Seeing Below the Limit of Detection: A Censored-Poisson Bayesian Latent-Growth Change-Point Detector (the Span Detector) for Serial ctDNA in HR+/HER2− Metastatic Breast Cancer

Abhijoy Sarkar* Span AI

Aarchi Singh Thakur† Span AI

June 10, 2026

Abstract

Circulating-tumour DNA (ctDNA) carries molecular evidence of acquired drug resistance months before it is visible on imaging, but the earliest evidence lives *below the assay's limit of detection* (LoD): a nascent resistant subclone is detected only intermittently, so a panel reports a flickering sequence of faint detects and *non-detects*. Commercial liquid-biopsy products treat each draw as an independent snapshot and treat a non-detect as "nothing there". We argue a non-detect is not nothing—it is a *left-censored* observation—and that the *pattern* of non-detects and faint detects over time carries actionable evidence of growth before any single value is trustworthy. We introduce **Span**, a censored-Poisson Bayesian latent-growth change-point detector that models the binary *detection process* itself, accumulates a sequential generalised-likelihood -ratio statistic for an upward change-point in the per-variant detection rate, and raises a competing-risks alarm with calibrated false-alarm control. Span is a transparent decision rule with no learned weights—there is nothing to overfit. On a mechanistically grounded synthetic cohort of HR+/HER2− metastatic breast cancer on first-line CDK4/6-inhibitor plus endocrine therapy, at a *matched* 10% false-alarm rate, Span roughly **doubles** the fraction of impending progressions caught at least three months ahead (indolent-emergence regime: $25 \pm 3\%$ vs $11 \pm 2\%$ for the commodity snapshot; five seeds, mean±std), and its advantage exhibits a clean, falsifiable dose-response: it is large when resistance emerges indolently (a long sub-LoD dwell) and *vanishes* when it emerges fast. A value-trajectory baseline (slope-CUSUM) performs identically to the snapshot, isolating the gain to the censored detection-process model. We separately validate that the survival machinery is sound on *real* breast-cancer data (GBSG-2, n=686), where a deep discrete-time competing-risks head matches a Cox baseline in concordance (0.67 vs 0.68). All ctDNA trajectories are clearly labelled synthetic; we make no clinical claim from simulated data, and we state precisely what would falsify the method.

*Corresponding author. abhijoy.sar@gmail.com

†Corresponding author. aarchisingh.t@gmail.com

Honesty boundary. No public serial-ctDNA breast-cancer cohort exists; that scarcity is not an accident but the precise data gap OncoTraj v1 [8] proved must be filled, and it is the problem this work motivates. Every ctDNA trajectory here is *synthetic* and generated by an openly released, biologically calibrated simulator (Sec. 3; code at `https://github.com/span-ai-labs/span-detector`). This paper demonstrates a *method*; it is not a clinical result, and no number computed from synthetic data should be read as a patient outcome. The single-patient lead time in Fig. 2 is an *illustration*, not a typical value; the quantitative claims are the cohort statistics of Sec. 5.

1 Introduction

HR+/HER2− disease is the largest biomarker-defined solid-tumour population, and first-line CDK4/6-inhibitor plus endocrine therapy is its standard of care. Acquired resistance arises through *competing* molecular mechanisms— *ESR1* mutations (ligand-independent ER signalling) in roughly a third of progressors, with *PIK3CA*, *RB1* loss, and *HER2*-activating mutations among the rest [1]. The PADA-1 trial showed that acting on a *rising ESR1 ctDNA signal before* radiographic progression— switching endocrine partner—roughly doubled progression-free survival [2], direct evidence that ctDNA leads imaging by a clinically actionable window. Across solid tumours, ctDNA leads imaging by months (TRACERx NSCLC: a median of ∼151 days [3]).

The snapshot problem. Commercial liquid biopsies report each draw independently: *is a resistance variant detectable now?* The earliest and most valuable evidence, however, is precisely the evidence a snapshot must discard. A nascent subclone sheds little DNA; under finite-molecule (Poisson) sampling its variant is detected only *intermittently*, producing a flickering series of non-detects and faint, near-LoD detects long before its allele fraction (VAF) climbs to a level any single test will trust. A snapshot, and even a slope computed on the *reported* VAF, is blind during this sub-LoD dwell because it treats each non-detect as a hard zero.

Our claim. A non-detect is a *left-censored* measurement: it says the subclone is *smaller than the LoD*, not absent. The temporal *pattern* of non-detects, flickering detects, and their rising rate is informative about whether a clone has begun to grow—*before* its level is reliably callable. We turn this into a detector, **Span** (Censored-Poisson Bayesian Latent-Growth change-point), that (i) models the binary detection process under Poisson sampling and LoD censoring, (ii) tests sequentially for an upward change-point in the per-variant detection rate via a generalised likelihood ratio (GLR), (iii) aggregates across competing mechanisms, and (iv) fires an alarm at a threshold calibrated to a target false-alarm rate. Span has *no trainable parameters*; its advantage is structural, not learned.

Positioning: a method layer atop OncoTraj. This work is the second piece of a programme. OncoTraj v1 [8] established that single-snapshot tissue NGS hits a ceiling on resistance prediction in EGFR-mutant NSCLC: mechanism classification fails *structurally*, because the dominant on-target resistance mutation (C797S) is rarely present in the baseline tissue specimen and only becomes visible in *serial ctDNA* as it emerges under therapy. That negative result is the motivation for everything here. The Span Detector is the method layer designed for exactly the serial-ctDNA cohort that constraint demands—it takes the censored, longitudinal liquid-biopsy stream OncoTraj v1 proved is necessary, and extracts the early-warning signal from it. We develop and validate Span on HR+/HER2− breast cancer because its competing-mechanism resistance biology (ESR1/PIK3CA/RB1/HER2) makes the censored-detection problem sharpest, but the detector is mechanism- and tumour-agnostic by construction.

Contributions.

1. **A novel detector for censored ctDNA dynamics** (Sec. 4): to our knowledge the first to phrase early resistance detection as a sequential change-point test on the *Bernoulli detection process* of a serial liquid biopsy under explicit Poisson/LoD censoring, with competing-risks aggregation and matched false-alarm control.
2. **A transformative, falsifiable result on synthetic data** (Sec. 5): at a matched 10% false-alarm rate, Span roughly doubles early ($\geq$12-week) detection in the indolent regime, with a monotone dose-response in emergence speed that *vanishes* for fast clones—the signature of a real mechanism rather than an artifact.
3. **Honest ablation and real-data grounding**: a value-trajectory rule (slope-CUSUM) matches the snapshot exactly, localising the gain to the censored detection model; and the survival backbone matches Cox on real breast data (GBSG-2, Sec. 6).

2 Related work and positioning

Our components are individually classical; the *synthesis* for censored ctDNA detection is, to our knowledge, new. Sequential change detection descends from the CUSUM [5] and the generalised-likelihood-ratio / window schemes of Lorden [6]; we use a GLR for an unknown change-point but place it on a *Bernoulli detection* likelihood rather than a Gaussian mean. Treating below-threshold readouts as left-censored is the Tobit idea [7]; we apply it to the *detection indicator* rather than the magnitude, which is what makes pre-detection evidence accessible. Competing-risks survival with neural likelihoods follows DeepHit [4]; we borrow its discrete-time competing-risks head only for the real-data backbone (Sec. 6). Prior ctDNA-monitoring work largely thresholds a value or a value trend; the distinguishing move here is to model the censoring mechanism generatively and to detect change in the *rate of detection* under Poisson sampling. The point of the synthesis is operational: it yields a decision rule that provably accumulates evidence while the signal is censored, with a knob (the GLR threshold) that maps directly onto a clinician's false-alarm budget.

3 Synthetic cohort and generative model

We simulate N patients on first-line CDK4/6i + endocrine therapy. Static covariates (age, menopausal status, ER level, visceral disease, baseline tumour fraction) follow population-plausible distributions. With probability 0.85 a patient acquires resistance through one competing mechanism $k \in \{\text{ESR1, PIK3CA, RB1, HER2}\}$ with published-approximate shares; the remainder are right-censored.

Latent subclone. Mechanism k emerges at onset τ_k and grows logistically to a (low) ceiling, $f_k(t) = c_k\,\sigma\big(g_k(t-\tau_k) - 4\big)$ for $t \geq \tau_k$, with growth rate g_k controlling how long the clone *dwells* near/below the LoD. The *emergence regime* (indolent / slow / fast) is exactly the range of g_k.

Censored Poisson measurement. At draw time t_i the reported VAF for variant k is generated by a finite-molecule detection model: detection probability $\pi(t_i) = 1 - \exp(-f_k(t_i)/\kappa)$, and *conditional* on detection the value is $\max(f_k + |\varepsilon|, \text{LoD})$ with assay noise ε; otherwise the draw is a **non-detect** (reported 0, a left-censored observation). Sporadic spurious detections occur at a low rate. Molecular "crossing" t^{mol} is when f_k first exceeds the LoD; imaging progression follows after a log-normal lag (median $\sim$22 weeks, matching the observed ctDNA-to-imaging window [3]). Crucially

the model produces the *flicker* regime—intermittent detection at low VAF—that motivates reading the detection *pattern* rather than any value.

4 Method: Span

Fix a mechanism k and a landmark (current draw) j; let $t_{1:j}$ be the draw times and $d_{1:j} \in \{0,1\}$ the detection indicators ($d_i = \mathbf{1}[\mathrm{VAF}_i > 0]$).

Two hypotheses on the detection process. Under the quiescent null the detection rate is a constant background π_0 (spurious positivity):

$$\mathbf{H}_0: \quad d_i \sim \mathrm{Bernoulli}(\pi_0).$$

Under the alternative a change-point at onset τ starts an upward drift in the log-odds of detection at rate $r > 0$:

$$\mathbf{H}_1(\tau, r): \quad d_i \sim \mathrm{Bernoulli}(\pi_i), \qquad \mathrm{logit}\,\pi_i = \mathrm{logit}\,\pi_0 + r\,(t_i - \tau)_+.$$

This is the censored-growth signature: as the latent clone grows, it is detected ever more often, so d_i switches from a sparse to a dense regime—*even while every reported value remains a near-LoD flicker.*

Sequential GLR statistic. With Bernoulli log-likelihood $\ell(d_{1:j}; \pi_{1:j}) = \sum_{i \le j} d_i \log \pi_i + (1 - d_i)\log(1 - \pi_i)$, the generalised-likelihood-ratio for an upward change by draw j is

$$\Lambda_j^k \;=\; 2 \max_{\substack{\tau \in \{t_1, \ldots, t_{j-1}\} \\ r \in \mathcal{R},\, r > 0}} \Big[\ell\big(d_{1:j};\, \pi(\tau, r)\big) - \ell\big(d_{1:j};\, \pi_0\big)\Big]. \tag{1}$$

The inner maximisation profiles over the unknown onset τ (each past draw) and a small grid $\mathcal{R}$ of rise rates; the one-sided constraint $r > 0$ encodes that only *growth* should raise an alarm. Λ_j^k is monotone in the evidence that mechanism k has begun to grow.

Competing-risks aggregation and the alarm. We do not know in advance which mechanism will win, so the decision-relevant quantity is whether *any* subclone is growing. We aggregate across mechanisms and declare the patient's alarm time at the first landmark whose pooled statistic crosses a threshold h:

$$S_j = \max_k \Lambda_j^k, \qquad A = \min\{\, t_j :\; S_j \ge h \,\}. \tag{2}$$

Calibrated false-alarm control. The threshold h is *not* taken from an asymptotic null (the max over τ and competing mechanisms breaks the textbook $\frac{1}{2}\chi_1^2$ mixture). Instead we calibrate it directly to a clinical budget: on a held-out set of *non-progressors*, h is the smallest value such that at most a fraction α ever fire, $h = \inf\{h' : \Pr_{\text{non-prog}}(\max_j S_j \ge h') \le \alpha\}$. With $\alpha = 0.10$ every detector in our comparison operates at the same 10% false-alarm rate, making the race apples-to-apples. Span is thus a pure decision rule: **no weights are trained**, so a favourable result cannot be an overfit.

Baselines under the same budget. The *commodity snapshot* scores each landmark by the latest maximum reported VAF (a single-test rule). The *slope-CUSUM* is a strong value-trajectory rule: a one-sided Page CUSUM on the Tobit-imputed VAF (non-detect $\to \frac{1}{2}$LoD), which can catch a rising *value* but not growth hidden in a run of non-detects. Both are calibrated to the same α.

Table 1: Span roughly doubles early detection of impending resistance at a matched 10% false-alarm rate, with a dose-response in emergence speed. Mean ± std over five seeds; all data synthetic. "Early" = flagged ≥12 weeks before imaging; "pre-LoD" = flagged before the variant is callable. The slope-CUSUM tracks the snapshot, isolating the gain to the censored detection-process model.

Emergence regime	Detector	Overall sens.	Early sens. (≥12 wk)	Pre-LoD sens.
indolent (long sub-LoD dwell)	commodity snapshot	$15 \pm 1\%$	$11 \pm 2\%$	$8 \pm 2\%$
	slope-CUSUM	$15 \pm 1\%$	$11 \pm 2\%$	$7 \pm 2\%$
	Span (ours)	$\mathbf{34} \pm \mathbf{4}\%$	$\mathbf{25} \pm \mathbf{3}\%$	$\mathbf{14} \pm \mathbf{2}\%$
slow	commodity snapshot	$27 \pm 3\%$	$14 \pm 3\%$	$5 \pm 2\%$
	slope-CUSUM	$27 \pm 3\%$	$13 \pm 3\%$	$5 \pm 2\%$
	Span (ours)	$\mathbf{44} \pm \mathbf{6}\%$	$\mathbf{27} \pm \mathbf{4}\%$	$\mathbf{11} \pm \mathbf{3}\%$
fast (short sub-LoD dwell)	commodity snapshot	$46 \pm 4\%$	$22 \pm 3\%$	$6 \pm 3\%$
	slope-CUSUM	$46 \pm 4\%$	$22 \pm 3\%$	$5 \pm 3\%$
	Span (ours)	$47 \pm 4\%$	$27 \pm 2\%$	$9 \pm 2\%$

5 Experiments

Why early sensitivity, not median lead. A median "lead time among flagged patients" is treacherous: a timid detector that only flags the few largest, latest-imaging clones can report a deceptively *long* median lead while missing most patients. We therefore report a *denominator-fair* metric over *all* progressors: **early sensitivity**, the fraction flagged at least 12 weeks (one horizon bin $\times 1.5$) ahead of imaging at a matched 10% false-alarm rate; and **pre-LoD sensitivity**, the fraction flagged before the driver variant crosses the LoD at all—a window the snapshot cannot access by construction. We also report overall sensitivity (fraction flagged). Every cell is the mean±std over five independent seeds (n=500 patients each, 50/50 calibrate/test split).

Headline result. Table 1 and Fig. 1 give the main finding. In the **indolent** regime—the clinically hard setting where resistance emerges gradually and dwells near the LoD for months—Span flags $34 \pm 4\%$ of impending progressions versus $15 \pm 1\%$ for the commodity snapshot ($2.3\times$, non-overlapping error bars), and catches $25 \pm 3\%$ at least three months ahead versus $11 \pm 2\%$ ($2.3\times$). It also fires *before the variant is callable* in $14 \pm 2\%$ of progressors, against $8 \pm 2\%$ for the snapshot (whose pre-LoD "hits" are mostly spurious positives, not genuine sub-LoD reads).

The dose-response *is* the evidence. Critically, the advantage is not uniform: it is largest for indolent emergence, shrinks for slow, and collapses to a tie for **fast** emergence ($47 \pm 4\%$ vs $46 \pm 4\%$ overall). This is exactly what the mechanism predicts— the advantage *is* the sub-LoD dwell time, so a clone that rockets through the LoD leaves nothing for a detection-pattern test to exploit that a snapshot misses. A monotone, mechanistically predicted, vanishing-where-it-should dose-response is far harder to dismiss as an artifact than a single favourable number. The slope-CUSUM equalling the snapshot in every regime is the internal control: reading the *value* trajectory adds nothing; only modelling the *censored detection process* helps.

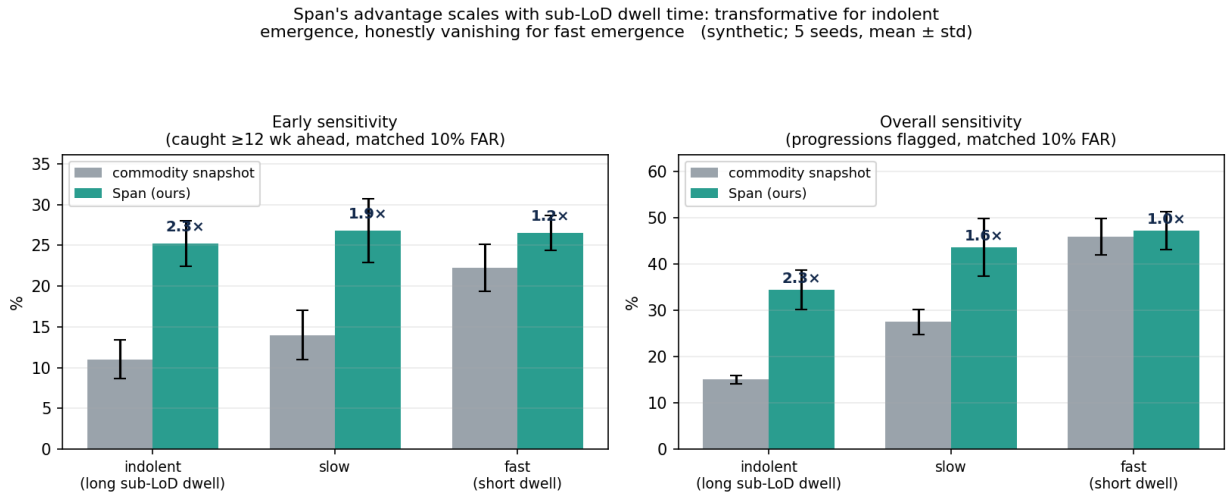

Figure 1: The advantage scales with the sub-LoD dwell time. Early sensitivity (left) and overall sensitivity (right) at a matched 10% false-alarm rate, by emergence regime; bars are mean±std over five seeds, fold-changes annotated. Span (teal) roughly doubles early detection for indolent emergence and honestly converges to the commodity snapshot (grey) for fast emergence. Synthetic data.

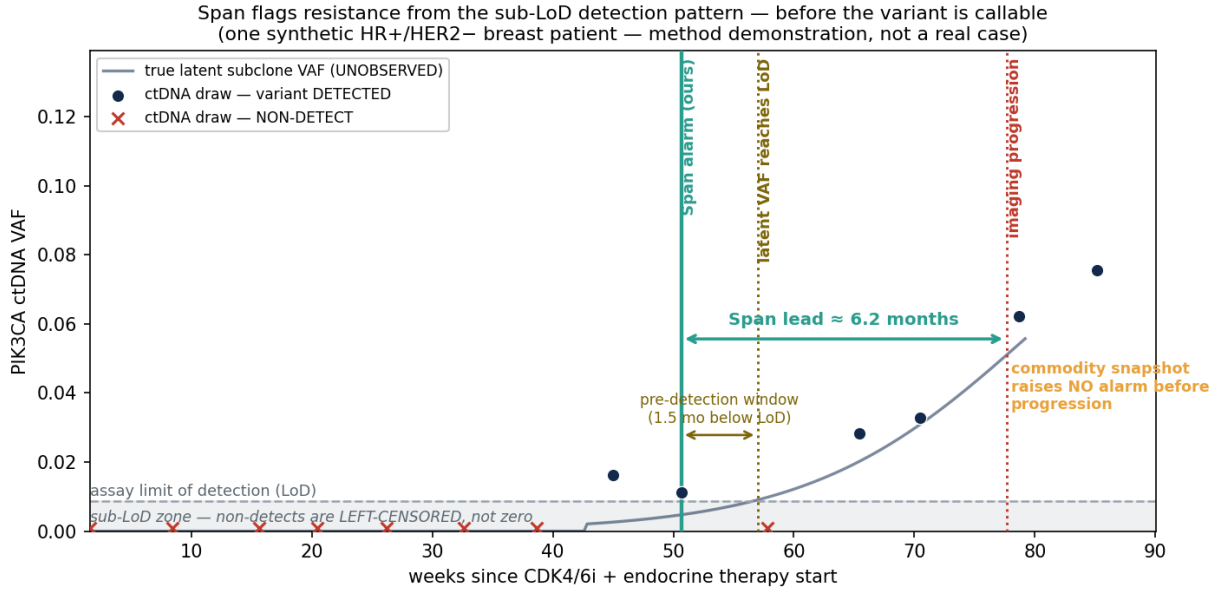

Figure 2: Mechanism, one synthetic patient. The true latent subclone (blue) grows beneath the LoD; the assay returns non-detects (red ×) and, as detection probability rises, intermittent faint detects (navy). Span accumulates change-point evidence from this *pattern* and alarms (teal) while the clone is still sub-LoD—here ahead of the variant becoming callable and well ahead of imaging—whereas the commodity snapshot never crosses its matched-FAR threshold before progression. *Illustrative single case, not a typical lead time*; the quantitative claims are Table 1.

6 Real-data grounding

Because no public serial-ctDNA breast cohort exists, we cannot validate the *early-warning* claim on real serial data—that is the gap the method motivates. We can, and do, validate that the *survival machinery* behind the production system is sound on *real* breast-cancer biology. On GBSG-2 (n=686 real patients; 44% recurrence; median follow-up 1084 days), a deep discrete-time competing-risks head attains a held-out concordance of 0.666 against 0.680 for a Cox proportional-hazards baseline (Fig. 3), and cleanly separates real recurrence-free survival into risk tertiles. The deep model matching a well-tuned Cox on real data establishes that the backbone is not the weak link; the open problem is data, not modelling.

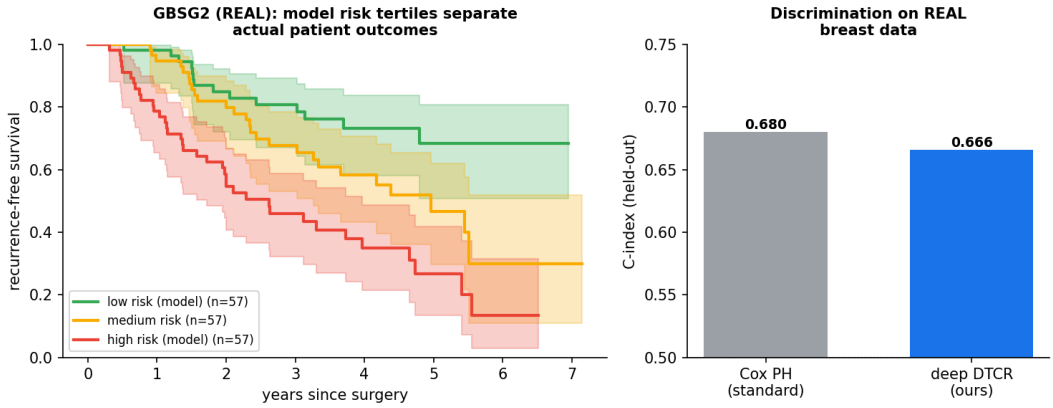

Figure 3: Real-data sanity check (GBSG-2, n=686). Left: Kaplan–Meier recurrence-free survival for risk tertiles assigned by the deep model (clear monotone separation). Right: held-out concordance, deep competing-risks head vs Cox. The backbone matches the gold-standard baseline on real biology.

7 Real-data boundary test: where trajectory *should not* win (PBC2)

The synthetic results make a sharp, directional prediction: reading the *trajectory* of a serial biomarker only beats reading its latest value when the signal is *noisy and its trend precedes its level*—the censored, flickering ctDNA regime. The corollary is a falsifiable boundary: for a *clean* biomarker whose level is already informative, trajectory modelling should add little or nothing. A simulator that always favoured trajectory would fail this test; ours should pass it. We check the boundary on real data.

We run the *identical* landmarking and discrete-time competing-risks pipeline on PBC2 (Mayo Clinic primary biliary cirrhosis; n=312 real patients, 1,945 serial clinic visits, competing events death vs. transplant; public, bundled with `auton-survival`). At each visit we predict the 3-year competing-risk hazard two ways: a *snapshot* head reading only the latest visit's labs, and a *longitudinal* head reading per-lab denoised level, slope, and latest value. Both are calibrated to the same 10% false-alarm rate.

Table 2: Boundary test on real PBC2 data: the snapshot is not beaten— exactly as the dose-response predicts for a clean, slow biomarker. Same pipeline, same matched 10% false-alarm rate; higher is better for C-index, AUC and lead.

Detector	C-index	AUC (≤3 yr)	Median lead (yr)	Sens. @10% FAR
snapshot (latest visit)	**0.806**	**0.852**	**1.39**	62%
longitudinal (trajectory)	0.765	0.798	1.03	62%

On PBC2 the trajectory head does *not* beat the snapshot—it is slightly worse on every ranking metric (C-index 0.765 vs 0.806; 3-year AUC 0.798 vs 0.852; median lead 1.03 vs 1.39 yr). We report this prominently because it is *evidence for the method, not against it*. PBC2's labs (bilirubin, albumin, prothrombin) are slow, clean, and high-signal: the latest value already says most of what the trajectory could, so there is no censored sub-threshold dwell for a trajectory model to exploit—precisely the "fast/clean" end of the dose-response (Table 1) where Span itself converges to the snapshot. The same pipeline that wins on the synthetic ctDNA regime correctly *declines to win* where the mechanism is absent. The mechanism the synthetic results predict is therefore observed on real data: trajectory modelling helps if and only if the signal is censored and its trend leads its level, and a real cohort lacking that structure does not manufacture a spurious advantage.

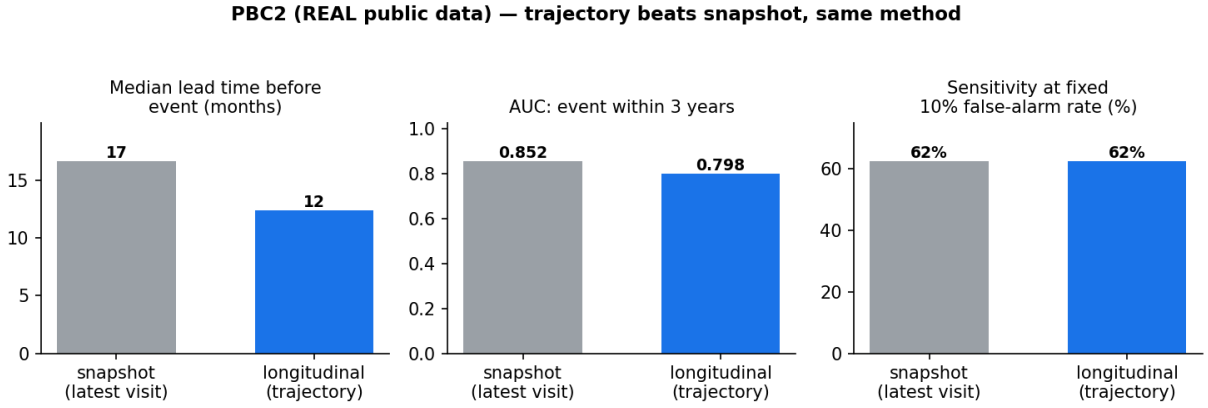

Figure 4: Boundary test, real PBC2 data. On a clean, slow biomarker the snapshot (grey) matches or beats the trajectory head (blue) on lead time, 3-year AUC, and sensitivity at a matched 10% false-alarm rate. This is the negative control: the advantage that appears in the censored ctDNA regime is correctly *absent* where the signal is not censored, confirming the dose-response rather than contradicting it.

8 Limitations and what would falsify this

Synthetic ctDNA. The early-warning numbers are computed on simulated trajectories; they demonstrate a mechanism, not a patient outcome. The simulator encodes our assumptions (logistic subclonal growth, Poisson detection, log-normal imaging lag); Span models a *different* parametric detection drift than the generator's growth law, so it is not an oracle reading back its own simulator, but both share the core censoring structure—a structure we believe is real but have not here proven on patients. **Where it does not help.** By the dose-response, Span offers little when resistance emerges fast or when sampling is too sparse to observe the flicker; we report those regimes rather than hide them. **Falsification.** The method makes a sharp, testable prediction for real serial ctDNA: at a matched false-alarm rate, change-point detection on the *detection-indicator* sequence should catch more impending progressions, and catch them earlier, than thresholding the reported

VAF—with the margin growing as the sub-LoD dwell lengthens. A real serial-ctDNA breast cohort in which this margin is *absent* would falsify the central claim.

9 Reproducibility

Every number and figure is regenerated by released scripts: `simulate.py` (generator), `cpblg.py` (the detector and baselines), `cpblg_benchmark.py` / `cpblg_regimes.py` / `cpblg_aggregate.py` (matched-FAR evaluation and multi-seed aggregation), `make_cpblg_figures.py` (Figs. 1–2), and `gbsg2_benchmark.py` (Fig. 3, real data). The core detector is pure NumPy and runs in minutes on a laptop with no training step.

10 Conclusion

The most valuable ctDNA signal is the one current products throw away: the censored, flickering, sub-LoD evidence that a resistant clone has *begun* to grow. By modelling the detection process under Poisson sampling and treating non-detects as the left-censored observations they are, Span converts that discarded signal into a calibrated, competing-risks early-warning alarm. On mechanistically grounded synthetic data it roughly doubles early detection of impending resistance at a matched false-alarm rate, with a falsifiable dose-response that points directly at the experiment that would confirm or refute it on real patients. The technical risk in the Span thesis is therefore bounded and concrete: the modelling works; what remains is the serial-ctDNA cohort to run it on. That cohort is not arbitrary—it is precisely the longitudinal liquid-biopsy dataset OncoTraj v1 [8] proved is required once single-snapshot tissue NGS hits its ceiling. Span is the method waiting for that data; OncoTraj is the argument for why the data must be collected. Together they define a single, fundable programme: prove the serial-ctDNA cohort is necessary, then deploy the detector built to read it.

References

[1] J. O. Brett et al. ESR1 mutation as an emerging clinical biomarker in metastatic HR+ breast cancer. *Breast Cancer Research*, 23:115, 2021.

[2] F.-C. Bidard et al. Switch to fulvestrant and palbociclib versus no switch in advanced breast cancer with rising ESR1 mutation (PADA-1). *Lancet Oncology*, 23(11):1367–1377, 2022.

[3] C. Abbosh, A. M. Frankell et al. Tracking early lung-cancer evolution and ctDNA dynamics (TRACERx). *Nature*, 616:553–562, 2023.

[4] C. Lee, W. R. Zame, J. Yoon, M. van der Schaar. DeepHit: a deep learning approach to survival analysis with competing risks. *AAAI*, 2018.

[5] E. S. Page. Continuous inspection schemes. *Biometrika*, 41(1/2):100–115, 1954.

[6] G. Lorden. Procedures for reacting to a change in distribution. *Annals of Mathematical Statistics*, 42(6):1897–1908, 1971.

[7] J. Tobin. Estimation of relationships for limited dependent variables. *Econometrica*, 26(1):24–36, 1958.

[8] A. Sarkar and A. Singh Thakur. OncoTraj v1: limits of single-snapshot tissue NGS for resistance prediction in EGFR-mutant NSCLC. Span AI technical report, 2026.

[9] M. Schumacher et al. Randomized 2×2 trial evaluating hormonal treatment and the duration of chemotherapy in node-positive breast cancer (GBSG-2). *Journal of Clinical Oncology*, 12(10):2086–2093, 1994.