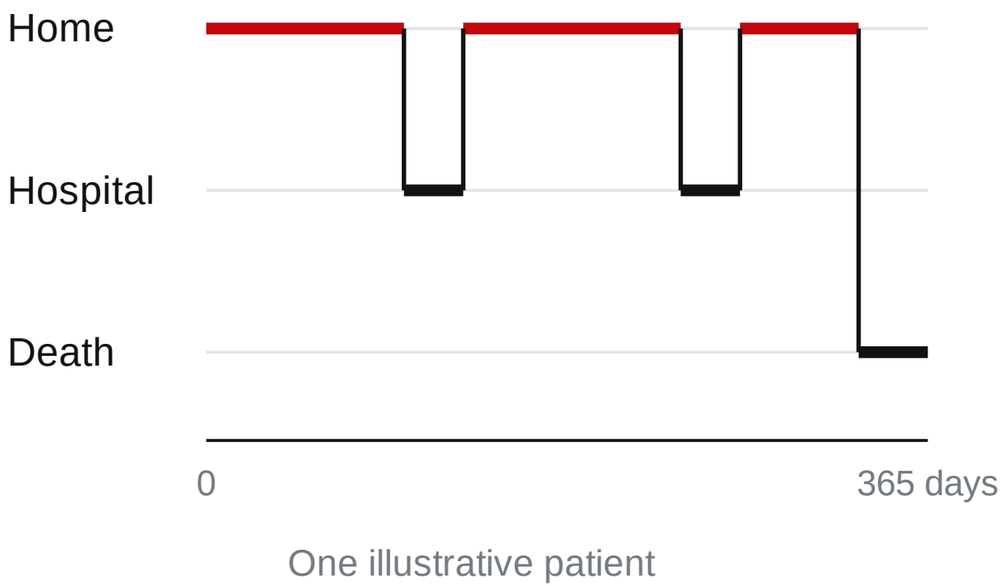
Conformal prediction for censored life history data

Lu Mao

Department of Biostatistics and Medical Informatics,
University of Wisconsin–Madison

Nov 17, 2026

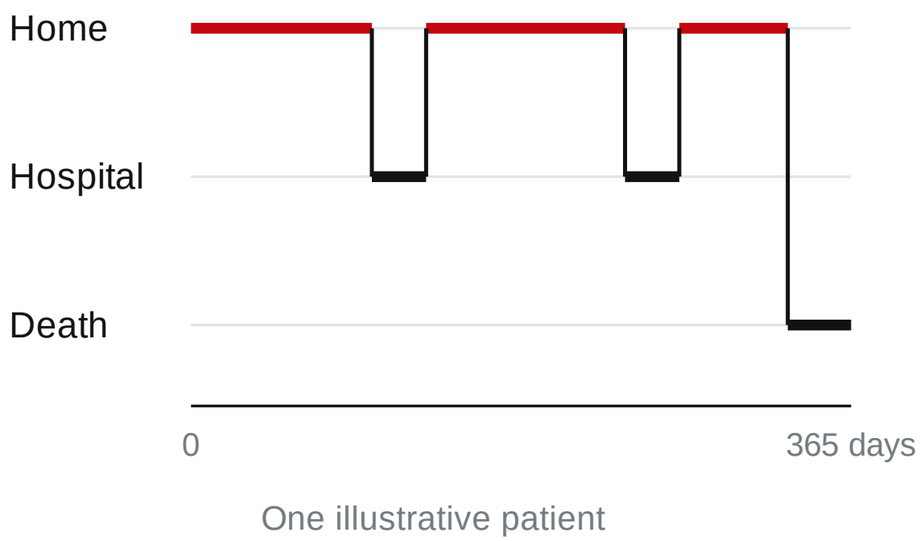
https://doi.org/10.48550/arXiv.2611.15724



One illustrative patient

Days at home after a heart failure visit

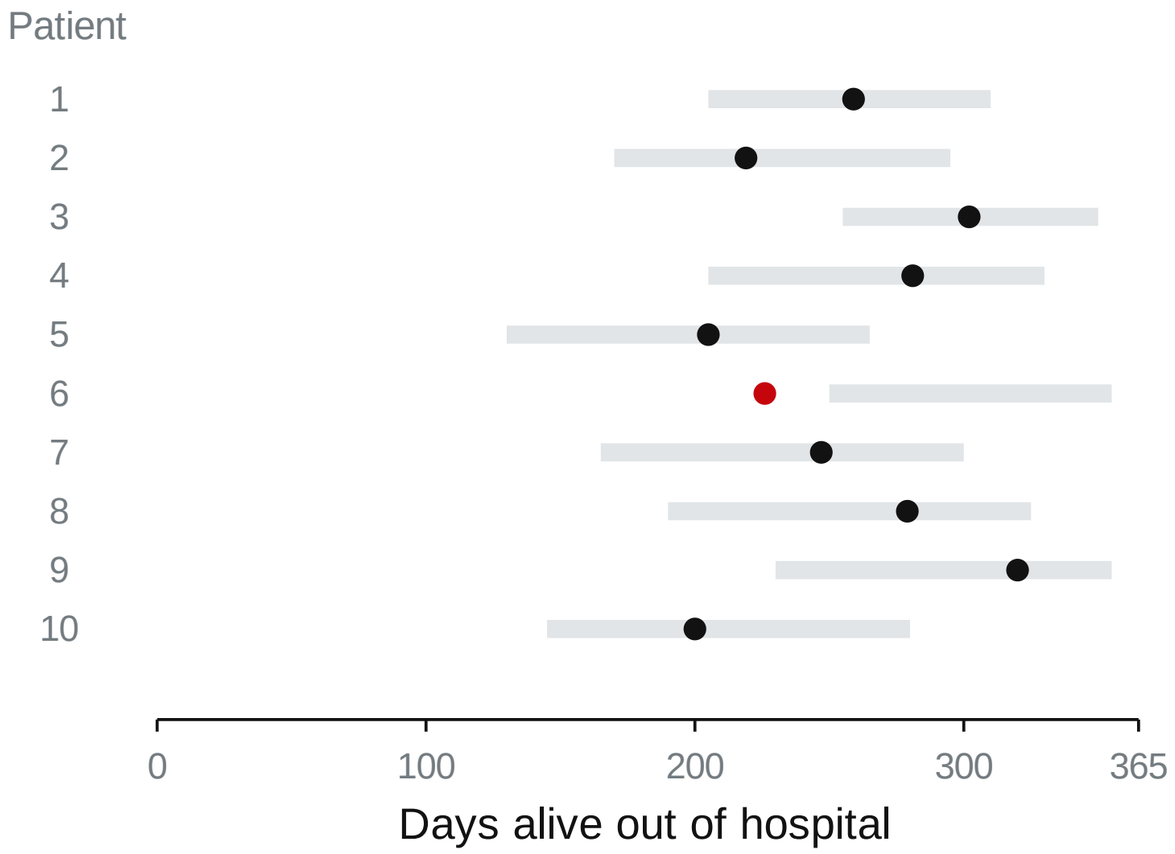
- **Patient outcome** — total days alive out of hospital over the next year
- **Trial summaries** — mean days at home, restricted mean survival time
- **Individual prediction** — a plausible range of home days for *a new patient* with a given baseline profile



One illustrative patient

Bühler, Cook & Lawless (2023), 2025 · Uno et al. (2014) · McCaw et al. (2026)

Coverage of a 90% prediction interval



Need a procedure for constructing prediction intervals that guarantee $1 - \alpha = 90\%$ coverage for new patients

- **Line** — one patient's baseline prediction interval
- **Dot** — home days observed during follow-up
- **Coverage** — probability of the outcome falling inside its interval

Shafer & Vovk (2008) · Angelopoulos, Barber & Bates (2026)

Split conformal prediction in general

1. **Training sample** — model for days at home from baseline characteristics
   - Supplies predicted mean $\hat{f}(z)$ for each covariate profile z
2. **Separate calibration sample**
   - (Y_i, Z_i) $(i = 1, \ldots, n)$ with observed outcome Y_i and covariates Z_i
   - **Nonconformity score** $S_i = s(Y_i, Z_i) = |Y_i - \hat{f}(Z_i)|$ — distance from the fitted prediction
3. **Calibrate threshold**
   - $q = S_{(k)}$, the kth smallest score, with $k = \lceil (n+1)(1-\alpha) \rceil$ and $S_{(n+1)} = \infty$
4. **New patient** (Y^*, Z^*) — all outcomes within q of the fitted prediction
   - Interval $\hat{\mathcal{C}}(Z^*) = [\hat{f}(Z^*) - q, \hat{f}(Z^*) + q]$ for unknown Y^*

> **Under exchangeability** — $\Pr\{Y^* \in \hat{\mathcal{C}}(Z^*)\} \geq 1 - \alpha$ (marginal over the training, calibration, and new data)

Lei et al. (2018)

Conformal prediction with censored event time

Earlier target — a *lower* prediction bound (identifiability)

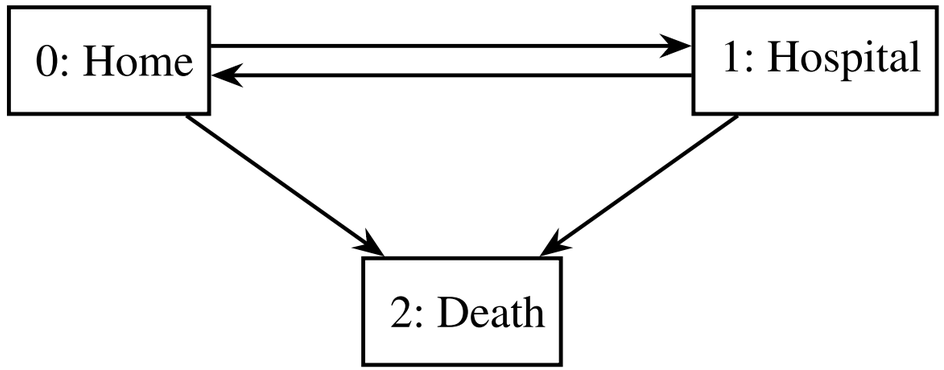
Study	Censoring information	Calibration method
Candès et al. (2023)	Known potential censoring times	Weighting patients with sufficient potential follow-up
Gui et al. (2024)	Same setting	Patient-specific follow-up cutoffs
Sesia & Svetnik (2025)	Censoring time hidden by an earlier event (death)	Imputation of the censoring time, followed by weighting
Farina et al. (2026)	Ordinary right censoring	Weighting and augmentation to estimate $F(q) = \Pr\{s(T_i, Z_i) \leq q\}$ and find cutoffs

Life histories: emulate Farina et al. (2026)

- Predict functionals of life history (state occupation times)
- Two-sided intervals over a fixed horizon

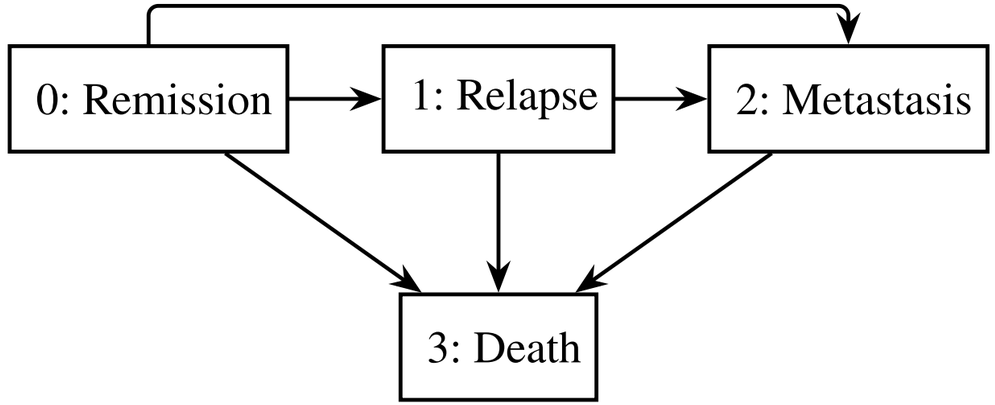
Two multistate models

A. Repeated hospitalisation



- Repeated admissions and discharges
- Time accumulated across all stays in each state

B. Cancer progression



- Metastasis after remission or relapse
- Death possible from every living state

From a life history to time in a state

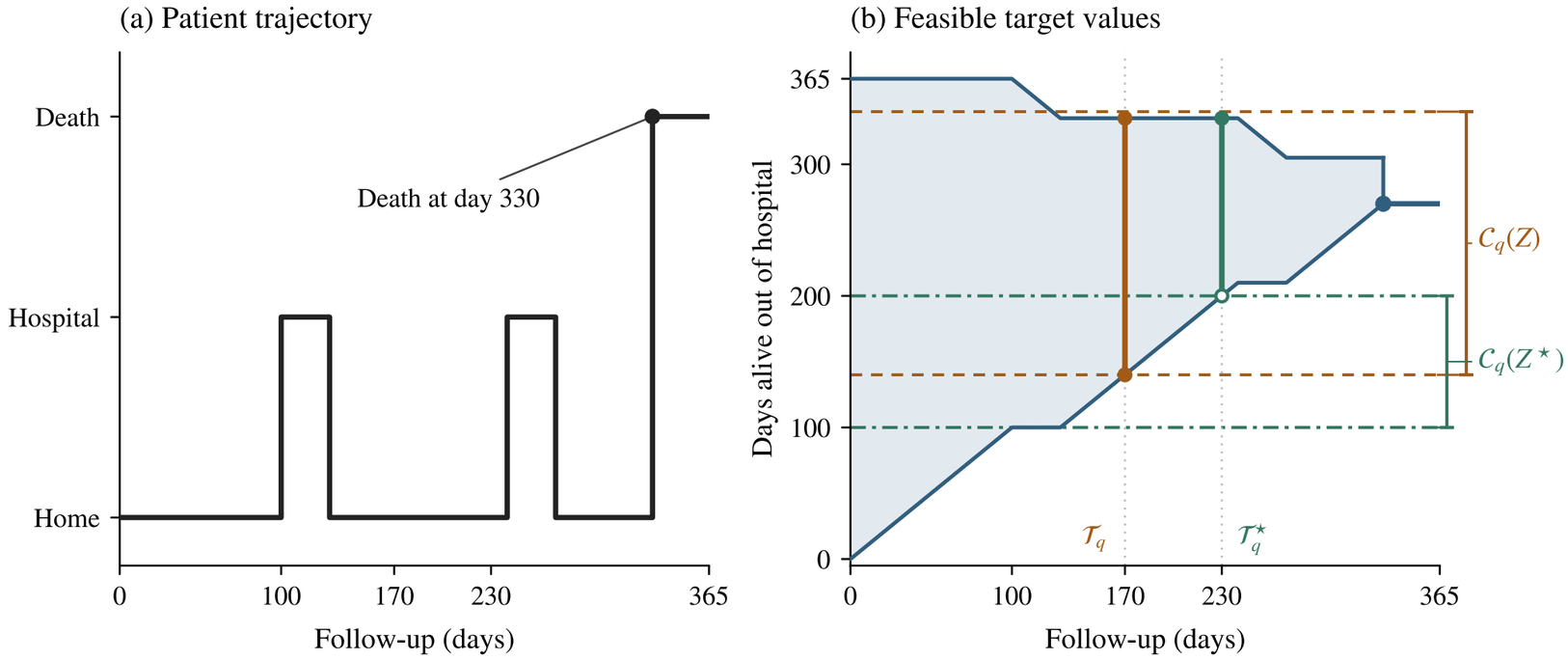
- **Patient** i — baseline characteristics Z_i, state $Y_i(t)$ at follow-up time t, clinical horizon τ
- **Life history** $\mathcal{H}_i(t) = \{Y_i(u) : 0 \leq u \leq t\}$ — all states and transitions through t
- **Target** $\Gamma_{\tau,i}$ — a scalar or vector summary of this history through τ

$$\Gamma_{\mathcal{A},i}(\tau) = \int_0^\tau I\{Y_i(t) \in \mathcal{A}\}\, dt$$

- $\mathcal{A}$ — selected states · $I\{\cdot\}$ — indicator, 1 when true and 0 otherwise
- **Heart failure** — days at home or in hospital by one year
- **Breast cancer** — time in initial remission by five years

Coverage can be determined before horizon

- As follow-up progresses, the feasible range of final home days shrinks



Candidate [140, 340] days

- Day 170 · every feasible final value inside

Candidate [100, 200] days

- Day 230 crossing · accrued home days passing 200

The coverage determining time

- **Candidate value** γ for a patient with baseline profile z
  - **Nonconformity score** $s(\gamma, z)$ — departure from the fitted prediction
- **Candidate region** $\mathcal{C}_q(z) = \{\gamma : s(\gamma, z) \leq q\}$ — all outcomes that conform closely enough to the fitted prediction
- For each calibration patient i, decide whether $\Gamma_{\tau,i} \in \mathcal{C}_q(Z_i)$ before the horizon τ
- **Feasible set** $\mathcal{K}_i(t)$ — all final target values from admissible continuations of patient i's history through t

> $$\mathcal{T}_{qi} = \inf\{t \in [0, \tau] : \mathcal{K}_i(t) \subseteq \mathcal{C}_q(Z_i) \text{ or } \mathcal{K}_i(t) \cap \mathcal{C}_q(Z_i) = \varnothing\}$$

First decisive time — all feasible values inside the region (coverage) or all outside (a miss)

- Coverage known even if the patient is censored afterward

Coverage and observability under censoring

Patient i in a calibration sample of size n · censoring time C_i · determining time $\mathcal{T}_{qi}$

Quantity	Definition	Interpretation
Coverage decision	$A_{qi} = I\{\Gamma_{\tau,i} \in \mathcal{C}_q(Z_i)\}$	1 = covered, 0 = missed
Decision observed	$R_{qi} = I\{C_i \geq \mathcal{T}_{qi}\}$	1 = known, 0 = unresolved
Observed duration	$X_{qi} = \min(C_i, \mathcal{T}_{qi})$	Earlier of the two times
Population coverage	$F(q) = \Pr\{\Gamma_{\tau,i} \in \mathcal{C}_q(Z_i)\} = E(A_{qi})$	Chance of coverage

Censoring on day 180 · $R_{qi} = 1,\ A_{qi} = 1$ for [140, 340] · $R_{qi} = 0$, A_{qi} still unknown for [100, 200]

Inverse probability weighting (IPW)

- **Censoring survival** $G_0(t \mid z) = \Pr(C_i \geq t \mid Z_i = z) \cdot \widehat{G}$ fitted on the independent training sample
- **Observation probability** $\Pr(R_{qi} = 1 \mid Z_i, \mathcal{H}_i(\tau)) = G_0(\mathcal{T}_{qi} \mid Z_i)$ under independent censoring given Z_i

Estimated population coverage at cutoff q in n calibration patients

$$\widehat{F}_{\mathrm{IPW}}(q) = \frac{1}{n} \sum_{i=1}^{n} \underbrace{\frac{R_{qi}}{\widehat{G}(\mathcal{T}_{qi} \mid Z_i)}}_{\text{inverse observation probability}} \underbrace{I\{\Gamma_{\tau,i} \in \mathcal{C}_q(Z_i)\}}_{\text{observed coverage decision } A_{qi}}$$

- **With true weights —** $E\{R_{qi} A_{qi} / G_0(\mathcal{T}_{qi} \mid Z_i) \mid Z_i, \mathcal{H}_i(\tau)\} = A_{qi}$
  - unbiased for the population coverage $F(q)$

Assumptions · $C_i \perp \mathcal{H}_i(\tau) \mid Z_i$ · censoring survival bounded away from zero through τ

Augmented inverse probability weighting (AIPW)

- **Conditional coverage** $Q_{0q,i}(t) = \Pr\{A_{qi} = 1 \mid Z_i, \mathcal{H}_i(t-)\}$ before determination
  - $\mathcal{H}_i(t-)$ the history strictly before t
- **Fitted coverage probability** $\widehat{m}_{qi}(t) = \widehat{Q}_{0q,i}(t)$ based on an outcome model (Markov, survival tree ensemble, etc.) trained on the independent training sample

$$\widehat{F}_{\text{AIPW}}(q) = \widehat{F}_{\text{IPW}}(q) + \frac{1}{n}\sum_{i=1}^{n}\left[\underbrace{\frac{(1-R_{qi})\,\widehat{m}_{qi}(X_{qi})}{\widehat{G}(X_{qi} \mid Z_i)}}_{\text{correction at observed censoring}} - \underbrace{\int_0^{X_{qi}} \frac{\widehat{m}_{qi}(t)}{\widehat{G}(t \mid Z_i)}\,\widehat{\lambda}_C(t \mid Z_i)\,dt}_{\text{expected correction during follow-up}}\right]$$

- $X_{qi} = \min(C_i, \mathcal{T}_{qi}) \cdot 1 - R_{qi}$ indicating censoring before determination
- $\widehat{\lambda}_C$ — fitted censoring hazard, with $\widehat{G}(t \mid z) = \exp\{-\int_0^t \widehat{\lambda}_C(u \mid z)\,du\}$

A cutoff for the prediction region

- **Joint score** for d outcomes, with candidate components γ_j, fitted means $\widehat{\mu}_j(z)$ and standard deviations $\widehat{\sigma}_j(z) > 0$

$$s(\gamma, z) = \max_{1 \le j \le d} \frac{|\gamma_j - \widehat{\mu}_j(z)|}{\widehat{\sigma}_j(z)}$$

- **Training sample —** fit a working Markov model or a survival-tree ensemble to estimate $\widehat{\mu}_j(z)$ and $\widehat{\sigma}_j(z)$ for each outcome.
- **Desired coverage** $1 - \alpha$ · calibration size n · adjusted rank $k = \lceil (n+1)(1-\alpha) \rceil$
- **First qualifying cutoff** and **region for a new profile** z

$$\hat{q} = \inf\{q : \widehat{F}_{\mathrm{AIPW}}(q) \ge k/n\}, \qquad \widehat{\mathcal{C}}(z) = \mathcal{C}_{\hat{q}}(z)$$

- Full feasible range if no finite cutoff qualifies or $k = n + 1$

Estimating $F(q)$ at fixed q

With probability at least $1-\delta$ over calibration samples

$$\left|\widehat{F}_{\mathrm{AIPW}}(q) - F(q)\right| \le \underbrace{e_n(\delta, c)}_{\text{sampling error of order } n^{-1/2}} + \underbrace{r_C r_Q / c}_{\text{model bias bound}}$$

- **Double robustness** — consistent for $F(q)$ if either the censoring or outcome model is correct
- **Efficiency** — minimum asymptotic variance among regular estimators when both fits are consistent and $r_C r_Q = o_p(n^{-1/2})$
- **Training score** s — any score fixed before calibration, even from a misspecified model
  - A better fit can yield narrower prediction intervals

r_C, r_Q — root-mean-square errors in the censoring hazard and conditional coverage through determination, maximised over cutoffs · $c > 0$ — lower bound on fitted censoring survival

Coverage for a new patient

- **Estimated score quantile** $\hat{q} = \inf\{q : \widehat{F}_{\mathrm{AIPW}}(q) \geq k/n\}$, with $k = \lceil (n+1)(1-\alpha) \rceil$
- **Region for a new patient** $\hat{\mathcal{C}}(Z^*) = \mathcal{C}_{\hat{q}}(Z^*)$, calibrated on n patients in $\mathcal{D}$
- **Given the training sample —** with probability at least $1 - \delta$ over calibration samples

$$\Pr\{\Gamma_\tau^* \in \hat{\mathcal{C}}(Z^*) \mid \mathcal{D}\} \geq 1 - \alpha - \underbrace{\left(\tfrac{2}{c} - 1\right)\left(\sqrt{\tfrac{\pi}{2n}} + \sqrt{\tfrac{\log(1/\delta)}{2n}}\right)}_{e_n(\delta, c)\ \text{(sampling error)}} - \underbrace{r_C r_Q / c}_{\text{model-bias allowance}}$$

- **More calibration patients —** smaller sampling allowance
- **Either nuisance fit consistent, other error bounded —** vanishing model bias
- **Limiting marginal coverage —** at least $1 - \alpha$, averaged over calibration samples and new patients, under the stated regularity conditions

Simulations of recurrent hospitalization

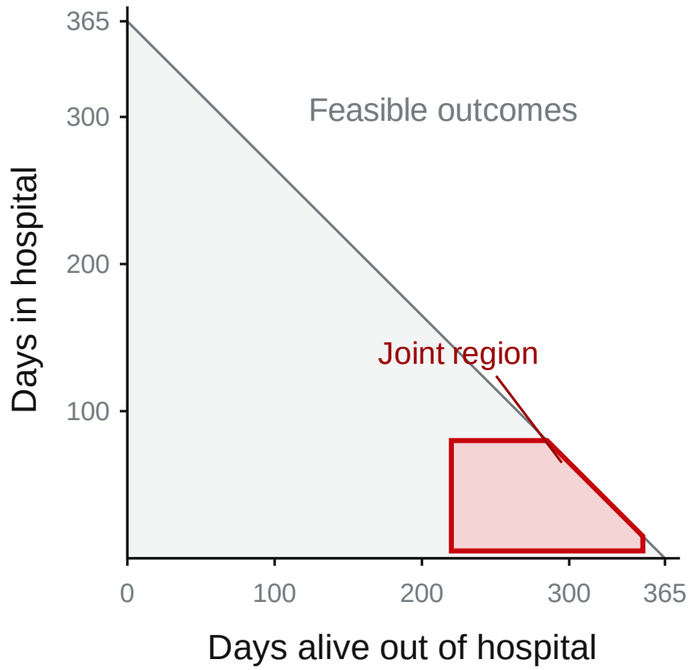
- **Prediction target** — one-year days alive and days at home, jointly
- **Design** — 60% censoring · 400 calibration patients · 90% joint coverage target

Calibration method	Joint coverage	Mean width for home days
Complete outcomes	89.9%	50.3 days
IPW at death or one year	89.9%	128.2 days
IPW at coverage determination	88.8%	75.0 days
AIPW at coverage determination	**89.8%**	**51.5 days**

With augmentation — home-time width within 1.2 days of the complete-outcome benchmark

HF-ACTION · days at home and in hospital

- **2,127 patients** with systolic heart failure and follow-up for hospitalisation and death
- **Joint target** — days at home and days in hospital during the year after randomisation
- **Feasible region** — nonnegative home and hospital days, sum at most 365
- **Working model** — Cox models for admission, discharge and death · 14 baseline predictors



Schematic joint region · O’Connor et al. (2009)

HF-ACTION · similar coverage with little censoring

- **Within one year** — 99 deaths · 79 patients (3.7%) with joint outcomes unresolved by censoring
- **Evaluation** — same held-out patients for all methods · 40 data splits · 90% target

Calibration method	Joint coverage	Home-time coverage	Mean home-time width
IPW at death or one year	89.8%	90.0%	28.9 days
IPW at coverage determination	89.9%	89.9%	28.8 days
AIPW at coverage determination	**90.4%**	**90.4%**	**29.3 days**

Separate home-time interval — about 29 days wide under all three methods

Held-out coverage adjusted for censoring

GBSG · five-year relapse-free time

German Breast Cancer Study Group · 686 women with node-positive breast cancer and complete baseline data

Five-year outcomes

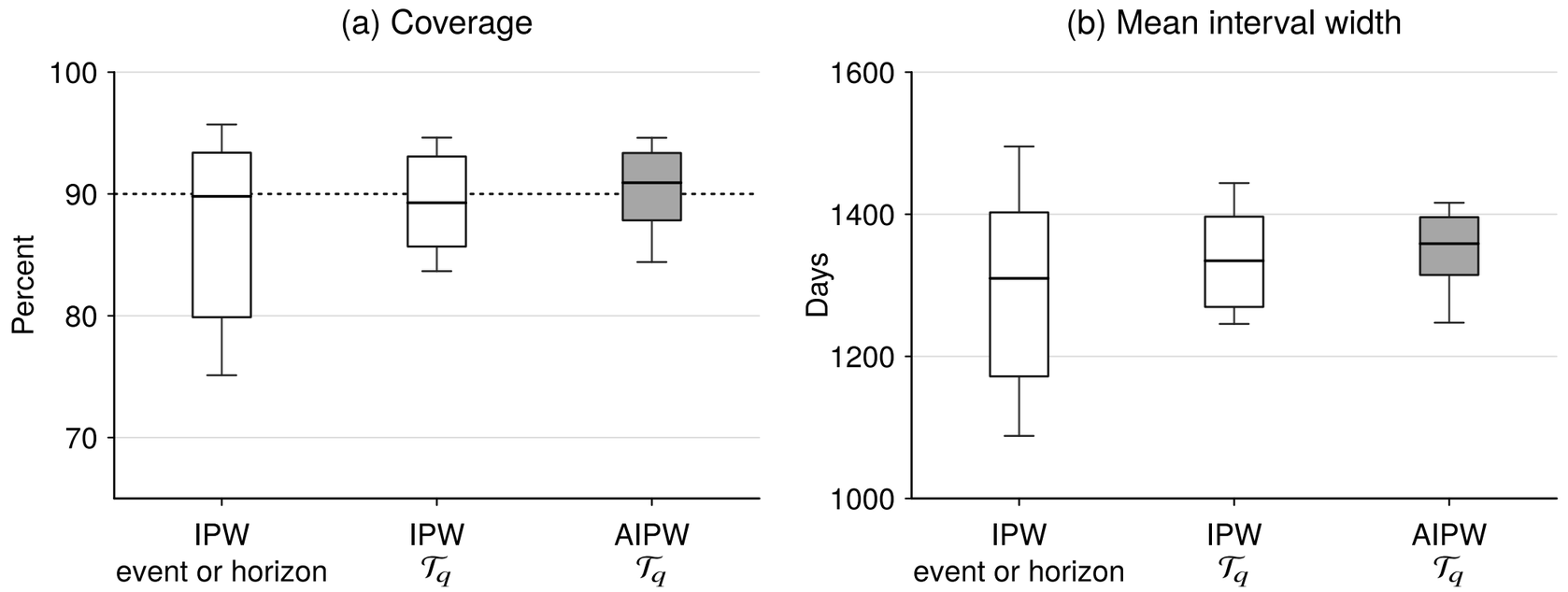
- 285 recurrences or deaths
- 123 women observed in remission through year five
- **278 (40.5%)** censored earlier in remission

Illustrative interval [2, 5] years

- Target — time to recurrence or death, capped at five years
- Remission through year two — every feasible target between two and five years
- Coverage known at year two, even with later censoring

Schumacher et al. (1994)

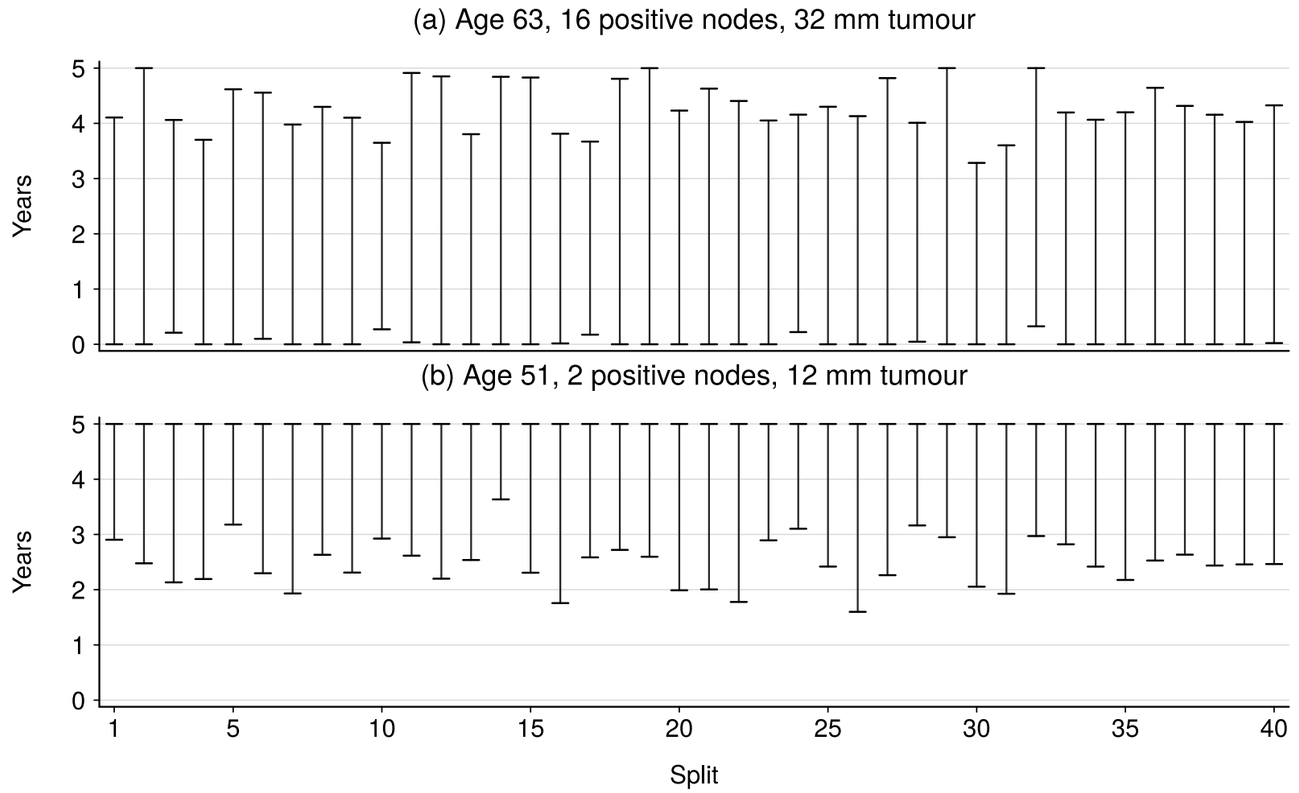
GBSG · estimated coverage and interval width



- **Mean coverage** — 86.8% with event-or-horizon IPW · 89.2% with determining-time IPW · **90.3% with AIPW**
- **Mean interval width** — less variation across data splits with determining-time calibration

90% target · 40 training–calibration–evaluation splits

GBSG · individual relapse-free time predictions



Age 63 · 16 positive nodes

- Upper endpoint below five years in **37 of 40** splits
- Very early recurrence or death within most intervals

Age 51 · 2 positive nodes

- Upper endpoint at five years in **all 40** splits
- Generally later lower endpoints

One 90% AIPW interval per bar · both women without hormonal therapy

Discussion

- **Known coverage before the final outcome** — earlier use of observed admissions, discharges and remission histories
- **Unresolved coverage** — history-based probabilities for the augmented estimator
- **Clinical evidence** — similar intervals in HF-ACTION with 3.7% censored · closer-to-nominal coverage in GBSG with 40.5% censored
- **Dynamic conformal prediction** — updated predictions at a return visit, using the history since baseline

References

Angelopoulos AN, Barber RF, Bates S (2026). Theoretical foundations of conformal prediction. Book manuscript, arXiv 2411.11824v5.

Bühler A, Cook RJ, Lawless JF (2023). Multistate models as a framework for estimand specification in clinical trials of complex processes. *Statistics in Medicine* 42, 1368–1397.

Bühler A, Cook RJ, Lawless JF (2025). Specification of estimands for complex disease processes using multistate models and utility functions. *Statistics in Medicine* 44, e70269.

Candès E, Lei L, Ren Z (2023). Conformalized survival analysis. *Journal of the Royal Statistical Society, Series B* 85, 24–45.

Farina R, Tchetgen Tchetgen EJ, Kuchibhotla AK (2026). Doubly robust and efficient calibration of prediction sets for right-censored time-to-event outcomes. *Biometrika* 113, asag036.

Gui Y, Hore R, Ren Z, Barber RF (2024). Conformalized survival analysis with adaptive cut-offs. *Biometrika* 111, 459–477.

Lei J, G'Sell M, Rinaldo A, Tibshirani RJ, Wasserman L (2018). Distribution-free predictive inference for regression. *Journal of the American Statistical Association* 113, 1094–1111.

McCaw ZR, Armbruster S, Tian L, et al. (2026). Using days alive and out of hospital as the study endpoint in cardiovascular heart failure clinical trials. *JACC: Heart Failure* 14, 103013.

O'Connor CM, Whellan DJ, Lee KL, et al. (2009). Efficacy and safety of exercise training in patients with chronic heart failure: HF-ACTION randomized controlled trial. *JAMA* 301, 1439–1450.

References

Schumacher M, Bastert G, Bojar H, et al. (1994). Randomized 2 × 2 trial evaluating hormonal treatment and the duration of chemotherapy in node-positive breast cancer patients. *Journal of Clinical Oncology* 12, 2086–2093.

Sesia M, Svetnik V (2025). Doubly robust conformalized survival analysis with right-censored data. *Proceedings of the 42nd International Conference on Machine Learning*, PMLR 267, 53989–54029.

Shafer G, Vovk V (2008). A tutorial on conformal prediction. *Journal of Machine Learning Research* 9, 371–421.

Uno H, Claggett B, Tian L, et al. (2014). Moving beyond the hazard ratio in quantifying the between-group difference in survival analysis. *Journal of Clinical Oncology* 32, 2380–2385.