [4–7]: those studies predict from a single readout per vehicle with no battery-health measurements, over a different failure definition, and report error rate and concordance rather than fixed-horizon AUC. The comparison that is like-for-like — same dataset, same metric, same protocol — is against Aitio and Howey [1] in § 6.5.

6.2 Ablation: does the telemetry earn its place?

Table 2
and Fig. 3 report a feature ablation on the at-risk population for all five models.

Model	Features	Uno C	AUC@4wk	AUC@12wk
Cox PH	all	0.754	0.845	0.827
	age only	**0.534**	**0.545**	0.561
	telemetry only	0.755	0.844	0.824
Coxnet	all	0.756	0.847	0.828
	age only	0.534	0.545	0.561
	telemetry only	0.755	0.845	0.824
Random survival forest	all	0.747	0.843	0.829
	age only	0.533	0.546	0.565
	telemetry only	0.747	0.843	0.828
Extra survival trees	all	0.737	0.819	0.812
	age only	0.533	0.546	0.565
	telemetry only	0.739	0.823	0.813
Gradient-boosted survival	all	0.749	0.848	0.832
	age only	0.531	0.542	0.564
	telemetry only	0.750	0.848	0.833

Table 2. Feature ablation on the at-risk population, all five models.

Within the at-risk population, age alone is close to random guessing (AUC 0.542–0.546) and telemetry alone recovers the full model to within the reported precision. The result is consistent across five structurally different estimators, so it is a property of the data rather than of any one algorithm.

This resolves an apparent paradox in our own experiments. Across the whole fleet, age dominates — it separates “too young to fail” from “old enough to be at risk” trivially, and a one-line age heuristic beats a 627 k-parameter network at RUL regression (§ 7.1). Within the at-risk population, that separation is already made and the sensors do the remaining work. Both statements are true; only the second is relevant to a deployed system, which is only ever asked about batteries currently in service.

6.3 Calibration

Table 3

Brier score under a nested protocol: outer grouped CV for evaluation; inner 70/30 battery-level split within each training fold, with 70% fitting the model and 30% fitting the calibrator. Values differ slightly from Table 1 (e.g. GBS 0.0704 vs 0.0697 at 4 wk) because the nested protocol reserves 30% of each training fold for the calibrator, reducing the data available for model fitting.

Model / horizon	Uncalibrated	Isotonic	Platt
GBS, 4 wk	**0.0704**	0.0744	0.0766
GBS, 12 wk	**0.1457**	0.1491	0.1492
RSF, 4 wk	**0.0704**	0.0732	0.0782
RSF, 12 wk	0.1496	**0.1488**	0.1501

Recalibration does not improve the Brier score: the uncalibrated model wins three of four comparisons, and the single isotonic win (0.0008) is within noise; Platt scaling is worst throughout. There is nonetheless a genuine trade-off, visible in the calibration-error metrics. For the random survival forest at four weeks, isotonic regression reduces ECE from 0.0273 to 0.0165 and maximum calibration error from 0.1196 to 0.0357 — a 3.3× reduction in worst-case error — at a cost of approximately 0.008 AUC, attributable to ties introduced by the isotonic step function. This is the Murphy decomposition operating as expected: isotonic regression improves calibration and pays in refinement.

Practical rule. Where a probability is quoted contractually — warranty pricing, subscription alerting — apply isotonic regression and accept the AUC cost, because worst-case error governs. Where the output is only a ranking — which site to visit first — use the uncalibrated score.

Validity check. Isotonic regression is a monotone non-decreasing map and therefore cannot change the ranking of distinct scores; any AUC change after calibration arises solely from ties introduced by its step function. The released code asserts that pre- and post-calibration AUC agree up to this tie-attributable difference, and the assertion holds.

6.4 What drives the prediction

Table 4
Permutation importance, gradient-boosted survival, at-risk population, four-week AUC: the seven most important features of 97, together with the highest-ranked temperature feature (12th).

Feature	Δ AUC
ah_out_last — recent discharge throughput	+ 0.0122
eq_cycles_last — recent equivalent full cycles	+ 0.0122
frac_rest_last — recent fraction of time at rest	+ 0.0119
v_range_mean — mean daily voltage span	+ 0.0082
ah_in_slope — trend in charge accepted	+ 0.0064
ah_in_last — recent charge accepted	+ 0.0054
ah_out_slope — trend in discharge throughput	+ 0.0047
t_max_mean — mean daily peak temperature (ranked 12th)	+ 0.0031

Recent throughput and rest fraction dominate. Absolute magnitudes are small because the features are correlated and permutation importance divides credit among them; the ordering is the informative content.

Temperature ranks twelfth and is weak. This is consistent across every analysis we performed. In a univariate comparison over the 524-battery subset for which extracted capacity and resistance traces are available (§ 7.3; 352 failed, 172 surviving), mean temperature did not discriminate ($p = 0.81$), nor did extracted internal-resistance growth ($p = 0.72$), while charge acceptance ($p = 0.0009$), time below 50% SOC ($p = 0.0018$) and coulombic efficiency ($p = 0.0048$) did. We interpret the temperature result as an absence of variance rather than an absence of effect: this fleet occupies a single climate band (mean 28.6°C, p95 33.5°C), so temperature cannot explain differential failure within it even though the Arrhenius dependence of VRLA float life is well established. A multi-climate fleet would be required to recover it.

6.5 Head-to-head with the published result

Aitio and Howey [1] report, on this dataset, 82% balanced accuracy at the point of failure against a 66% benchmark, and 73% eight weeks in advance against a 51% benchmark, under 5-fold stratified cross-validation, using a Gaussian-process classifier over estimated SoH, its gradient, and usage stress factors computed from the full 60-second sequences.

Three protocol elements must match for the comparison to be meaningful, and we align all three: balanced accuracy as the metric; failure-anchored windows, in which each case is anchored h days before its own failure rather than at a fixed calendar landmark; and age-matched controls, since cases are near end of life and therefore old by construction. The decision threshold is selected by Youden's J on the training folds only. Table 5 and Fig. 6 report the horizon sweep for the class-weighted logistic

regression, the stronger of our two classifiers; the gradient-boosting classifier tracks 3–6 points below it and also exceeds both published operating points.

Table 5
Balanced accuracy versus prediction horizon on the published protocol (class-weighted logistic regression). Grouped 5-fold cross-validation; approximately 483 cases (failed units with complete feature windows at the given horizon) and an equal number of age-matched controls at each horizon.

Horizon	This work (logistic)	Published [1]	Published benchmark [1]
0 d (at failure)	**89.7%**	82%	66%
7 d	88.6%	—	—
14 d	88.6%	—	—
28 d	85.9%	—	—
42 d	83.6%	—	—
56 d (8 weeks)	**84.1%**	**73%**	51%
84 d (12 weeks)	77.8%	—	—

We improve on the published figure by 7.7 points at the point of failure and 11.1 points at eight weeks, on the same dataset and the same metric, while using daily aggregates and no state-of-health estimate — strictly less input information than the original method consumed.

Table 6
Confound control for Table 5. Age-matched control sets are drawn independently for this experiment; the small differences from Table 5 (e.g. 82.6% versus 84.1% at 56 d) reflect control-resampling variability, quantified in § 9 (item 3).

Horizon	All features	Age only	No age
0 d	89.7%	**57.5%**	89.1%
28 d	85.5%	54.9%	85.7%
56 d	82.6%	53.3%	82.4%
84 d	77.9%	52.5%	77.4%

What we could not match, and why it does not favour us. Aitio and Howey do not publish their fold assignments, so our folds are necessarily different; we use grouped 5-fold cross-validation with the battery as the grouping unit, which is at least as strict as stratified 5-fold because it additionally forbids a battery contributing rows to both train and test. We use daily aggregates where they used the full 60-second sequences, and we supply no state-of-health estimate where their classifier consumed one — both are strictly less information, so neither can inflate our figure. The failure definition, the failure-

anchored windowing and the age-matched control construction follow their description. The one element that could favour us is threshold selection; we therefore select the Youden's J threshold on training folds only, never on the evaluation fold.

Age matching leaves a residual 16–36 day difference between cases and controls. Table 6 indicates this is immaterial: age alone is near-random, and removing age entirely leaves performance essentially unchanged.

7. Negative results

We report three findings that contradict reasonable prior expectations. Each cost substantial effort, and each would otherwise be rediscovered by the next group.

7.1 A deep sequence model is beaten by a one-line heuristic

We implemented a bidirectional selective state-space (Mamba [19]) network for RUL regression — 627 k parameters, two bidirectional blocks, cross-attention over four learned degradation anchors, and a heteroscedastic output head — and trained it on 47,718 causal 30-day windows under the same leakage guards. Censored units enter the training objective through a one-sided censoring hinge in the v1 variant and a right-censored log-normal (accelerated-failure-time) likelihood in the v2 variant. Table 7 and Fig. 7 report mean absolute error on a held-out test set of 155 unseen batteries.

Table 7
RUL regression, held-out test on 155 unseen batteries.

Predictor	MAE (days)
Predict the training-mean RUL	121.4
Ridge regression on the flattened window	**57.2**
Gradient boosting on the flattened window	59.6
Gradient boosting on age alone	59.9
mean lifetime − age (one line, no fitting)	63.8
BiMamba, v1	83.3
BiMamba, v2 (log target, AFT censored likelihood)	129.4

A ridge regression fitted in under a second outperforms both network variants, and age alone comes within 2.7 days of the best result. The v2 variant, which added the age feature and replaced the censoring hinge with the right-censored log-normal likelihood, achieved better calibration (prediction-interval coverage 0.63 → 0.88) but worse accuracy ($R^2 = -1.18$, worse than a constant predictor), which

we attribute to the log-space target amplifying errors at long horizons through the exponential inverse transform.

Counting all experiments in this paper, linear models matched or beat non-linear ones in four independent comparisons on this dataset: (i) ridge regression versus BiMamba (Table 7); (ii) the age heuristic versus BiMamba (Table 7); (iii) Cox proportional hazards versus the random survival forest (Table 1); and (iv) the logistic classifier versus the gradient-boosting classifier (§ 6.5). We take this as evidence that the achievable signal is close to linear in these features, not that state-space models are unsuitable in general.

7.2 Excluding battery age as a leakage precaution is an error

In our first experiment we excluded battery age from the feature set, reasoning that the target (lifetime − age) is a function of it. This is incorrect: at inference the age of a deployed battery is always known, and only its total lifetime is unknown. Because the observed duration in this fleet is 533 ± 89 days, age is a strong prior, and its exclusion is the principal reason the v1 network underperformed a trivial baseline. (The mean of observed durations pools failure times with censoring times and therefore underestimates the true mean lifetime; for the baseline heuristic in Table 7 this simply makes the baseline slightly weaker than it could be.) We record the error because the reasoning that produced it is superficially plausible.

7.3 OCV-anchored capacity estimation is biased low

Field data contain no capacity test, so we attempted to infer state of health by anchoring SOC at two rested points via the measured OCV curve and integrating the current between them, giving capacity = $\int I\,dt / \Delta SOC$. The chemistry is favourable — lead-acid OCV is near-linear in SOC, with none of the plateau that defeats this method on LFP — and rest periods are abundant (median 425 rests exceeding 1 h per battery; 147 exceeding 4 h).

The estimator nonetheless returns a median baseline capacity of 15.5 Ah against a 20 Ah nameplate — a 22% low bias — and yields a usable trace for only 524 of the 1,027 batteries. Diagnosis: at 8 h of rest the same batteries read 21–22 Ah, consistent with incomplete relaxation inflating ΔSOC and deflating the capacity estimate. This is a physical limitation of field data rather than an implementation defect: solar home systems do not rest long enough, often enough. Pairing consecutive anchors additionally starves the estimator, since a solar battery rests at a similar state of charge most nights; pairing over a forward window of anchors is required to find the large excursions that make the ratio well conditioned.

Consequence for the field. State-of-health regression on field lead-acid data without capacity ground truth should be treated as unvalidated. Failure classification, which requires no capacity label, is the defensible target.

8. Discussion

8.1 Why the framing dominates the architecture

The results of § 6.1 and § 7.1 come from the same batteries, the same telemetry and largely the same features; the difference lies in how the question is posed. Regression on remaining life must either discard 52% of the cohort or fabricate a label for it, whereas survival analysis consumes the censoring bound as information. That single change moves the task from worse than a one-line heuristic to better than the published result on this dataset. We suspect this generalises: field prognostics datasets are censored by construction — assets in service have not yet failed — and the lead-acid literature's near-total reliance on regression-style MLPs may therefore understate what its own data support.

8.2 Implications for instrumented battery monitoring

Table 2 is the commercially decisive result. If age had remained dominant within the at-risk population, monitoring hardware would be difficult to justify: a maintenance calendar would suffice. That telemetry alone recovers the full model, consistently across five estimators, indicates that measurement adds information a calendar cannot. The operating characteristic supports fleet triage directly: the top risk quintile exhibits a 31.5% observed four-week failure rate against 0.6% in the bottom quintile — a 52-fold ratio. At 70% recall, precision is 30% against a 9% base rate. For a subscriber-facing alert we would advise operating at lower recall and higher precision, since false alarms erode trust faster than missed detections in a service context.

8.3 On temperature

Temperature's absence from the discriminative features (§ 6.4) should not be read as evidence against the Arrhenius dependence of VRLA float life, which is well established. It is an absence of variance in a single-climate fleet. A study spanning climate zones — which the Indian market would naturally provide — would be required to recover the effect, and we regard this as the single most valuable extension of the present dataset.

9. Limitations

1. *Single fleet, single design, single climate.* All 1,027 units are the same 12 V/20 Ah VRLA product on the same 50 Wp system in one climate band. Generalisation to tubular flooded cells, telecom float duty or multi-climate fleets is untested.

2. *No capacity ground truth.* The dataset contains no reference capacity test, so state of health cannot be validated (§7.3) and only failure timing is supervised.

3. *The comparison of §6.5 is a replication, not a re-execution.* We match metric, horizon sweep, cross-validation scheme and case anchoring, but the original controls were drawn from field workshop repair records while ours are age-matched survivors, and the original features included an SoH estimate we do

not compute. Resampling the controls moves the 56-day figure by approximately 1.5 points (82.6–84.1 %), which bounds the sensitivity of Tables 5 and 6 to control selection.

4. *Deep survival methods are not evaluated.* DeepHit, Deep Survival Machines and Deep Recurrent Survival Analysis [14–16] are the natural next comparison and are absent. Given §7.1, we would not expect large gains, but the experiment remains to be done.

5. *Licence.* The dataset is CC BY-NC. Results are reproducible for research; commercial deployment of a model trained on it requires separate permission.

6. *Failure definition.* "Failure" is entry into the operator's repair system, a partly operational event. Some variance may be irreducible from telemetry alone.

10. Conclusion

This work establishes, to our knowledge, the first openly specified and reproducible benchmark for lead-acid battery prognostics on field data, addressing a gap the field's own review literature identifies explicitly [3]. The central result is methodological: with 52% right censoring, the task is substantially better posed as survival analysis than as regression. Under the regression framing, a 627 k-parameter bidirectional state-space network is beaten by a one-line age heuristic; under a landmark survival framing on the identical data, gradient-boosted survival attains AUC 0.848 (95% CI 0.819–0.877) at four weeks and 0.832 at twelve, and the replication of the published protocol yields 84.1% balanced accuracy eight weeks before failure against a published 73% on the same dataset and metric.

A five-model ablation indicates that, within the at-risk population, telemetry rather than battery age carries the predictive signal — the empirical justification for instrumenting a battery at all. We additionally report that probability recalibration is unnecessary under a correctly nested protocol, that linear models matched or beat non-linear ones in four independent comparisons here, and that OCV-anchored capacity estimation on field data is biased 22% low by incomplete relaxation.

The immediate extensions are a multi-climate fleet, which would recover the temperature dependence a single-climate dataset cannot express, and controlled cycling with capacity ground truth, which would make state of health a supervised rather than an inferred quantity.

Declarations

Data and code availability

Data. The battery dataset is the Oxford University Research Archive release accompanying Aitio and Howey [2], DOI 10.5287/bodleian:aVR4oDV4N, comprising 1,027 field-deployed 12 V/20 Ah VRLA batteries. It is distributed under CC BY-NC 4.0; this work is non-commercial academic research and complies with that licence. No raw data are redistributed.

Code. The full pipeline — raw telemetry ingestion, the 97-feature landmark construction, the deterministic battery-level splits, the four leakage guards asserted in code, and the evaluation harness for all five survival models — is released at [insert repository URL].

Reproducibility. All splits are deterministic and seeded at the battery level; no battery appears in more than one fold. Every numerical result reported here is reproduced from logged experiment output rather than restated from an earlier draft.

References

1. Aitio A, Howey DA (2021) Predicting battery end of life from solar off-grid system field data using machine learning. *Joule* 5(12), 3204–3220 10.1016/j.joule.2021.11.006. Preprint: arXiv:2107.13856
2. Aitio A, Howey D (2021) Univ Oxf Res Archive. 10.5287/bodleian:aVR4oDV4N. Licence: CC BY-NC 4.0 Data repository for 'Predicting battery end of life from solar off-grid system field data using machine learning'
3. Chevtchenko SF, da Silva Rocha E, Cruz B, de Andrade EC, de Araújo (2023) D. R. B. A mapping study of machine learning methods for remaining useful life estimation of lead-acid batteries. arXiv:2307.05163
4. Voronov S, Frisk E, Krysander M (2018) Data-driven battery lifetime prediction and confidence estimation for heavy-duty trucks. IEEE Trans Reliab 67(2):623–639. 10.1109/TR.2018.2803798
5. Voronov S, Krysander M, Frisk E (2020) Predictive maintenance of lead-acid batteries with sparse vehicle operational data. Int J Prognostics Health Manage 11(1). 10.36001/ijphm.2020.v11i1.2608
6. Frisk E, Krysander M, Larsson E (2014) Data-driven lead-acid battery prognostics using random survival forests. *Annual Conference of the PHM Society* 6(1) 10.36001/phmconf.2014.v6i1.2370
7. Frisk E, Krysander M (2015) Treatment of accumulative variables in data-driven prognostics of lead-acid batteries. IFAC-PapersOnLine 48(21):105–112. 10.1016/j.ifacol.2015.09.512
8. Cox DR (1972) Regression models and life-tables. J Roy Stat Soc B 34(2):187–202
9. Kaplan EL, Meier P (1958) Nonparametric estimation from incomplete observations. J Am Stat Assoc 53(282):457–481
10. Ishwaran H, Kogalur UB, Blackstone EH, Lauer MS (2008) Random survival forests. Annals Appl Stat 2(3):841–860. 10.1214/08-AOAS169
11. Uno H, Cai T, Pencina MJ, D'Agostino RB, Wei LJ (2011) On the C-statistics for evaluating overall adequacy of risk prediction procedures with censored survival data. Stat Med 30(10):1105–1117. 10.1002/sim.4154
12. Graf E, Schmoor C, Sauerbrei W, Schumacher M (1999) Assessment and comparison of prognostic classification schemes for survival data. Stat Med 18(17–18):2529–2545
13. Katzman JL, Shaham U, Cloninger A, Bates J, Jiang T, Kluger Y (2018) DeepSurv: personalized treatment recommender system using a Cox proportional hazards deep neural network. BMC Med Res Methodol 18:24. 10.1186/s12874-018-0482-1

14. Lee C, Zame WR, Yoon J, van der Schaar M (2018) DeepHit: a deep learning approach to survival analysis with competing risks. *Proceedings of the AAAI Conference on Artificial Intelligence* 32(1)
15. Nagpal C, Li X, Dubrawski A (2021) Deep Survival Machines: fully parametric survival regression and representation learning for censored data with competing risks. IEEE J Biomedical Health Inf 25(8):3163–3175. 10.1109/JBHI.2021.3052441
16. Ren K, Qin J, Zheng L, Yang Z, Zhang W, Qiu L, Yu Y (2019) Deep recurrent survival analysis. *Proceedings of the AAAI Conference on Artificial Intelligence* 33(1), 4798–4805
17. Pölsterl S (2020) scikit-survival: a library for time-to-event analysis built on top of scikit-learn. J Mach Learn Res 21(212):1–6
18. Pedregosa F et al (2011) Scikit-learn: machine learning in Python. J Mach Learn Res 12:2825–2830
19. Gu A, Dao T (2023) Mamba: linear-time sequence modeling with selective state spaces. arXiv:2312.00752
20. Sulzer V, Chapman SJ, Please CP, Howey DA, Monroe CW (2019) Faster lead-acid battery simulations from porous-electrode theory: Part I. Physical model. J Electrochem Soc 166(12):A2363–A2371. 10.1149/2.0301912jes
21. Sulzer V, Marquis SG, Timms R, Robinson M, Chapman SJ (2021) Python Battery Mathematical Modelling (PyBaMM). J Open Res Softw 9(1):14. 10.5334/jors.309
22. van Houwelingen HC (2007) Dynamic prediction by landmarking in event history analysis. Scand J Stat 34(1):70–85. 10.1111/j.1467-9469.2006.00529.x

Figures

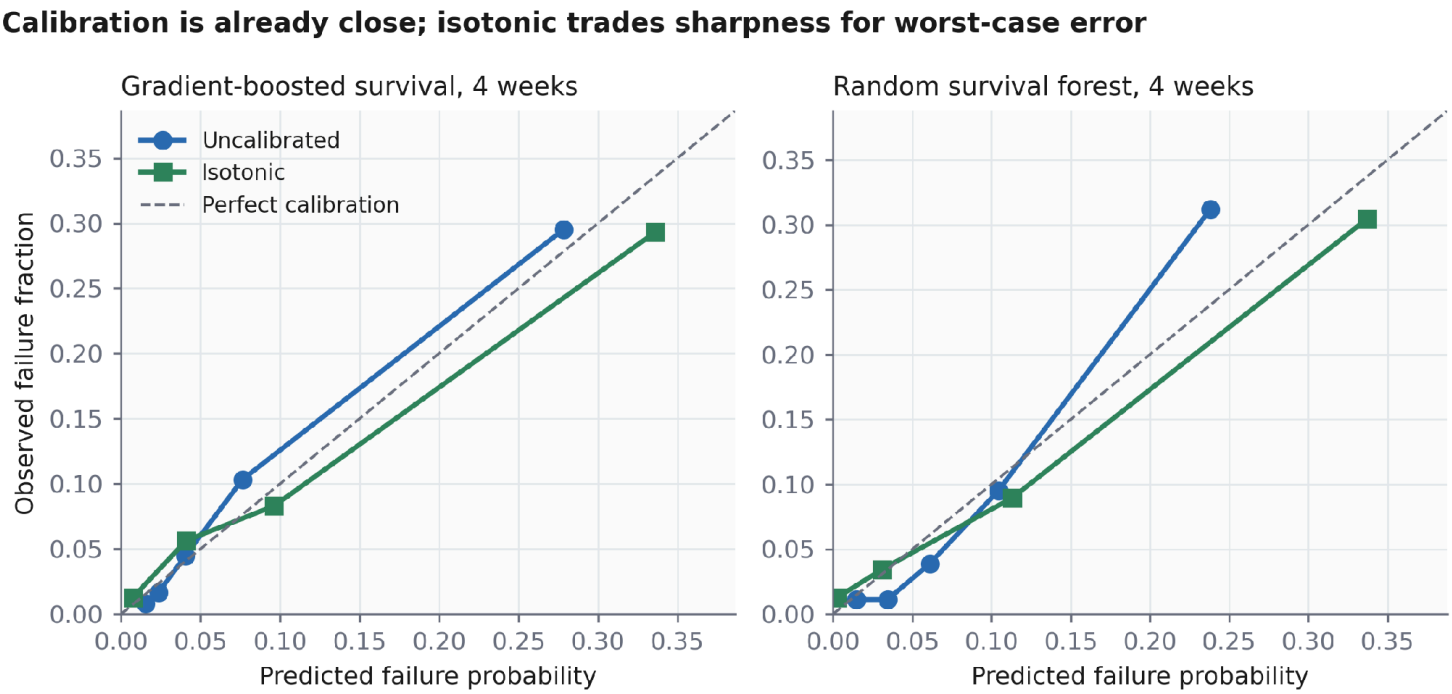


Figure 1

Four-week AUC for each survival model on the full landmark set (grey) and the at-risk population (blue). The gap between the two bars is the inflation contributed by landmarks at which failure within the horizon is arithmetically impossible (§5.3).

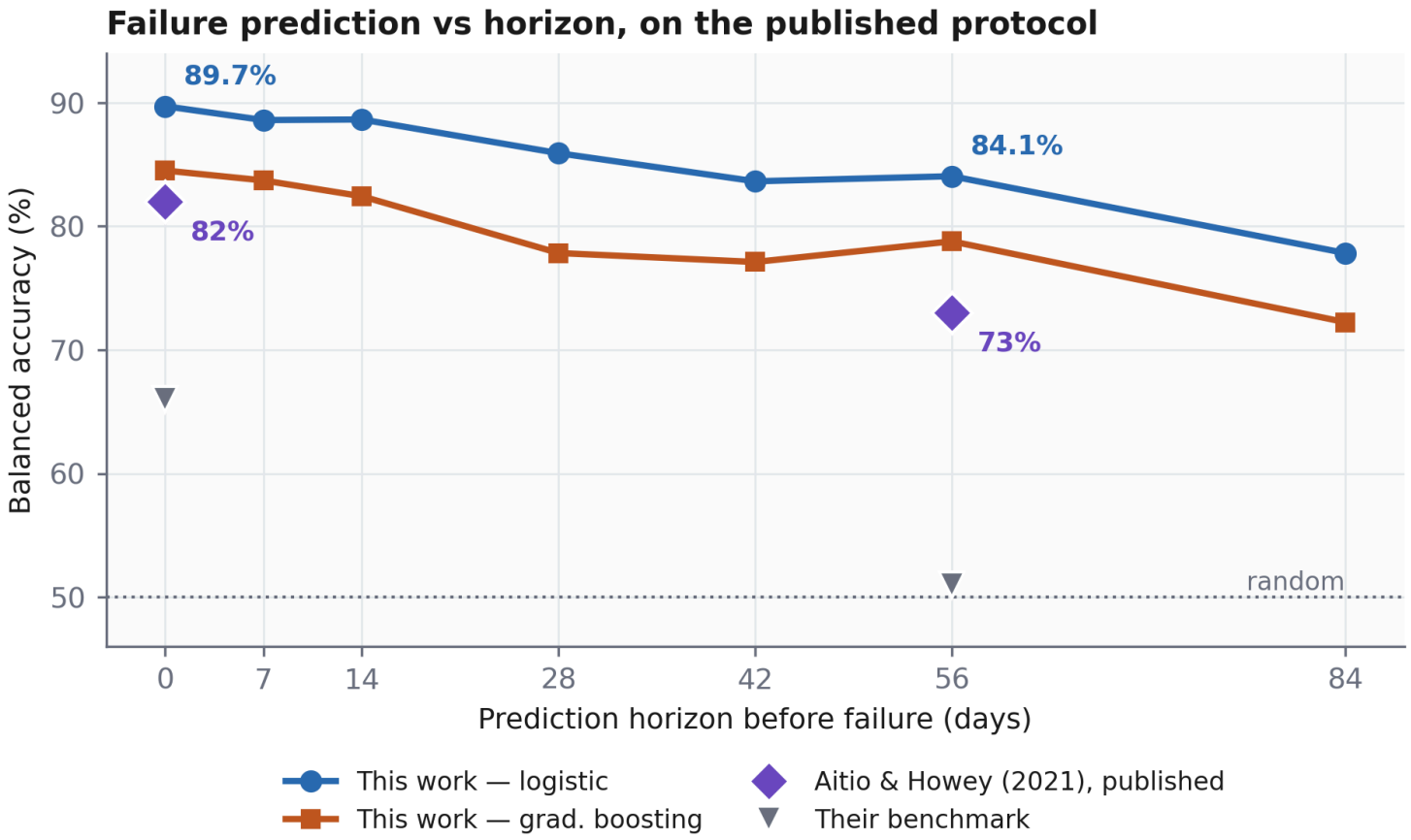


Figure 2

Kaplan–Meier survival curves [9] for held-out batteries grouped into quintiles of predicted risk. Separation is monotone across all five quintiles and is already established by the four-week horizon.

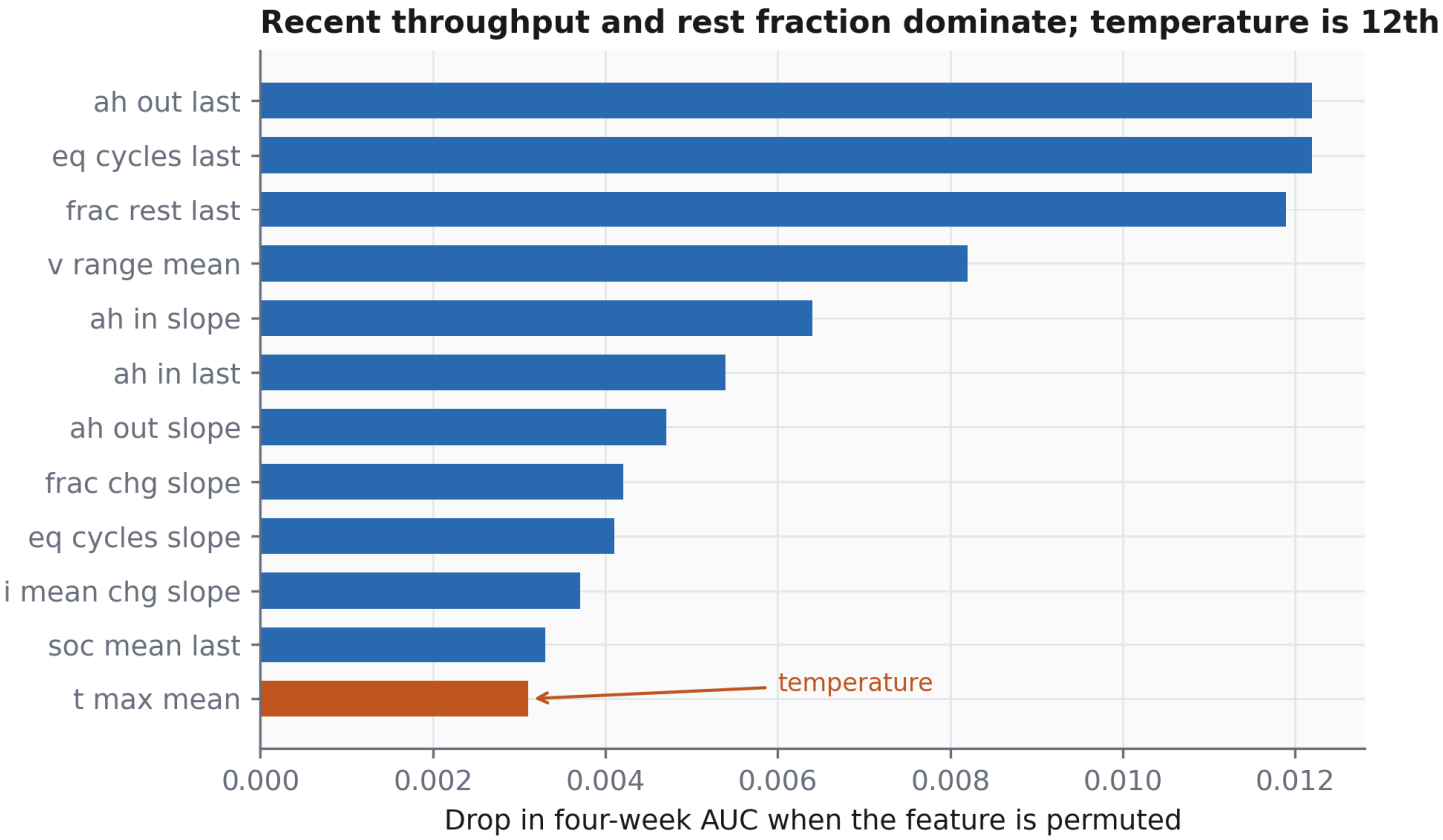


Figure 3

Feature ablation on the at-risk population. Age alone (orange) sits at the random line for every one of the five estimators, while telemetry alone (green) recovers the full model (blue). The result is algorithm-independent.

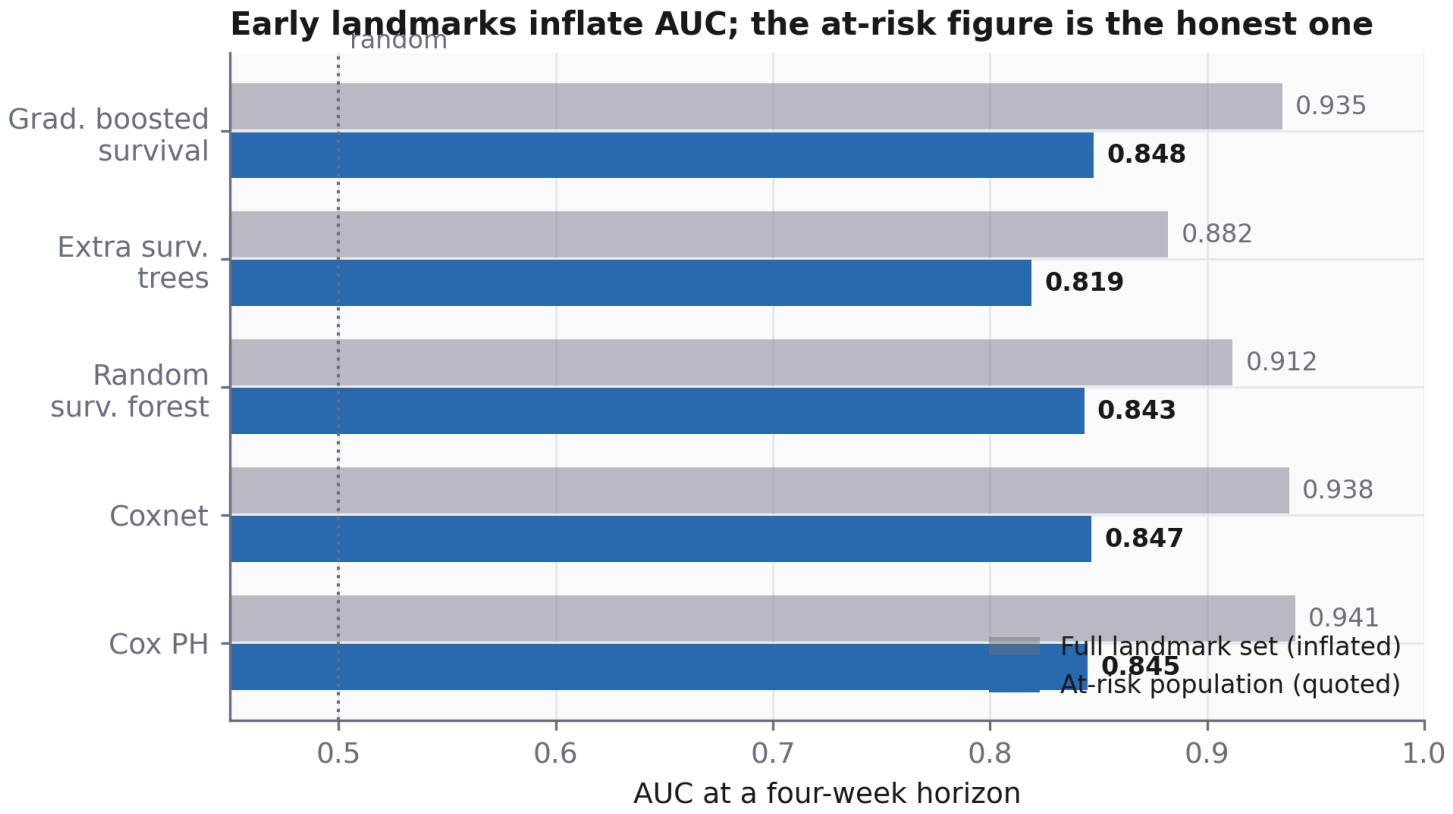


Figure 4

Reliability diagrams at the four-week horizon. Uncalibrated predictions already track the diagonal closely; isotonic regression shifts the highest-risk bin rightward, improving calibration at the cost of sharpness.

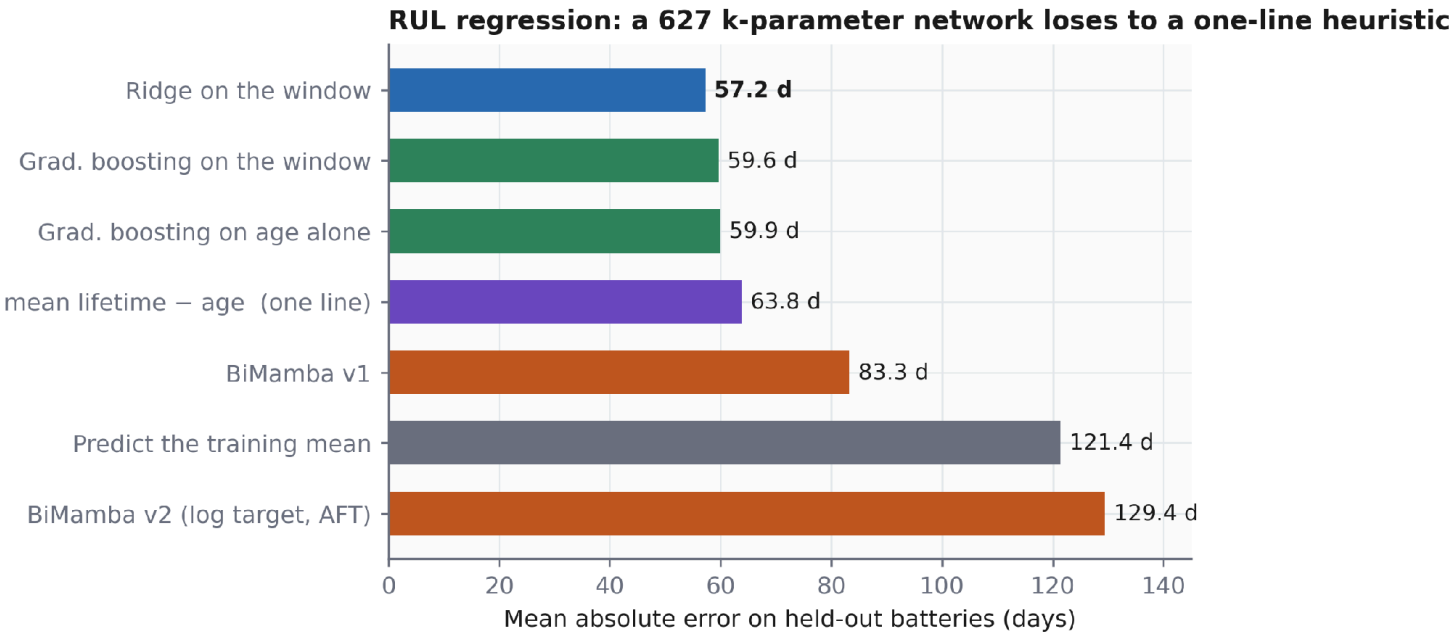


Figure 5

Permutation importance for the top twelve of 97 features. Recent discharge throughput, equivalent cycles and rest fraction dominate; mean peak temperature (orange) ranks twelfth.

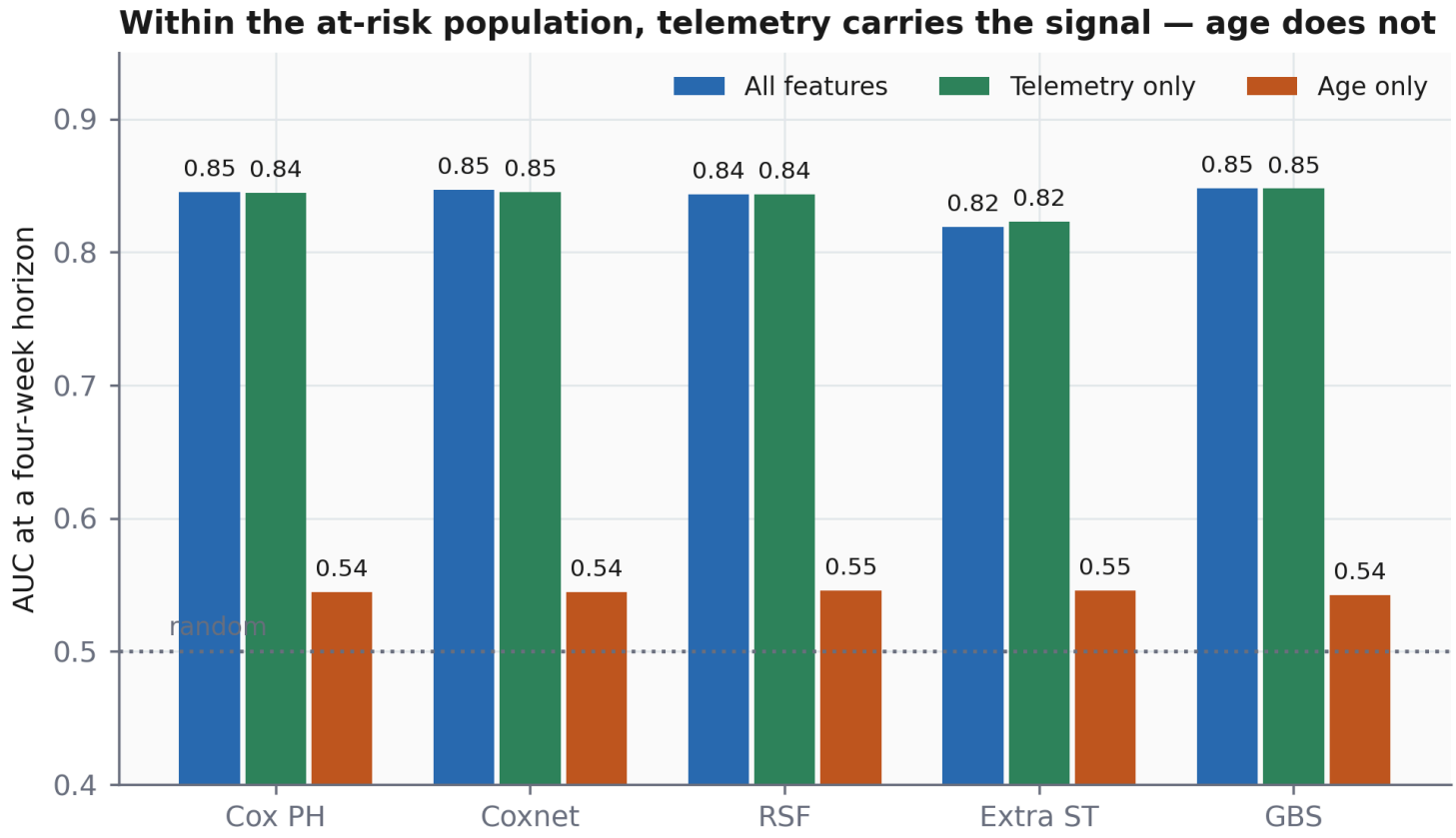


Figure 6

Balanced accuracy against prediction horizon on the published protocol. Both of our models exceed the two published operating points (purple diamonds) across the horizon sweep; grey triangles mark the benchmarks reported alongside them [1].

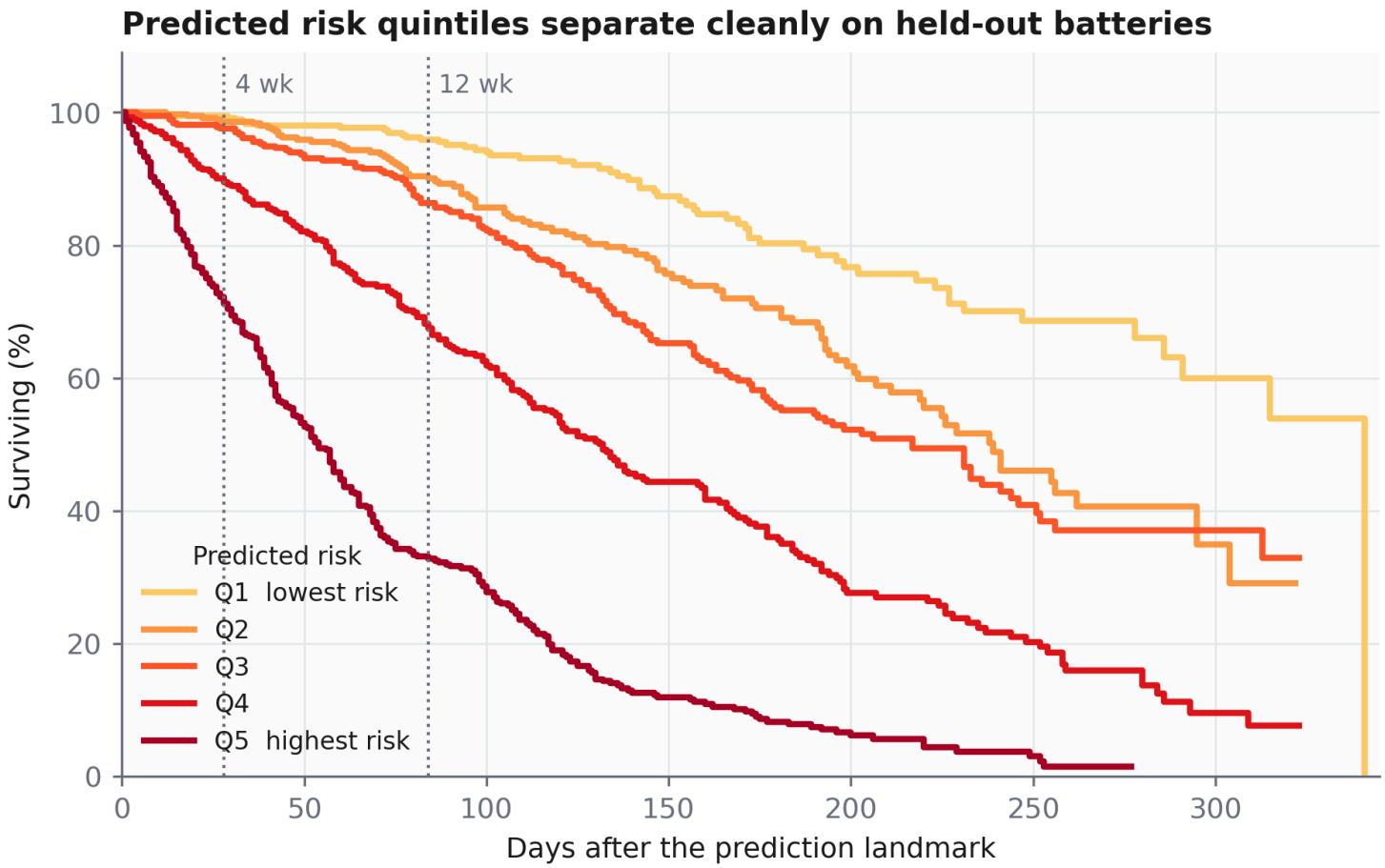


Figure 7

Mean absolute error for remaining-useful-life regression on held-out batteries. A ridge regression on the flattened window (blue) outperforms both bidirectional state-space variants (orange) and the one-line age heuristic (purple).