

Endpoint definitions and censoring in battery lifetime prediction across sources

Jufeng Li

q3529416243@gmail.com

China University of Petroleum, East China

Wenjie Li

China University of Petroleum, East China

Tao Li

China University of Petroleum, East China

Research Article

Keywords: battery lifetime prediction; reliability evaluation; right censoring; endpoint definition; calibration

Posted Date: October 27th, 2026

DOI: <https://doi.org/10.21203/rs.3.rs-11508938/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.

Endpoint definitions and censoring in battery lifetime prediction across sources

Wenjie Li¹ | Jufeng Li² | Tao Li³

¹College of Mechanical and Electronic Engineering, China University of Petroleum (East China), Qingdao 266580, China

²College of Science, China University of Petroleum (East China), Qingdao 266580, China

³Qingdao Institute of Software and College of Computer Science and Technology, China University of Petroleum (East China), Qingdao 266580, China

Correspondence

Jufeng Li, College of Science, China University of Petroleum (East China), 66 West Changjiang Road, Huangdao District, Qingdao 266580, China. Email: q3529416243@gmail.com. Telephone: +86 19863720183. ORCID: <https://orcid.org/0009-0009-2137-0020>. Coauthor emails: dvien5162@gmail.com; 19863708112@163.com.

Abstract

Reliable comparison of battery lifetime predictors requires a common endpoint and explicit treatment of incomplete trajectories. We examine how capacity references, right censoring and access to completed target lifetimes affect evaluation across experimental sources. A seven-source registry contains 220 physical cells; 214 contribute 318,707 cycle records. At an 80% threshold, changing from effective nominal to first-observation references changes SNL observed events from 50 to 35 and HUST events from 0 to 77. For SNL predictions at cycle 100, event-only mean absolute percentage error is 36.31%, whereas 11 censored cells imply a full-cohort identification range of 45.37–47.80% without imputing lifetimes. At the observed 85% HUST endpoint, three-label calibration reduces allocation-averaged error from 75.64% to 16.25%, compared with 20.25% for a matched target-only constant and 14.23% for ordinary unshrunk calibration. Exhaustive evaluation of all 73,150 HUST and 19,600 SNL three-label allocations reveals consistent improvement on HUST but a harmful tail on SNL, where average event-only error rises to 48.50%. The allocations describe selection variability in fixed cohorts, not independent replications. These retrospective case studies support an engineering evaluation procedure that declares endpoint references, separates observed and censored outcomes, reports cohort-level error bounds, and compares adaptation methods using identical label budgets and test cells.

KEYWORDS

battery lifetime prediction; reliability evaluation; right censoring; endpoint definition; calibration

1 | INTRODUCTION

Predicting the lifetime of a lithium-ion cell from its first hundred cycles has been widely studied within individual experimental campaigns^{1,2}. Public resources have moved the field towards larger and more heterogeneous collections: BatteryLife assembles sixteen datasets, eighteen baseline models and explicit cross-condition and cross-domain tasks¹, and inter-cell deep models extend prediction across ageing conditions and chemistries³. Against that background, the contribution of the present study is deliberately narrow. We do not introduce a dataset, a network architecture, or a new cross-source task. We ask what has to be reported for a cross-source lifetime number to mean the same thing to two different readers.

Three evaluation decisions require explicit reporting. The first is the *reference capacity* that turns a percentage threshold into an absolute capacity. A "80% end of life" defined against the nominal capacity, against the capacity of the first measured cycle, or against nominal capacity multiplied by the width of the state-of-charge (SOC) window can select different absolute capacities, and therefore different events. The second is the treatment of cells whose record ends before the threshold is reached. A benchmark that reports accuracy only on cells that reached the endpoint answers a question about those cells; it does not answer the same question about the full cohort that a deployment would face. The third is the label budget. When a small number of target cells with completed lifetimes are used to adapt a transferred model, the evaluation must exclude exactly those cells, hold the remaining test set fixed for every comparator, and include constants computed from the same labels; otherwise a large improvement over a transferred model may reflect nothing more than access to the target scale.

Threshold-flexible battery prediction and evaluation with incomplete lifetimes already have direct precedents. Guo et al.⁴ use criterion-adaptive supervision and a configurable state-of-health threshold for prediction with scarce end-of-life samples. Qi

et al.⁵ study absolute-error evaluation under right censoring and propose pseudo-observation-based MAE estimation. Our contribution is an empirical audit of endpoint definitions, censoring and matched label budgets in cross-source battery evaluation. We do not claim a new threshold-conditioned learning method or a new general solution to censored-error estimation.

Our study measures these three effects on data that are already public. Section 2 defines an endpoint contract and applies it to seven sources. Section 3 fixes the prediction protocol, including two historically frozen source-held-out records, and specifies a matched comparison of eight simple predictors. Section 4 reports the endpoint-policy sensitivity, the full-cohort evaluation for SNL, and an exhaustive allocation analysis for both external records. Section 5 turns the results into a short reporting checklist, and Section 6 states what the evidence cannot support.

Two clarifications about scope are needed at the outset. First, this is an evaluation and protocol study. The predictors used here are intentionally simple; their role is to hold the modelling constant while the reporting decisions change. Second, the two external records were produced at different times, with different training builds and different primary endpoints. They are two case studies, and the contrast between them is not a controlled experiment that isolates chemistry, protocol heterogeneity, or lifetime scale.

Figure 1 summarizes the information flow. Predictions use only the first 100 cycles, whereas endpoint auditing and retrospective scoring use the released trajectories. Completed target lifetimes enter only the declared calibration comparisons; the same calibration cells are removed from every comparator. The diagram separates prediction inputs from retrospective outcome information.

2 | DATA AND ENDPOINT CONTRACT

2.1 | Source registry and cell accounting

The registry contains seven sources: CALB, CALCE, HNEI, MICH_EXP, UL_PUR, SNL and HUST. It lists 220 physical cells. The processed curve table used for all analyses contains 214 cells and 318,707 cycle records; six registered MICH_EXP cells do not enter the processed table because their whole-record discharge-capacity range is below 0.5% of their maximum, i.e. they carry no measurable fade signal. Table 1 gives the per-source counts.

Table 1 distinguishes the registry count (220), the number of cells with curves (214), and the number of cells eligible for a particular task (for example 61 for SNL and 77 for HUST). Cycle records describe trajectory length and are not independent cell samples. HUST contains one chemistry, but not one cycling protocol: BatteryLife reports 77 different multistage discharge protocols with a common fast-charging protocol at 30°C¹ Appendix A.3.

SNL originates from the cycling study of Preger et al.⁶, and HUST from Ma et al.⁷. The processed and raw BatteryLife records and its source-specific processing code are identified separately^{8–10}.

CALB deserves a note. The released processed archive for this source stores at most 100 cycles per cell, while the official CALB lifetime labels lie between 104 and 1411 cycles. We therefore rebuilt CALB from the official raw capacity workbook distributed with BatteryLife Raw⁹. The rebuild is validated in two ways: of the 2,791 rows that also exist in the processed archive, 2,592 agree to 0.0 Ah, and 25 of the 27 official CALB labels are reproduced, with the two exceptions explained by documented source-specific handling (one extrapolated label and one skipped single-cycle anomaly). The processed table used here contains 40,481 CALB cycle records. An earlier parse of the same workbook reported 41,361 rows; the difference has not been reconciled inside this analysis, so we report only the 40,481 rows that the shipped table reproduces, and we do not attribute the difference to any cause.

2.2 | The retirement event as a contract

For cell i , let $Q_i(c)$ be the observed discharge-capacity series on the documented cycle coordinate. An endpoint contract specifies a reference capacity $Q_{i,\text{ref}}$, a threshold θ , a crossing rule and a preprocessing convention. The event is the first cycle at which the capacity falls to $\theta Q_{i,\text{ref}}$.

In the shipped tables, the policy named "nominal" uses an effective nominal reference: $Q_{\text{ref}} = Q_{\text{nameplate}}(\text{SOC}_{\text{high}} - \text{SOC}_{\text{low}})$. The first-observation policy uses the first retained discharge capacity. Unscaled nameplate capacity is a distinct convention; it coincides with the effective nominal reference for a full SOC window. All numerical results labelled "nominal" below use the

effective reference. The reference is a modelling choice, not a physical constant. If two references Q_A and Q_B are intended to describe the same absolute capacity, the thresholds must satisfy

$$\theta_B = \theta_A \frac{Q_A}{Q_B}. \quad (1)$$

Consequently, two tasks both called "80% lifetime" need not predict the same event. The nominal capacity, the first observed capacity, and the nominal capacity multiplied by the SOC-window width are all legitimate conventions, but they are different ones, and the difference is measurable.

Table 2 reports the paired comparison on identical curves at $\theta = 0.80$. For each source we give the number of observed events under the nominal reference and under the first-observation reference, the number of right-censored cells under each, the number of cells whose status changes, the number of cells observed under both policies, and the median paired ratio of the two lifetimes on those jointly observed cells.

The two extreme rows are the informative ones. For HUST, the released records end between 81.0% and 82.4% of the nominal reference; the 80% endpoint is therefore never observed under the nominal reference, while the first-observation reference places all 77 cells past their endpoint. For SNL, both references produce events, but 15 of 61 cells change status and the paired lifetime ratio on the 35 cells observed under both policies has a median of 1.06. The largest paired change in the registry occurs at $\theta = 0.90$ for MICH_EXP, where the median ratio is 5.48 over 12 jointly observed cells.

Table 2 is a statement about the *same* curves read under two definitions. It is not evidence that any published label is wrong, and it does not identify a physically correct reference. It does establish that a benchmark must declare which contract it uses, because the sample size, the event set and the target quantity all change with it.

2.3 | Observed, censored and extrapolated outcomes

Three outcome types must be distinguished. An **observed event** is a crossing present in the released trajectory. A **right-censored** cell has a valid record that ends before the threshold; its lifetime satisfies $T > L$, where L is the last observed cycle. An **extrapolated label** is produced by a source-specific regression beyond the record end. Only the first is an observation of the event.

This distinction has practical consequences in both external records. The official HUST label file provides 77 values, but all of them are extrapolations: no HUST record contains an 80% crossing, and the labels were produced by extending the final twenty cycles. We do not use them as observed events. In SNL, two of the 52 official labels are of this extrapolated kind and one further label differs from the recomputed crossing by 18 cycles because the official convention counts list positions rather than cycle numbers; our primary analysis uses recomputed crossings and treats non-crossing records as censored.

Because the two records use different primary endpoints, the comparison between them must be read carefully. SNL is evaluated at $\theta = 0.80$, where 50 events and 11 censored cells are available. HUST is evaluated at $\theta = 0.85$, where all 77 cells are observed. The 85% choice was fixed before any prediction error on HUST was computed, precisely because the 80% endpoint has no observed event in that source; it is a consequence of the label structure, not a selection made after seeing results.

2.4 | Prediction time and the two historical records

Both records predict at a cutoff of 100 cycles. Features are computed only from cycles up to the cutoff, and capacity normalisation uses the same reference rule in training and in the target source. The SNL record was frozen in the V62 build and trained on CALCE, HNEI, MICH_EXP and UL_PUR; the HUST record was frozen in the V64 build and additionally includes CALB. Each record carries a locked protocol file with the archive checksum, the hashes of the data tables and predictor code, the variance components used for calibration, the random seed, and a written statement of what had been inspected before the lock.

Label-level audits preceded both locks: the number of cells, the chemistry mix, the SOC windows and the event/censoring counts were examined before freezing. We therefore describe these as **source-held-out historical evaluations with a disclosed label audit**, not as third-party sealed blind tests. Everything done with those predictions afterwards, including the exhaustive allocation analysis of Section 4, is explicitly post hoc.

3 | PREDICTION, CALIBRATION AND EVALUATION

3.1 | Frozen predictor

The predictor is a ridge regression of log lifetime on three prefix features — the relative capacity at the cutoff, the per-cycle slope fitted over the prefix, and the curvature of the prefix fitted against cycle/cutoff — plus the threshold θ . One model is fitted per cutoff. The curvature feature is normalised by the cutoff, so pooling two cutoffs in one design matrix would mix units; the protocol never pools them.

The two records differ only in their declared training population and endpoint, and both are used exactly as frozen. No architecture search, hyper-parameter tuning, or retraining on target errors was performed for this paper. The fixed simple model lets us evaluate changes in reporting while keeping each historical prediction record unchanged. Robustness across model architectures is not tested here.

3.2 | Few-label calibration and matched controls

Let T_j be an observed lifetime and p_j the frozen predicted lifetime for calibration cell j . For a calibration set C of three completed target cells,

$$d_C = \text{median}_{j \in C} [\log T_j - \log p_j], \quad (2)$$

and the corrected prediction is

$$p_i^{(C)} = p_i \exp(w d_C), \quad w = \frac{\tau^2}{\tau^2 + \sigma^2/3}. \quad (3)$$

τ^2 and σ^2 are inherited from each source’s frozen protocol (SNL: 0.0706 and 0.0560, giving $w = 0.791$; HUST: 0.3590 and 0.1838, giving $w = 0.854$). Because the label statistic is a median rather than a Gaussian mean, we describe this as a variance-informed shrunk median correction or an empirical-Bayes-style log-offset, not as an exact posterior.

Seven comparators are evaluated on exactly the same allocations and the same held-out cells: no correction; the ordinary unshrunk log-median correction ($w = 1$); a fixed half correction; and four target-only constants computed from the same three labels — the arithmetic mean, the median, the geometric mean, and the constant that minimises calibration-sample MAPE. The target-only constants matter because a transferred model plus correction can only be credited with information beyond the target scale if it beats them.

Every method excludes the same three physical calibration cells from evaluation. The calibration cells are also the source of the labels; they are never scored.

3.3 | Exhaustive allocation sensitivity

Three labels can be chosen in $C(77,3) = 73,150$ ways for HUST and $C(50,3) = 19,600$ ways for SNL, giving 92,750 allocations in total. For each allocation we evaluate all eight methods on the remaining cells: 74 test events for HUST (the 77 observed cells minus the three calibration cells) and 47 test events plus the 11 censored cells for SNL.

The main summary is the arithmetic mean of the test MAPE over all allocations, with the allocation median and the 2.5th and 97.5th percentiles retained. These allocations overlap heavily: each cell appears in many allocations, and the 92,750 values are not independent experiments. The distribution describes label-selection variability within a fixed cohort. As a consistency check, the average zero-shot risk over all complements equals the full event-cohort zero-shot mean, because each event cell appears in the same number of test complements.

For SNL, calibration labels can only be drawn from cells that completed their lifetime; the analysis is therefore an event-conditional retrospective adaptation. It is not a simulation of a deployment in which three arbitrary cells are selected at cycle 100 and their final lifetimes are known immediately.

For descriptive visualization of allocation sensitivity, let $r_j = \ln(T_j/p_j)$. In addition to the existing median correction statistic $d_C = \text{median}_{j \in C}(r_j)$, define the unscaled dispersion $\text{MAD}_C = \text{median}_{j \in C} |r_j - d_C|$. No factor of 1.4826 is applied. The paired response is $\Delta_C = \text{MAPE}_{\text{shrunk}, C} - \text{MAPE}_{\text{zero}, C}$, in percentage points, on exactly the same held-out events. The residual map groups

all allocations into hexagons and colors each occupied bin by its mean Δ_C . Marginal histograms display allocation counts; a shared symmetric-logarithmic color scale has a linear region of ± 5 percentage points. The two source panels retain their own coordinate ranges. MAD describes disagreement among three labels and is not used by the correction. These retrospective summaries do not tune the predictor, establish a causal role for dispersion, or validate an allocation-selection rule.

3.4 | Full-cohort error identification under censoring

For a positive prediction p and a censored lifetime $T > L$, the absolute percentage error $APE = |p/T - 1|$ is bounded on the permitted range:

$$lo(p, L) = \max\left(0, 1 - \frac{p}{L}\right), \quad u(p, L) = \max\left(1, \left|\frac{p}{L} - 1\right|\right). \quad (4)$$

Observed events contribute their exact losses; censored cells contribute these bounds; averaging gives an identification range for the full fixed-cohort mean under the stated constraints. No independence-of-censoring assumption and no imputed lifetime is used. This differs from estimating a single censored-data MAE using pseudo-observations⁵: we report the MAPE values compatible with the observed bounds for this fixed cohort. The range is a property of this cohort and these predictions. It is not a confidence interval for a new source, and it is labelled as an identification range throughout.

Interval coverage is reported separately. If the upper endpoint of a prediction interval lies below a censored cell's observed survival bound, that interval definitely fails to cover the final lifetime; other censored cells may remain indeterminate. Coverage, interval width and the number of determinate failures are reported together.

3.5 | Computational verification

The two frozen records were reproduced from their archived code and data before this paper was written: the HUST zero-shot predictions were recomputed for all 462 queries and agreed to floating-point precision, and the SNL record was re-checked against its archived predictions. The exhaustive allocation tables are stored as compressed per-allocation records with an index, so that any single allocation can be reconstructed. The accompanying numerical audit checks cell isolation, matched comparisons, censoring bounds, allocation counts and sampled reconstruction of calibration predictions.

4 | RESULTS

4.1 | The reference policy changes what is observable

Figure 2 displays Table 2. Panel (a) shows the number of observed crossings at the same nominal threshold under the two references; panel (b) shows the paired lifetime ratio on the cells observed under both. The differences are not marginal: the HUST source moves from 0 to 77 events, SNL changes the status of 15 cells, and CALB's paired ratio is 2.49 on five jointly observed cells.

The practical implication is that the sample size of a "retirement" benchmark is itself a function of the endpoint contract. A study that reports 50 events for SNL and a study that reports 35 events for the same 61 cells are not reporting different measurement campaigns; they are answering different questions. Neither is wrong, but the two numbers cannot be used interchangeably, and a transfer experiment that mixes them will attribute a definitional difference to a model.

The same endpoint audit at all three thresholds is shown as a compact source-by-threshold heatmap in Supplementary Figure S1. It separates the net event-count difference from the total number of status changes, which can include changes in both directions.

4.2 | SNL: the event-only score and the full cohort

At a cutoff of 100 cycles and $\theta = 0.80$, the frozen SNL predictor has an event-only MAPE of 36.31% on 50 cells. Table 3 adds the 11 right-censored cells.

For the primary endpoint, the event-only score of 36.31% understates the cohort-level error by roughly nine percentage points: the identification range compatible with the observed survival information is 45.37–47.80%. Every one of the 11 censored cells has a prediction below its observed no-event cycle, which is precisely what drives the range upwards. The predictor is systematically short on long-lived cells.

The same cells change the coverage picture. The nominal 80% prediction interval covers 30 of the 50 observed events, i.e. 60.0% on the event set, but its upper endpoint lies below the observed survival bound for all 11 censored cells. Coverage over the full 61-cell cohort is therefore identifiable as $30/61 = 49.18\%$, not 60.0% (Figure 3). The median interval ratio is 2.20, so the intervals are also wide; the coverage statement must be read together with that width.

4.3 | HUST: comparison with matched target constants

At $\theta = 0.85$ all 77 HUST cells are observed. The historically specified three-label allocation (HUST_2-3, HUST_3-8, HUST_9-5) leaves 74 test cells. On those cells the zero-shot MAPE is 75.58%, the shrunk correction gives 12.77%, and the matched three-label geometric-mean constant gives 22.34%. An ordinary unshrunk correction gives 11.90%.

Table 4 places the fixed allocation inside the exhaustive analysis.

Figure 4 displays the individual predictions behind the historical fixed-allocation columns of Table 4. It retains the same held-out cells and distinguishes censored survival bounds from exact endpoints. These examples show where aggregate errors arise; the exhaustive distributions below measure sensitivity to the calibration-cell choice.

Three observations follow.

First, on HUST the shrunk correction improves on the matched zero-shot prediction in **every one** of the 73,150 allocations, with a mean paired difference of -59.39 percentage points (2.5th–97.5th percentile of the paired difference: -63.96 to -47.31). It also beats the matched geometric-mean constant on average by 4.00 percentage points and has lower error in 80.30% of allocations. This supports an incremental benefit of the calibrated transferred predictions over the matched constant within HUST. The full reduction from the uncalibrated score cannot be attributed to a new calibration algorithm or to transferable predictive skill alone.

Second, the ordinary unshrunk correction is *better* than the historically specified shrinkage on HUST (14.23% versus 16.25% in allocation mean). The unshrunk variant is a post hoc comparator, so its lower error does not make it a validated selection; it does show that the specific shrinkage rule is not the source of the gain. We therefore describe the result as a practical calibration opportunity, not as the superiority of an empirical-Bayes rule.

Third, fixing the three cells gives a more flattering number than the allocation average: 12.77% for the fixed allocation against 16.25% averaged over every allocation. Reporting only the fixed allocation would hide the sensitivity to which three cells are completed.

4.4 | SNL: an average can hide a harmful tail

On SNL the same family of correction behaves differently. Averaged over all 19,600 allocations it *increases* the event-only MAPE, from 36.31% to 48.50%, on the 47 test events remaining in each allocation. Yet the matched zero-shot score is improved in 58.78% of the allocations. The two statements are consistent because the distribution is asymmetric: many allocations produce a small improvement, while a minority produce very large deterioration. The 97.5th percentile of the corrected allocation distribution is 215.98%, and the median is 36.11%, i.e. slightly below the zero-shot value.

The 11 censored cells remain in the 58-cell test cohort for every allocation. Averaging the per-allocation full-cohort lower and upper bounds gives 54.37–58.27% for the shrunk correction, compared with 45.84–48.39% for its matched zero-shot comparator (Table S6). These use the same 58-cell complements; the 61-cell zero-shot range of 45.37–47.80% in Table 3 instead includes the three cells later used for calibration. Thus the deterioration is also visible when the censored cells are retained.

Figure 6 shows the event-only allocation distributions. The unshrunk correction is worse still (mean 60.57%, 97.5th percentile 358.54%), the half correction is comparatively benign (mean 38.25%, 97.5th percentile 91.94%), and all four target-only constants are far worse than the zero-shot predictions on average.

The contrast concerns these two fixed cohorts: the correction improves every HUST allocation relative to its matched zero-shot score, whereas SNL has a slightly lower allocation median, a higher allocation mean and a long harmful tail. The difference

cannot be attributed to the correction rule alone, because the two records differ in chemistry mix, cycling conditions, training build and primary endpoint. What can be said is that a single reported number, whether the mean or the median, is insufficient to describe the behaviour of a label-based adaptation on a heterogeneous cohort; the allocation distribution has to be shown.

Figure 7 places the same allocation records in the space of median calibration residual and residual MAD. The coordinates are computed from the three completed calibration lifetimes, while color uses the corresponding held-out event errors. HUST occupies a region of consistent paired improvement; SNL includes both improvement and deterioration. The ECDF in Supplementary Figure S2 shows the complete paired error-change distributions without treating allocations as independent trials. The map provides a descriptive diagnostic rather than a validated boundary between useful and harmful calibration.

4.5 | What is and is not established

Table 5 summarises the information each method uses and the corresponding result. The distinction is not cosmetic: the zero-shot row is the only one available before any target lifetime is observed.

Two boundaries follow. The HUST gain requires three completed target lifetimes; obtaining them may require a long calendar experiment, and the present analysis does not quantify that cost. The SNL full-cohort range is an evaluation quantity computed from observed event lifetimes and censoring bounds. It requires no additional calibration labels, is available only retrospectively, and does not supply a lifetime prediction.

5 | A PRACTICAL REPORTING PROTOCOL

The results support six checks that can be applied to any cross-source battery-lifetime study without new data.

1. **Declare the reference.** State the reference capacity and the SOC-window convention, and give the equivalence relation $\theta_B = \theta_A Q_A / Q_B$ when comparing thresholds. Report the number of cells whose status changes between the conventions used in the comparison (Table 2).
2. **Separate observed, censored and extrapolated outcomes.** Label extrapolated endpoints explicitly and never include them in the event-only denominator (Section 2.3).
3. **Report the cohort, not only the events.** Give the event-only score and the full-cohort identification range together, and state the number of determinate lower-bound violations (Table 3, Figure 3).
4. **Report interval width with coverage.** A coverage figure without the median interval ratio is uninterpretable; here the ratio is 2.20 and full-cohort coverage is 49.18%.
5. **Match every calibration comparator.** Use the same three labels, the same excluded calibration cells and the same test cells for the transferred model, the correction variants and the target-only constants (Table 4).
6. **Show the allocation sensitivity.** Report the mean, the median and the upper tail over the evaluated label allocations. A fixed allocation can give a more favourable result, as on HUST (Figure 5 and Figure 6). For censored cohorts, also report allocation-level full-cohort bounds (Table S6).

These checks are inexpensive: all of them can be computed from tables that an evaluation study already has.

6 | LIMITATIONS AND DISCUSSION

Two case studies, not a controlled contrast. SNL spans three chemistries and multiple cycling conditions; HUST uses one chemistry with 77 different multistage discharge protocols and common fast charging¹ Appendix A.3. HUST must not be interpreted as a single-protocol cohort. The records also differ in training build (CALB enters only HUST training) and primary endpoint (0.80 versus 0.85). The data do not isolate why calibration behaves differently, and do not establish chemistry, protocol diversity, source homogeneity or lifetime scale as its cause.

Both external sources are now available for development. Neither can serve as a fresh blinded test in a future revision. Any further use of them is a post-hoc analysis, and the locks in the source package are preserved to keep that boundary auditable.

Allocations are dependent. The 92,750 allocations reuse the same cells. Their percentiles describe label-selection variability in a fixed cohort; they are not confidence intervals and cannot be used as independent replications in a significance test.

The SNL adaptation is event-conditional. Its three calibration labels come from cells that completed their lifetime. A deployment that must choose three cells at cycle 100 and wait for their endpoints faces a selection problem that this analysis does not model.

Interval statements are empirical. The 80% intervals are built from development residuals. They are not conformal guarantees and carry no distribution-free coverage claim; the reported 49.18% is a finite-cohort count, not a population coverage probability.

Endpoint comparison is definitional. We show that different references can select different events; we do not identify a physically preferred reference. Threshold-adaptive supervision and configurable threshold inputs are already studied in⁴. We add no new threshold-conditioned predictor; the contribution is to make the endpoint reference and observed-event population explicit in evaluation.

Censored-error identification has a limited aim. Work such as⁵ addresses estimation and comparison of survival-model errors with censored outcomes. Our algebraic bounds instead display the MAPE uncertainty compatible with this cohort without imputing event times. They neither establish a generally superior survival metric nor resolve population-level model ranking when the ranges overlap.

Development accuracy depends on the build. On the current development table, leave-one-source-out error at the 80% endpoint and a 100-cycle cutoff is 27.39%. An earlier build of the same project reported 18.7%. The two builds differ in sample composition and preprocessing, so the change cannot be attributed to any single source, and we do not use it as evidence about CALB or about any other dataset.

Label cost is not quantified. "Three labels" means three completed target lifetimes, not three cycles of work. The value of an adaptive measurement policy is not established here.

Reproducibility boundaries. The SNL record belongs to an earlier data build; its tables are archived unchanged and its predictions are not recomputed on the current build. The HUST record matches the shipped tables bit-for-bit. Both are documented in the source package, together with the archive checksums and the frozen protocol files.

7 | CONCLUSION

A cross-source battery-lifetime number is only as interpretable as the endpoint contract behind it. On identical curves, the choice between a nominal and a first-observation capacity reference changes the number of observed events from 0 to 77 in one source and the status of 15 of 61 cells in another. On SNL, the event-only error of 36.31% corresponds to a full-cohort identification range of 45.37–47.80% once 11 censored cells are acknowledged, and the same intervals cover 49.18% of the full cohort rather than 60.0% of the events. On HUST, three completed labels reduce the allocation-averaged error from 75.64% to 16.25% and beat matched three-label constants on average, while the same operation on SNL increases the allocation-averaged error and produces a long harmful tail. The practical output is a six-item reporting contract and a matched comparison design that make these effects visible before a model is credited with transferable skill.

DECLARATIONS

Funding. This research received no specific grant from any funding agency in the public, commercial or not-for-profit sectors.

Conflict of interest. The authors declare no conflicts of interest.

Author contributions. Wenjie Li: conceptualization, data curation, formal analysis, software, and writing the original draft. Jufeng Li: methodology, supervision, validation, and reviewing and editing the manuscript. Tao Li: software, validation, visualization, and reviewing and editing the manuscript.

Data availability. The underlying public battery data are available through the BatteryLife processed and raw data records and its official code repository, as cited in this article. The accompanying Supporting Data and Code archive supplies the derived analysis tables, historical protocol records, frozen predictions, per-allocation records and scripts needed to verify the reported numbers. The upstream raw archives are not redistributed. The original source licences continue to apply. No new public repository DOI is claimed. Any later public deposit must identify the actual deposited version and its applicable licences.

AI assistance. OpenAI Codex assisted numerical consistency checks, correction of data-driven figures, manuscript revision and document preparation for this version. The supplied earlier project also records AI assistance with analysis-code development and manuscript drafting; its earlier tool versions were not retained. The framework is a conceptual schematic; the other figures are deterministic plots of the supplied data and are not generated experimental images. The authors reviewed the final content and take responsibility for it.

Ethics. Not applicable. This analysis uses previously published battery cycling data and involves no human participants or animal subjects.

References

1. Tan R, Hong W, Tang J, Lu X, Ma R, Zheng X, et al. BatteryLife: A comprehensive dataset and benchmark for battery life prediction. In: Proceedings of the 31st ACM SIGKDD Conference on Knowledge Discovery and Data Mining; 2025. Available from: <https://doi.org/10.1145/3711896.3737372>.
2. Severson KA, Attia PM, Jin N, Perkins NG, Jiang B, Yang Z, et al. Data-driven prediction of battery cycle life before capacity degradation. *Nature Energy*. 2019;4:383–391. Available from: <https://doi.org/10.1038/s41560-019-0356-8>.
3. Zhang H, Li Y, Zheng S, Lu Z, Gui X, Xu W, et al. Battery lifetime prediction across diverse ageing conditions with inter-cell deep learning. *Nature Machine Intelligence*. 2025;7:270–277. Available from: <https://doi.org/10.1038/s42256-024-00972-x>.
4. Guo R, Zhang K, Ke S, He S, Tian J, Shen W, et al. Enabling accurate battery remaining useful life prediction and through-life degradation mapping with scarce end-of-life samples. *IEEE Transactions on Industrial Informatics*. 2026;22(5):4492–4503. Available from: <https://doi.org/10.1109/TII.2026.3658872>.
5. Qi SA, Kumar N, Farrokh M, Sun W, Kuan LH, Ranganath R, et al. An effective meaningful way to evaluate survival models. In: Proceedings of the 40th International Conference on Machine Learning. vol. 202 of Proceedings of Machine Learning Research; 2023. p. 28244–28276.
6. Preger Y, Barkholtz HM, Fresquez AJ, Campbell DL, Juba BW, Roman-Kustas J, et al. Degradation of commercial lithium-ion cells as a function of chemistry and cycling conditions. *Journal of The Electrochemical Society*. 2020;167(12):120532. Available from: <https://doi.org/10.1149/1945-7111/abae37>.
7. Ma G, Xu S, Jiang B, Cheng C, Yang X, Shen Y, et al. Real-time personalized health status prediction of lithium-ion batteries using deep transfer learning. *Energy & Environmental Science*. 2022;15(10):4083–4094. Available from: <https://doi.org/10.1039/D2EE01676A>.
8. BatteryLife processed data. Zenodo. Version 12, record 21149533. Available from: <https://zenodo.org/records/21149533>.
9. BatteryLife raw data. Zenodo. Version 2, record 17960956. Available from: <https://zenodo.org/records/17960956>.
10. BatteryLife official code repository. GitHub repository. Accessed 25 October 2026. Available from: <https://github.com/Ruifeng-Tan/BatteryLife>.

TABLE 1 Source and cell ledger. Event and censoring counts use the 80% nominal reference. The primary SNL task uses cutoff 100 and threshold 0.80 (50 events, 11 censored cells); the primary HUST task uses cutoff 100 and threshold 0.85 (77 events, no censored cells).

Source	Role	Registered	Excluded	Processed	Cycle records	Events	Censored
CALB	development	27	0	27	40,481	9	18
CALCE	development	13	0	13	14,298	13	0
HNEI	development	14	0	14	15,155	14	0
MICH_EXP	development	18	6	12	4,006	11	1
UL_PUR	development	10	0	10	2,245	5	5
SNL	external record 1	61	0	61	96,402	50	11
HUST	external record 2	77	0	77	146,120	0	77
Total		220	6	214	318,707	102	112

TABLE 2 Endpoint policy on identical curves at $\theta = 0.80$.

Source	Cells	Events, nominal ref.	Events, first ref.	Censored, nominal	Censored, first	Status flips	Jointly observed	Paired lifetime ratio (first/nominal)
CALB	27	9	7	18	20	6	5	2.49
CALCE	13	13	13	0	0	0	13	1.00
HNEI	14	14	14	0	0	0	14	1.41
MICH_EXP	12	11	10	1	2	1	10	1.33
UL_PUR	10	5	4	5	6	1	4	1.00
SNL	61	50	35	11	26	15	35	1.06
HUST	77	0	77	77	0	77	0	–

TABLE 3 SNL event-only and full-cohort evaluation (historical frozen predictions, reanalysed).

Cutoff	θ	Cells	Events	Censored	Event MAPE (%)	Full-cohort MAPE range (%)	80% coverage	Censored lower-bound violations	Median interval ratio
100	0.80	61	50	11	36.31	45.37–47.80	0.492	11	2.20
100	0.85	58	52	6	33.34	39.22–40.24	0.655	6	2.20
100	0.90	31	31	0	54.09	54.09	0.355	0	2.20
50	0.80	61	50	11	52.25	58.75–60.86	0.328	11	2.90

TABLE 4 Calibration methods on HUST ($\theta = 0.85$, 74 test events) and SNL ($\theta = 0.80$, 47 test events). "All-allocation mean" averages the test MAPE over every three-label allocation (73,150 for HUST, 19,600 for SNL).

Method	HUST fixed (%)	HUST all-allocation mean (%)	HUST all-allocation median (%)	SNL fixed (%)	SNL all-allocation mean (%)	SNL all-allocation median (%)
No correction	75.58	75.64	75.64	34.63	36.31	36.36
Shrunk log-median (EB-style)	12.77	16.25	14.31	51.80	48.50	36.11
Unshrunk log-median	11.90	14.23	12.79	68.19	60.57	38.48
Fixed half correction	30.12	33.65	32.98	36.31	38.25	35.03
Three-label arithmetic mean	22.93	20.47	19.00	275.45	141.13	67.46
Three-label median	25.55	21.63	19.46	63.77	92.17	54.59
Three-label geometric mean	22.34	20.25	18.83	154.97	91.59	60.96
Three-label MAPE-optimal constant	25.55	21.69	19.46	63.77	59.67	53.57

TABLE 5 Information requirements and label budget.

Method family	Training lifetimes	Target prefix up to cycle 100	Completed target lifetimes	Result that it supports
Zero-shot transfer	from other sources	required	none	SNL 36.31% event-only, 45.37–47.80% full cohort; HUST 75.64%
Target-only constant	none	none	3	HUST 20.25–21.69% allocation mean; SNL 59.67–141.13%
Shrunk/unshrunk log correction	from other sources	required	3	HUST 14.23–16.25% allocation mean; SNL 48.50–60.57%
Full-cohort identification	from other sources	required	observed events for scoring	SNL 45.37–47.80%, 11 lower-bound violations

A framework for evaluating transferred battery lifetime predictions

A common data portfolio, two information paths, and matched evaluation

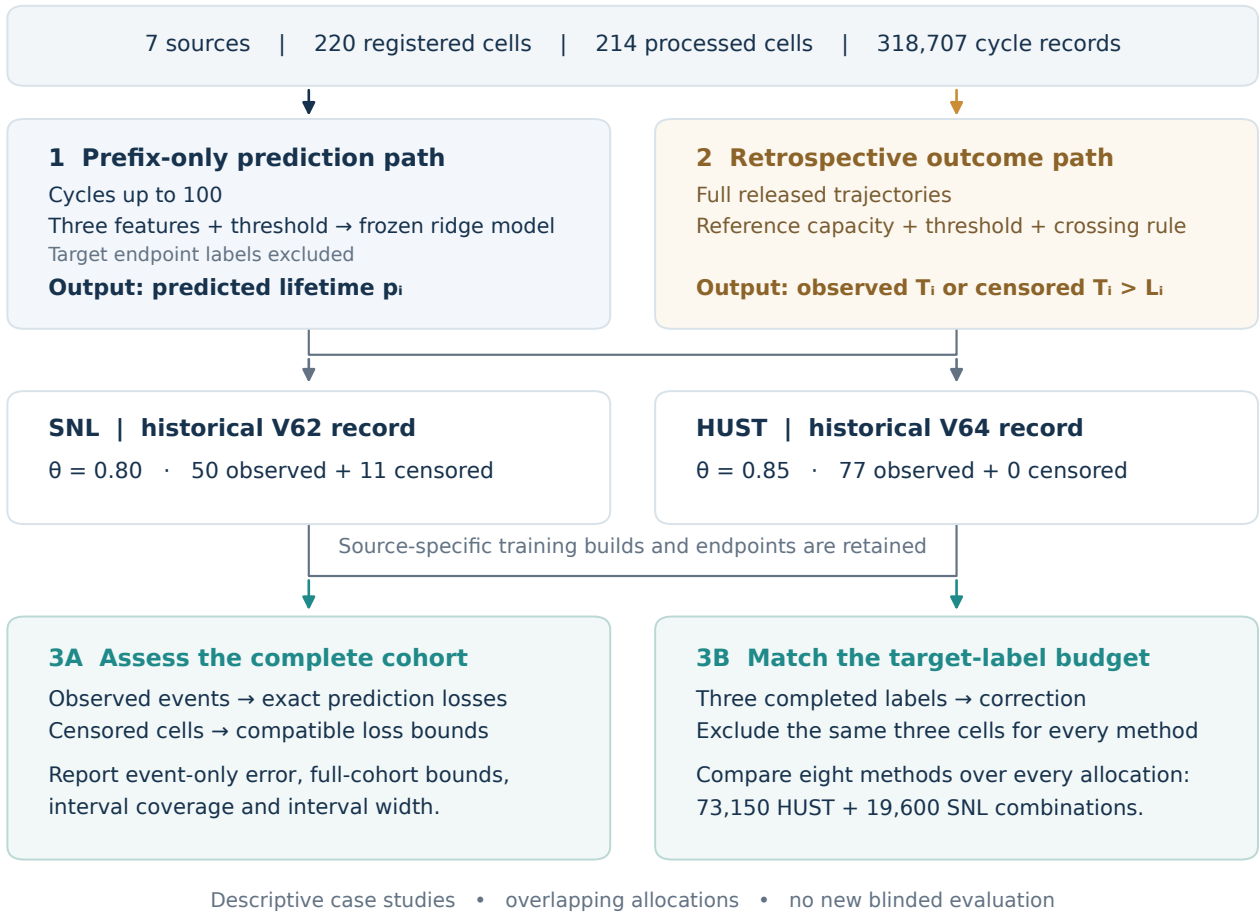


FIGURE 1 Information flow for the cross-source evaluation audit. Prefix data produce frozen zero-shot predictions; full released trajectories separately determine the endpoint and whether it is observed or censored. The two paths are paired within each source. Three completed target lifetimes enter only the retrospective adaptation comparisons, with the same calibration cells excluded from all comparators. Event-only errors, full-cohort identification bounds and allocation sensitivity are reported separately. The schematic describes historical analyses and does not depict a new learning architecture or a prospectively validated procedure.

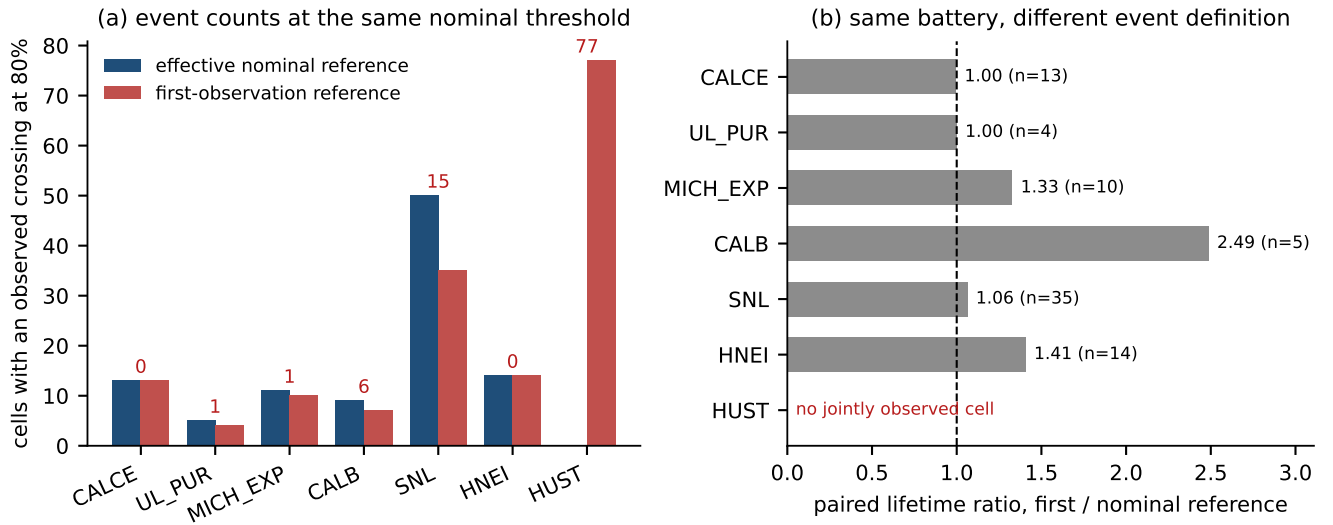


FIGURE 2 Endpoint policy on identical curves at $\theta = 0.80$. (a) Number of cells with an observed crossing under the nominal-capacity reference and under the first-observation reference; red labels give the number of cells whose event/censoring status changes. (b) Median paired lifetime ratio (first/nominal reference) on cells observed under both policies, with the number of such cells. HUST has no jointly observed cell.

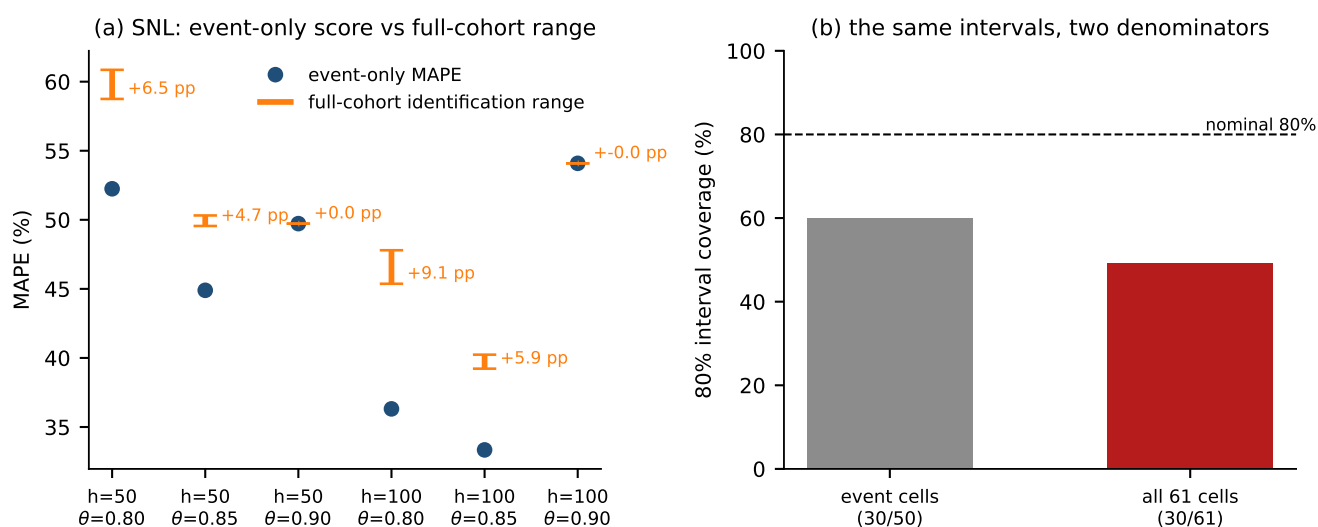


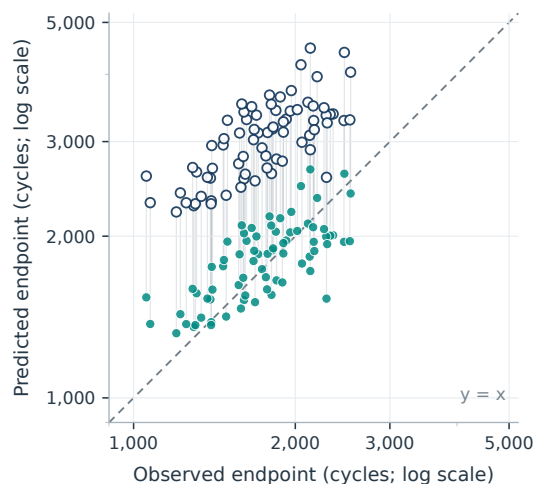
FIGURE 3 SNL full-cohort evaluation of the historical frozen predictions. (a) Event-only MAPE (dots) and the full-cohort identification range implied by the 11 censored cells (bars), at the stated cutoff and threshold; annotations give the width of the gap. (b) 80% interval coverage with the event denominators (30/50) and the full-cohort denominators (30/61).

Paired prediction diagnostics on held-out cells

Historical fixed three-label allocations | prediction cutoff: 100 cycles | no retraining or imputed endpoints

a HUST | 85% endpoint

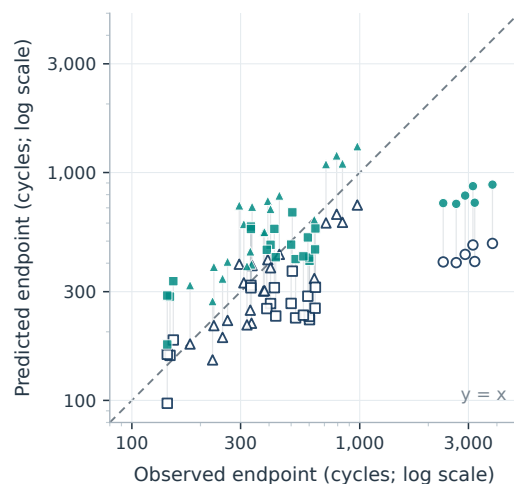
74 observed test cells | same 3 cells excluded
Event MAPE: 75.58% → 12.77%



○ Frozen zero-shot ● Fixed 3-label calibration — Same cell, paired predictions + Censored survival bound
○ LFP △ NCA □ NMC

b SNL | 80% endpoint

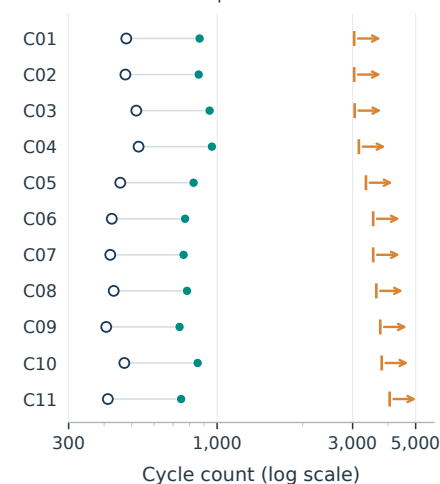
47 observed test cells | same 3 cells excluded
Event MAPE: 34.63% → 51.80%



Fixed-allocation scores are not allocation averages. Different source endpoints and training builds preclude a controlled source contrast.

c SNL censoring

11 test cells | all LFP
No observed endpoint to score



Orange: observed bound L; final $T > L$.
Arrow lengths are illustrative, not estimates.

FIGURE 4 Predictions for the historical fixed three-label allocations. (a) HUST: 74 observed test events at the 85% endpoint. (b) SNL: 47 observed test events at the 80% endpoint. Both use a 100-cycle cutoff. Hollow navy markers show zero-shot predictions; filled teal markers show the shrunk log-median correction. Fine vertical segments connect the same cell; shapes identify cathode chemistry. The dashed lines denote equality on source-specific logarithmic axes. (c) Eleven censored SNL test cells are displayed separately: orange ticks are observed survival bounds L, and right arrows mean $T > L$; their lengths are illustrative, not lifetime estimates. Calibration cells are excluded. The fixed-allocation event MAPEs are 75.58% and 12.77% for HUST, and 34.63% and 51.80% for SNL. These examples are not averages over all allocations.

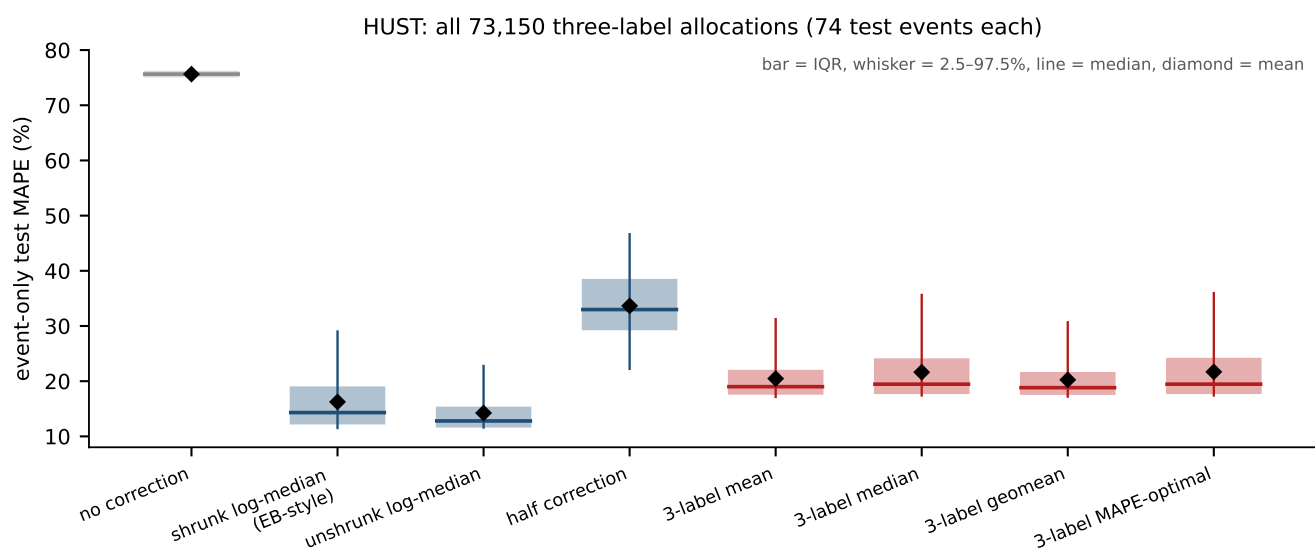


FIGURE 5 HUST: distribution of test MAPE over all 73,150 three-label allocations (74 test events each). Bars show the interquartile range, whiskers the 2.5th–97.5th percentile, horizontal lines the allocation median and diamonds the allocation mean.

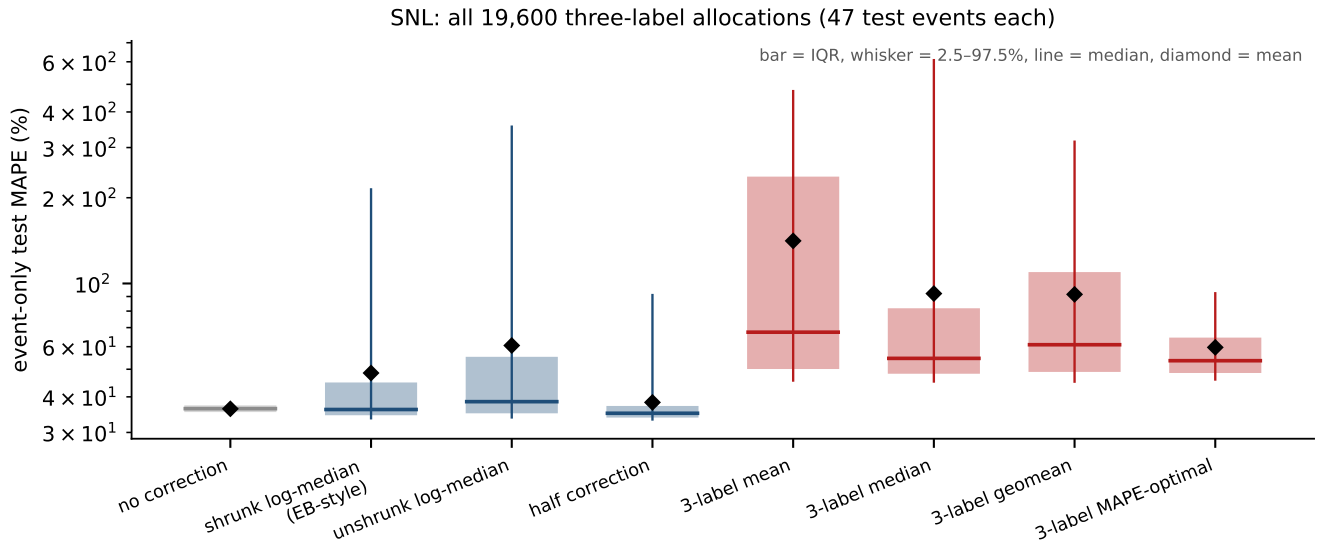


FIGURE 6 SNL: distribution of event-only test MAPE over all 19,600 three-label allocations (47 test events per allocation), on a logarithmic axis. The 11 censored cells do not enter this MAPE; full-cohort bounds for the 58-cell test complements are reported separately in Table S6. The corrected methods have medians close to the zero-shot value but long upper tails.

Where three-label calibration helps — and where it hurts

Every archived label allocation · same held-out cells for each paired comparison

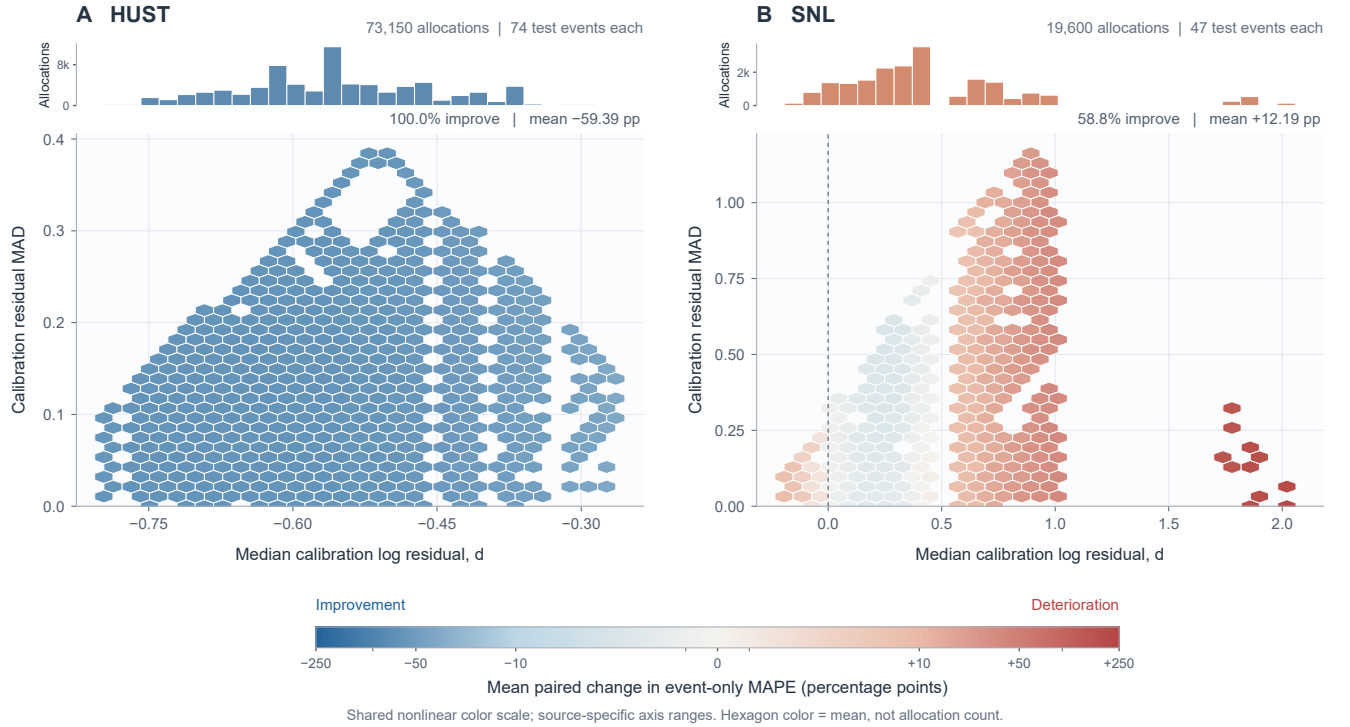


FIGURE 7 Retrospective calibration residual maps. All 73,150 HUST allocations and 19,600 SNL allocations are located by their median calibration log residual d and unscaled residual MAD. Color is the mean paired change in event-only test MAPE within each occupied hexagon: shrunk correction minus matched zero-shot, in percentage points. Blue denotes improvement and red deterioration; the shared color scale is symmetric-logarithmic, with a linear region of ± 5 percentage points. Top histograms show allocation counts. The panels have source-specific coordinate ranges. Each comparison excludes its three calibration cells and uses 74 HUST or 47 SNL test events. The eleven SNL censored cells do not enter the color metric. Empty bins are blank. The overlapping allocations are not independent replicates; MAD is not used by the correction or established as an independent risk factor.