CENSORING AS AN OUTCOME: ONE DIFFUSION FOR SURVIVAL PREDICTION AND COHORT SYNTHESIS

Anonymous authors
Paper under double-blind review

ABSTRACT

Survival models estimate when the event happens and treat when the patient left as a nuisance. That discard is why none of them yields a synthetic cohort a statistician can analyse: such a cohort needs the dropout mechanism too, so the target is the law of both latent times. Censoring at random factorises that law and makes it symmetric, an event record right-censoring the censoring time exactly as a censored record right-censors the event time. The censoring channel is then a fitted distribution rather than a weight, and follow-up becomes a quantity one can set instead of a property of the dataset: the model resamples a cohort under a dropout regime it never saw, which is what a power calculation or a planned extension asks for. A model of the observed pair (T^{obs}, E) has one lever, truncation, and truncation only shortens. We fit the law with a two-channel score-based diffusion on $(\log T^\star, \log C)$, trained by variational EM whose E-step draws the unobserved channel from a truncated normal in closed form; the objective upper-bounds the censored-channel negative log-likelihood, and the proposal's slack and the chain's consistency are bounded with it. On six real cohorts the same model leads nine baselines on macro integrated Brier score, time-dependent AUC and Brier resolution, four comparisons resolving at the cohort level, and composed with a tabular backbone it is best on every downstream-utility median we score and beats four competing generators in all six cohorts. Rescaling the fitted dropout curve leaves the event curve inside the estimator's own noise, where truncation matched to the same censoring rate moves it ten times further. Dropping the channel costs IBS, concordance, Brier resolution and reliability in five of six cohorts.

1 INTRODUCTION

Survival analysis models the time $T^\star$ to an event given covariates X, from n i.i.d. observations that are right-censored: per subject we record only $T_i^{\mathrm{obs}} = \min(T_i^\star, C_i)$ and $E_i = \mathbf{1}\{T_i^\star \le C_i\}$, with C_i a latent censoring time.[1] From Cox (Cox, 1972) through DeepSurv (Katzman et al., 2016) and DeepHit (Lee et al., 2018), the dominant paradigm targets $S(t \mid x) = P(T^\star > t \mid X = x)$ and treats C as a nuisance. That is enough to rank patients and not enough to write down a cohort, which is defined as much by when its subjects leave as by when their events occur. Aging biology needs such cohorts to fit and audit biological-age estimates (Kriukov et al., 2024; 2025).

The natural generative target is the latent law of both times

$$p(T^\star, C \mid X) = p(T^\star \mid X)\, p(C \mid X), \tag{1}$$

which factorises under censoring-at-random, $T^\star \perp C \mid X$, foundational to Kaplan–Meier, Cox and AFT estimators though untestable from observations alone (Tsiatis, 1975). Under CAR both marginals are non-parametrically identified, and the factorisation is structurally symmetric: an event record reveals $T^\star$ and right-censors C; a censored record reveals C and right-censors $T^\star$. The censored subjects are therefore not missing data for the event channel; they are observations of the other one. Almost nothing exploits that. Across discriminative estimators, tabular generators, IPCW pipelines and copula survival models, none yields a learned, sampleable $p(T^\star, C \mid X)$, and §2.1 says what each does instead. A misspecified $p(C \mid X)$ distorts the information fraction every downstream

[1] Right censoring throughout; left and interval censoring need different likelihood structures.

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

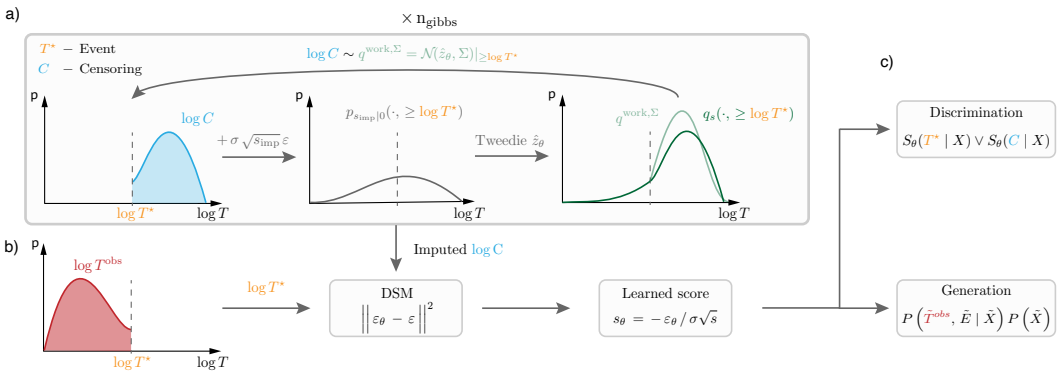


Figure 1: **FATUM in one picture.** **(a)** The E-step for an event record, whose censoring time C is unobserved and only lower-bounded by $\log T^{\text{obs}}$. The working proposal is the denoising Gaussian centred on the network's Tweedie estimate $\hat{z}_\theta$ and truncated to that half-line, and a draw from it replaces $\log C$, n_{gibbs} times over. The panel draws the corrected width Σ, for which the reported arm substitutes σ^2_{post}. The gap between the pale and dark curves is the slack Theorems 1 and 3 bound. **(b)** The M-step: the observed channel supplies $\log T^\star$ directly, the E-step supplies the imputed $\log C$, and both train a single denoising score $s_\theta = -\varepsilon_\theta/(\sigma\sqrt{s})$. **(c)** That one score serves both axes we evaluate: per-patient survival curves for either channel, and joint synthesis of $(\tilde{T}^{\text{obs}}, \tilde{E} \mid \tilde{X})$.

estimator sees. A joint object would also give what the field has no way to obtain today, a censoring law one can vary in order to ask whether a conclusion survives a different dropout regime.

We model that law directly with a score-based diffusion that denoises $(\log T^\star, \log C)$ in two channels. We call it FATUM, and Figure 1 summarises the construction. Log-space is what makes the censored channel tractable: a GBM schedule on T is a Gaussian diffusion on $z = \log T$, so the forward kernel $p_{s|0}(z_s \mid z_0) = \mathcal{N}(z_0 + \mu_{\text{drift}} s, \sigma^2_{\text{post}}(s))$ is Gaussian at every level, and the censoring constraint $T^\star \geq T^{\text{obs}}$ becomes the half-line $z_0 \geq \log T^{\text{obs}}$. Truncating the Tweedie-centred denoising Gaussian to that half-line gives a truncated normal with closed-form inverse CDF, so each E-step draw is exact, with no rejection step and no Metropolis correction, and the M-step minimises the resulting bound.

An EM step under censoring needs to sample the latent restricted to a half-line, in closed form. A parametric AFT can, at the price of fixing the shape of the log-time density, and among learned families a VAE (Kingma & Welling, 2014) reaches the constraint only through its evidence bound, a GAN (Goodfellow et al., 2014) has no density to truncate, and a normalising flow (Rezende & Mohamed, 2015; Papamakarios et al., 2021) sends the half-line to a curved region with no closed-form mass. A diffusion is Gaussian at every noise level while the law it composes to is not, so the flexibility and the exact truncation come together.

The same factorisation bounds what the decomposition can buy: it adds no identifying information about a functional of the observed law, so what it changes is estimation, where a proper scoring rule can see it. What it does add is an operation, since rescaling a fitted censoring clock acts on a latent.

Contributions. ***A two-channel diffusion over both times.*** We model $(T^\star, C)$ as one GBM diffusion on $(\log T^\star, \log C)$, so every record trains both channels where a likelihood for $p(T^\star \mid X)$ uses half of each.

An E-step whose slack is characterised. We bound the censored-channel NLL and the proposal's slack in closed form, and prove the imputation chain consistent up to a floor we price rather than assume away (Lemma 1, Theorems 1 and 2).

An operation the observed pair does not admit. Because $p(C \mid X)$ is fitted rather than applied as a weight, the cohort can be resampled under a dropout regime the model never saw, which no model of (T^{obs}, E) can do.

Evidence on two axes, at the cohort level. FATUM leads nine baselines on macro IBS, tdAUC and Brier resolution, and composed with a tabular backbone it leads every downstream-utility measure we score.

2 BACKGROUND

Score-based diffusion (Vincent, 2011; Song et al., 2020; Ho et al., 2020) trains a denoiser across a hierarchy of noise levels, its weighted objective bounds the negative log-likelihood from above (Song et al., 2021), and Tweedie's formula $\mathbb{E}[z_0 \mid z_s] = z_s + \sigma^2_{\text{post}}(s)\nabla_{z_s} \log p_s(z_s)$ (Efron, 2011) turns the score into a posterior-mean estimate, which §3.3 uses to centre the proposal. Variance-preserving and variance-exploding schedules treat all coordinates symmetrically (Song et al., 2020); for strictly positive outcomes geometric Brownian motion is the natural alternative, since diffusing $z = \log T$ is Gaussian at every level and lognormal in T (Stojkoski et al., 2020).

2.1 RELATED WORK

Discriminative survival models estimate $p(T^\star \mid X)$ and discard the censoring times: the Cox family (Cox, 1972), its neural extensions (Katzman et al., 2016; Kvamme et al., 2019), DeepHit (Lee et al., 2018), random survival forests (Ishwaran et al., 2008) and gradient-boosted analogues (Hothorn et al., 2006). None exposes a sampleable law over the pair $(T^\star, C)$. *Tabular generators* (Xu et al., 2019; Kotelnikov et al., 2022; Shi et al., 2024) fit mixed-type columns without treating time and event status as anything special, and *survival-aware generators* (Norcliffe et al., 2023; Brockschmidt et al., 2025; Ashhad & Henao, 2024) add survival-tailored losses to the same target: all of them fit the *observed* law of (X, T^{obs}, E) rather than the latent factorisation in equation 1. *Copula survival models* (Gharari et al., 2023; Zhang et al., 2024) are the one family that parameterises both latent marginals at once, and they do it to recover $p(T^\star \mid X)$ under dependent censoring: the censoring marginal is an identification device there, never a sampler, and is never evaluated as one. *Posterior sampling with a diffusion prior* is the other neighbour: Diffusive Gibbs Sampling (Chen et al., 2024) and the Plug-and-Play Split Gibbs sampler (Coeurdoux et al., 2023) formalise the alternating denoise-and-project scheme for a *fixed* prior, and our E-step is an instance of it on a half-line, with the prior learned jointly through variational EM, which is the case Theorem 2 has to cover.

3 METHOD

3.1 JOINT DIFFUSION ON LOG-TIMES WITH A GBM NOISE SCHEDULE

We model each latent log-time $z_0 \in \{\log T^\star, \log C\}$ as the clean signal in a forward diffusion. When $E = 1$, $\log T^\star$ is observed and $\log C > \log T^\star$; when $E = 0$ the roles reverse. Both channels are noised independently with the same kernel, so we state the theory for a generic scalar z_0. The forward kernel is a constant-coefficient geometric Brownian motion:

$$p_{s|0}(z_s \mid z_0) \;=\; \mathcal{N}\big(z_0 + \mu_{\text{drift}}\, s,\; \sigma^2_{\text{post}}(s)\big), \tag{2}$$

with $\mu_{\text{drift}} = \mu - \frac{1}{2}\sigma^2$ and $\sigma^2_{\text{post}}(s) = \sigma^2 s$. The score network $\varepsilon_\theta(z_s, s, x)$ is ε-parameterised, $s_\theta = -\varepsilon_\theta/\sigma_{\text{post}}$, with Tweedie mean

$$\hat{z}(z_s, x) \;=\; z_s - \mu_{\text{drift}}\, s - \sigma_{\text{post}}\, \varepsilon_\theta(z_s, s, x). \tag{3}$$

3.2 EXACT-CHANNEL LOSS

For the observed channel, $\log T^\star$ when $E = 1$ and $\log C$ when $E = 0$, the path-space ELBO of Song et al. (2021) combined with log-uniform absorption of the GBM $1/s$-weight gives

$$\mathbb{E}_{X,z_0}\big[-\log p_\theta^{s_{\min}}(z_0 \mid X)\big] \;\le\; L \cdot \mathcal{L}^{\text{exact}}_{\text{DSM}}(\theta) + C, \qquad L = \log(s_{\max}/s_{\min}), \tag{4}$$

with $p_\theta^{s_{\min}}$ the model marginal at noise level $s_{\min}$, C independent of θ, and $\mathcal{L}^{\text{exact}}_{\text{DSM}}(\theta) = \frac{1}{2}\mathbb{E}[(\varepsilon_\theta(z_s, s, X) - \varepsilon)^2]$ *unweighted*: the log-uniform sampling density cancels the GBM weight

pointwise, so that step is an identity. Nothing in it uses the prior beyond square-integrability of its score. Lemma 1 of Appendix B states the hypotheses and proves it.

Takeaway. Under CAR the observed-data likelihood splits into an event factor and a censoring factor. We optimise both, so the censoring channel comes back as a distribution the model can sample rather than a weight it applies.

3.3 CENSORED-CHANNEL LOSS: VARIATIONAL EM BOUND

For the unobserved channel, $\log T^\star$ when $E = 0$ and $\log C$ when $E = 1$, we know only that $z_0 \geq \log T^{\text{obs}}$. The per-sample contribution to the observed-data NLL is the integral

$$\mathcal{N}_{\text{cen}}(\theta; x, T^{\text{obs}}) \;:=\; -\log \mathbb{P}_\theta(z_0 \geq \log T^{\text{obs}} \mid x), \tag{5}$$

which the implicit diffusion prior leaves intractable. We bound it variationally: for any proposal $q(z_0 \mid x)$ supported on $\{z_0 \geq \log T^{\text{obs}}\}$, the Neal–Hinton identity (Neal & Hinton, 1999) gives

$$\mathcal{N}_{\text{cen}} \;=\; \mathbb{E}_{z_0 \sim q}[-\log p_\theta(z_0 \mid x)] - H(q) - \text{KL}(q \,\|\, q^\star), \tag{6}$$

with $q^\star := p_\theta(\cdot \mid x, z_0 \geq \log T^{\text{obs}})$ the true *marginal* truncated posterior. Since $\text{KL} \geq 0$, dropping it leaves a valid upper bound for any such q. Applying equation 4 inside the expectation (up to its $O(s_{\min})$ smoothing of $-\log p_\theta$) yields the trainable upper bound

$$\mathcal{N}_{\text{cen}} \;\leq\; -H(q) + L \cdot \mathbb{E}_q\big[\mathcal{L}^{\text{exact}}_{\text{DSM}}(\theta; z_0, x)\big] + C(q). \tag{7}$$

We adopt the *working-Gaussian* proposal at imputation noise level s_{imp},

$$q^{\text{work}}(z_0 \mid x, z_s, s_{\text{imp}}) \;:=\; \mathcal{N}\big(\hat{z},\, \sigma^2_{\text{post}}(s_{\text{imp}})\big)\big|_{z_0 \geq \log T^{\text{obs}}} \tag{8}$$

The proposal $q^{\text{work}}(\cdot \mid z_s)$ approximates the denoising posterior $q_s := p_\theta(\cdot \mid x, z_s, z_0 \geq \log T^{\text{obs}})$ at a fixed z_s, not the marginal $q^\star$. Theorems 1 and 3 quantify this per-step gap, and Appendix B.7 completes the justification of equation 7 once the chain is in place.

Definition 1 (Local Laplace coefficients)**.** Let $z_0^\dagger(z_s, x)$ denote the mode of q_s (unique under (A2); the assumptions are Appendix A) and $m(u) := \phi(u)/\Phi(u)$ the Mills ratio. Write $V_{\text{prior}} := -\partial^2_{z_0} \log p_\theta(z_0 \mid x)\big|_{z_0^\dagger}$ and $g_k := \partial^k_{z_0} \log p_\theta(z_0 \mid x)\big|_{z_0^\dagger}$ for $k = 3, 4$, all evaluated at the mode and all functions of (z_s, x), and set

$$\Sigma(z_s, x) := \frac{\sigma^2_{\text{post}}(s_{\text{imp}})}{1 + \sigma^2_{\text{post}}(s_{\text{imp}})\, V_{\text{prior}}(z_s, x)}, \qquad u^\star(z_s, x) := \big(z_0^\dagger(z_s, x) - \log T^{\text{obs}}\big)/\sqrt{\Sigma(z_s, x)}.$$

V_{prior} has no closed form, so the algorithm substitutes the denoiser Jacobian estimate $\widehat{V}(z_s, x)$ and runs the *plug-in* variance

$$\widehat{\Sigma}(z_s, x) \;=\; \frac{\sigma^2_{\text{post}}(s_{\text{imp}})}{1 + \sigma^2_{\text{post}}(s_{\text{imp}})\, \widehat{V}(z_s, x)}. \tag{9}$$

The theorems below are stated for Σ; the arm that runs $\widehat{\Sigma}$ is FATUM$_\Sigma$, whose cost §5.3 measures and whose $\widehat{V}$ Appendix C.3 gives in closed form equation 21 with the clip it needs.

Theorem 1 (Proposal slack at q^{work})**.** *Under (A1)–(A5), writing $\sigma_{\text{post}} \equiv \sigma_{\text{post}}(s_{\text{imp}})$ throughout, the following bound holds in the small-noise regime $\sigma^2_{\text{post}}\, V_{\text{prior}} \to 0$:*

$$\text{KL}(q^{\text{work}} \,\|\, q_s) \;\leq\; C_{\text{var}}(u^\star)\, \sigma^4_{\text{post}}\, V^2_{\text{prior}} + \mathcal{K} + C_1 + C_2, \tag{10}$$

where

$$C_{\text{var}}(u^\star) \;=\; \tfrac{1}{4} - \tfrac{u^\star\, m(u^\star)}{8}\big[1 + u^\star\big(m(u^\star) + u^\star\big)\big], \tag{11}$$

$\mathcal{K} = \mathcal{K}(u^\star, m(u^\star) - \delta_\theta)$ of equation 16 the location term, which reduces to $\frac{1}{2}\delta^2_\theta$ where the boundary is far, for $\delta_\theta := \varepsilon_\theta - \varepsilon^\star$ the score error, $C_1 = O(\sigma^5_{\text{post}}|g_3|V_{\text{prior}} + \sigma^6_{\text{post}}(g_3^2 + V^3_{\text{prior}}))$ and $C_2 = O(\sigma^8_{\text{post}}(g_4^2 + V^4_{\text{prior}}) + \sigma^{12}_{\text{post}} g_3^4)$.

Proof in Appendix B.2.

Replacing the forward variance σ^2_{post} by the Laplace variance Σ cancels the variance term and lowers the vanishing part of the slack to $O(\sigma^6_{\text{post}} g_3^2)$. It leaves $\mathcal{K}$ untouched, which is what §5.3 finds. Appendix F gives the corrected proposal, its slack bound and the consistency rate it implies.

3.4 Consistency of Imputation Chain

Fix x and one censored channel. The E-step is the data-augmentation Gibbs sampler of Tanner & Wong (1987): its ideal kernel P draws $z_s \mid z_0 \sim p_{s|0}$ and then $z_0 \mid z_s \sim q_s$, and leaves $q^\star$ invariant. Write $\tilde{P}$ for the surrogate that puts q^{work} in place of q_s. Its conditionals are those of no joint density (Arnold & Press, 1989), so a fixed point of $\tilde{P}$, where one exists, is not $q^\star$, and $\tilde{P}$ is not reversible. The theorem uses neither. Whether a fixed point exists at all is a property of the trained checkpoint rather than of the method, which is why the theorem claims none; Appendix D.4 measures it and Appendix D lists the hypotheses that would remove the n. Beyond (A1)–(A5) the theorem adds the curvature ceiling (A6), which supplies the uniform $\bar{V}$ below.

Theorem 2 (Consistency of the imputation chain)**.** *Assume (A1)–(A6), and write $\nu_k := \nu_0 \tilde{P}^{\,k}$ for the law of the imputed draw after k sweeps from an initial law ν_0. Both parts hold for arbitrary ν_0, which the algorithm needs: with `prior_anchored` on, five of the six reported runs start the latent channel at $\max(\log T^{\text{obs}}_{\pi(i)}, \log T^{\text{obs}}_i)$ for a random permutation π, so ν_0 is a shuffled empirical law with an atom on the boundary rather than a point mass.*

(i) (Algorithmic error, unconditional.) *Write $\bar{\varepsilon}_{\text{TV}} := \sup_{z_0} \mathbb{E}_{z_s|z_0} \| q^{\text{work}}(\cdot \mid z_s) - q_s(\cdot \mid z_s) \|_{\text{TV}}$, the expectation being under the forward kernel $p_{s|0}(\cdot \mid z_0)$. Then for every $n \geq 1$,*

$$\bar{\varepsilon}_{\text{TV}} \;\leq\; \sqrt{\tfrac{1}{2}\bar{C}_{\text{var}}}\,\sigma^2_{\text{post}}\bar{V} + \sqrt{\bar{\mathcal{K}}/2} + \alpha + O(\sigma^3_{\text{post}}),$$

$$\left\| \nu_0 \tilde{P}^{\,n} - \nu_0 P^{\,n} \right\|_{\text{TV}} \;\leq\; n\,\bar{\varepsilon}_{\text{TV}},$$

with α the exception probability of (A5), $\bar{C}_{\text{var}} := \sup_u C_{\text{var}}(u) = \frac{1}{2}$ (Lemma 2) and $\bar{\mathcal{K}} := \sup_{u \geq u_-} \mathcal{K}(u, m(u) - \delta_\theta)$, which is finite on the interior-mode event, where m is bounded, whenever the score error δ_θ is.

(ii) (Ergodic error of the ideal chain.) *For each fixed s_{imp}, $\| \nu_0 P^{\,n} - q^\star \|_{\text{TV}} \to 0$ as $n \to \infty$, from every ν_0.*

(iii) Hence $\| \nu_n - q^\star \|_{\text{TV}} \leq n\bar{\varepsilon}_{\text{TV}} + \| \nu_0 P^{\,n} - q^\star \|_{\text{TV}}$ at any fixed s_{imp}, and along a schedule with $n_{\text{gibbs}}(\sigma^2_{\text{post}}\bar{V} + \alpha) \to 0$ and $\| \nu_0 P^{\,n_{\text{gibbs}}} - q^\star \|_{\text{TV}} \to 0$ the bound is $n_{\text{gibbs}}\sqrt{\bar{\mathcal{K}}/2} + o(1)$, which vanishes with $\bar{\mathcal{K}}$ (Remark 2).

Part (i) needs no invariant law; Appendix D lists what it assumes and what it does not. The first form in (iii) is the one the E-step runs, s_{imp} being fixed there, and its growth in n is the price of using only non-expansion: the realised accumulation is geometric rather than linear and settles at a plateau, which Appendix D.2 measures. The schedule in the second form does not follow from (ii), which holds at a fixed s_{imp} and carries no rate, so we measure that there too.

Proof in Appendix B.4; Appendix D.2 measures the per-sweep behaviour the proof bounds.

Takeaway. We make one approximation and no others. The exact posterior q_s is out of reach, since the prior it conditions on is a learned score, so the E-step draws from a truncated Gaussian q^{work} that is a little too wide and a little off-centre. Theorems 1 and 2 bound the cost of the substitution.

4 Experimental Setup

We evaluate on six public right-censored cohorts spanning $n = 686$ to 8,873 patients and censoring rates from 32% to 92% (Table 3, Appendix C): AIDS (Hammer et al., 1997), GBSG2 (Schumacher et al., 1994), METABRIC (Curtis et al., 2012), SUPPORT (Knaus et al., 1995), FLCHAIN (Dispenzieri et al., 2012) and Rotterdam (Foekens et al., 2000). One stratified $80/20$ dev/holdout split is fixed per cohort and shared across all methods; hyperparameters are tuned per cohort by Optuna TPE (Akiba et al., 2019) on the dev fold and holdout metrics reported under the best dev-val configuration, repeated under ten fixed seeds. Two arms are reported: **FATUM**, on the uncorrected proposal q^{work}, which all headline numbers refer to, and **FATUM$_\Sigma$**, an *ablation* on $q^{\text{work},\Sigma}$. Architecture, search ranges, optimiser, sampler settings, seed list and hardware are in Appendix C.

Table 1: **Holdout discriminative performance**, macro over the six cohorts: each cell is the mean of the six cohort means. **Bold** is the best competing model; FATUM$_\Sigma$ (gray) is our own ablation and does not compete. The ordering is a ranking; §5.2 says which differences resolve at the cohort level.

Metric	FATUM	FATUM$_\Sigma$	CoxPH	CoxCC	CoxTime	WeibAFT	LogNAFT	RSF	GBSA	DeepSurv	DeepHit
IBS ↓	**0.1437**	0.1436	0.1477	0.1469	0.1491	0.1478	0.1485	0.1499	0.1539	0.1453	0.1559
tdAUC ↑	**0.7365**	0.7361	0.7260	0.7126	0.7150	0.7273	0.7254	0.7270	0.7008	0.7232	0.7017
C_{IPCW} ↑	0.6691	0.6683	0.6726	0.6619	0.6538	**0.6735**	0.6712	0.6639	0.6487	0.6682	0.6676
RES ↑	**0.0205**	0.0206	0.0183	0.0170	0.0155	0.0187	0.0201	0.0166	0.0174	0.0164	0.0167
REL ↓	0.0056	0.0056	0.0081	0.0067	**0.0051**	0.0095	0.0094	0.0056	0.0090	0.0052	0.0095

We report the integrated Brier score (Graf et al., 1999), its Murphy decomposition (Pohle, 2020) into reliability and resolution at the median holdout event time, Uno's inverse-probability-of-censoring concordance (Uno et al., 2011), which is consistent under censoring, and time-dependent AUC (Pantoja-Galicia et al., 2021), which scores the ordering separately at each time and so does not require non-crossing curves. Concordance runs on the risk score $-\log \hat{S}(t_{\mathrm{ref}} \mid x)$ at the median holdout event time, one ordering for all times.

4.1 Unit of analysis

The unit is the cohort ($m = 6$), not the cell ($n = 60$): the ten seeds inside a cohort share one split. Each cohort contributes its mean paired difference and we report the Student-t interval over the six means with the cohorts won. A sign test at $m = 6$ cannot fall below $2^{1-m} = 0.031$, so we report no stars, here or where a comparison goes against us (Appendix E).

5 Results

We report three things in turn: a controlled simulation, which answers what real data cannot since $T^\star$ is never seen for a censored subject; the six cohorts, which answer whether any of it transfers; and the ablations. FATUM predictions throughout are Monte-Carlo, $N_{\mathrm{mc}} = 500$.

5.1 Controlled simulation, where the latent is observable

Whether the E-step recovers the latent it imputes, and whether the Σ correction behaves as Theorem 3 derives, are unanswerable on censored data. We answer them on a simulation with known latents, $\log T^\star = x^\top \beta + \sigma_T \varepsilon$ and $\log C = g_0 + x^\top \gamma + \sigma_C \varepsilon'$ with g_0 set to the target censoring rate. The model does not generate these cohorts; both latent times come from a process we write down, so both are known per subject. Replications are independent datasets, so ordinary paired tests apply and the 2^{1-m} floor of §4.1 is not in force.

The E-step is calibrated: across three regimes, ten seeds and five imputation scales the 90% intervals run 0.812 to 0.920. Under CAR a consistent fit of a correctly specified family reproduces the constrained posterior exactly, so calibration alone cannot separate an imputer that uses the covariates from one that happens to be the right width. The next comparison can.

The channel shows itself where its structure exists. Calibration can come from a proposal whose width happens to be right, so the comparison that isolates the conditional signal permutes covariates within event-status strata at E-step time and changes nothing else. The sweep varies γ at fixed β, over zero, β itself (aligned) and a direction orthogonal to it at the same norm, so *zero* is a level at which the censoring channel carries no conditional structure while the event channel still does, which makes that channel the one this axis reads. The null leads on W_1 throughout, where at matched width the covariates buy the location and our draws stay wider than the truth, as the coverage above the oracle's and the marginal distance of §5.2 record. On the per-subject error the channel shows itself: it beats the null where the structure exists and nowhere else, $+0.0116$ ($p = 0.001$) aligned and $+0.0190$ ($p = 0.004$) orthogonal at $s_{\mathrm{imp}} = 0.10$ against -0.0001 ($p = 0.94$) at zero, at all four scales and so at the operating range.

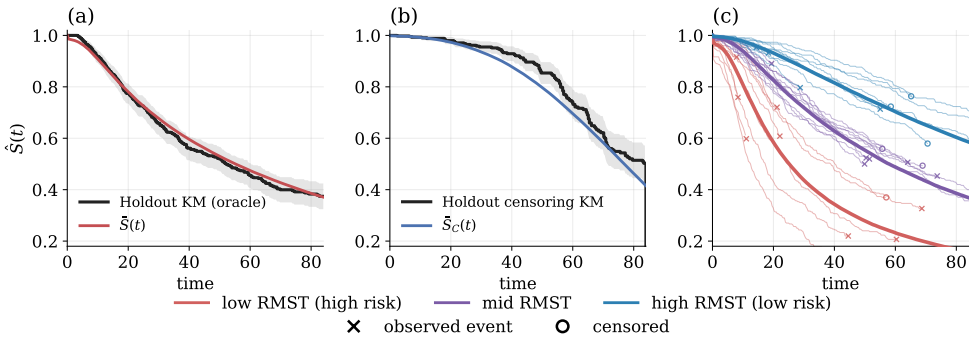


Figure 2: **Generated cohort on real holdout covariates (Rotterdam, seed 2960).** (a) holdout KM against the model's marginal $\bar{S}(t)$; (b) the same for the censoring channel; (c) survival by predicted-RMST tertile, thin curves per subject. Panel (b) is the one no discriminative model produces.

The Σ correction behaves as derived, and predicts its own size. The corrected-minus-uncorrected gap widens linearly in s_{imp}, as the leading $O(\sigma^2_{\text{post}} V_{\text{prior}})$ bias requires, on all six combinations of metric and channel, with $|r| = 0.987$–0.997 on five and 0.908 on $T^\star$ coverage (Fig. 7). Its benefit crosses zero between $s_{\text{imp}} = 0.05$ and 0.10 on the two error metrics and between 0.10 and 0.50 on coverage, and below its crossing the plug-in $\widehat{V}$ costs more variance than Σ removes. The six cohorts tune on both sides of that crossing, so no consistent direction is predicted across the panel (§5.3).

5.2 Six real cohorts: three predictions of the separability argument

§1 places the decomposition on the estimation side of the separability argument: every record trains both channels of one score network, so a censored subject informs the event channel instead of only reweighting it. Three predictions follow, and we test them in turn: a gain where latent structure is expressible (P1), an advantage over generators that lack the decomposition (P2), and an operation on the latent that a model of the observed pair cannot perform (P3). The same argument names where no gain is available: rank-only and purely marginal criteria are fixed by the observed law, and our objective has no ranking term and no marginal-shape term. A model that merely fit better would gain there too, so those columns are a refutation channel and we report them in full.

(P1) Gain where latent structure is expressible. FATUM attains the best macro IBS (0.1437, 95% CI 0.1422–0.1453), tdAUC (0.737, 0.732–0.741) and Brier resolution (0.0205) of the ten methods compared. Four cohort-level intervals clear zero, and they are the claims we make: IBS over DeepHit ($+0.0122$, $p = 0.011$) and over CoxTime ($+0.0054$, $p = 0.025$), both in all six cohorts; tdAUC over DeepHit ($+0.035$, $p = 0.033$), also unanimous; and Brier resolution over DeepSurv ($+0.0042$, $p = 0.036$) in five of six. Each p is the paired Student-t over the six cohort means, not the sign count beside it, and is not floored at 0.031; Table 8 has the intervals, and the rest sit under what $m = 6$ can resolve at the effect sizes the power analysis fixed in advance (Appendix E).

The Murphy decomposition locates the gain. Resolution, the across-subject variance of predicted risks, is the highest of the ten in the macro mean and in four of six cohorts, 12% above CoxPH although no term in the objective asks for ranking, while reliability (0.0056) stays within 10% of DeepSurv (0.0052): sharper stratification at comparable calibration. Table 7 gives IBS by cohort, where beating all nine baselines at once happens in three of six.

(P2) Advantage over generators lacking the decomposition. The test synthesises a whole cohort and scores it by training audit models on synthetic data alone and testing them on held-out real patients. FATUM enters as TabDiff for $p(X)$ composed with FATUM for $p(T^\star, C \mid X)$ (Fig. 2). Against the four generators that do not compose a survival-specific conditional onto a tabular sampler, FATUM wins in all six cohorts on nine metric-by-generator comparisons (Fig. 8). It leads every utility row of Table 2, and against SurvDiff the margin resolves on the RSF audit's IBS, unanimous at $p = 0.027$.

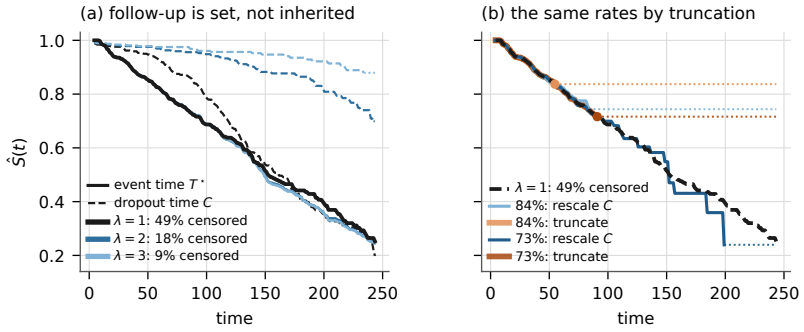


Figure 3: **Follow-up as a setting** (METABRIC, seed 2960, $n = 381$). Solid where events support the estimate, dotted over the plateau after the last one. **(a)** The fitted censoring clock puts the dropout curve (dashed) where it is asked; the event curve (solid) does not follow. **(b)** Truncation buying the same two rates stops at $t = 55$ and $t = 91$ (dot), where rescaling reaches $t = 84$ and $t = 199$; neither rate in (a) is available to it at all.

Table 2: **Synthetic cohort: fidelity and downstream utility.** Each cell is the median over the six cohorts of that cohort's mean over ten seeds; the median because two cohorts carry our W_1 (1.30 and 0.96 against 0.09–0.19), so a mean would report a distance no cohort has, and the competing models order the same either way. **Bold** is the best competing model; FATUM$_\Sigma$ (gray) is our own ablation and does not compete. Gray n/a marks too few cohorts for a median. Per-cohort counts are in Fig. 8.

Metric	TabDiff	SurvDiff	SurvivalGAN	TVAE	CTGAN	FATUM	FATUM$_\Sigma$
fidelity, times							
KM-MSE (train) ↓	1.32×10^{-3}	4.41×10^{-3}	5.34×10^{-3}	0.056	0.059	$\mathbf{7.65\times10^{-4}}$	8.14×10^{-4}
KM-MSE (holdout) ↓	3.05×10^{-3}	3.72×10^{-3}	5.20×10^{-3}	0.054	0.059	$\mathbf{1.84\times10^{-3}}$	1.56×10^{-3}
JS (time) ↓	**0.089**	0.256	0.260	0.246	0.258	0.168	0.160
W_1 (norm) ↓	**0.023**	0.374	0.228	0.186	0.226	0.184	0.159
downstream utility (TSTR)							
IBS, Cox audit ↓	0.182	0.183	0.213	n/a	n/a	**0.181**	0.182
IBS, RSF audit ↓	0.181	0.190	0.201	n/a	n/a	**0.181**	0.181
tdAUC, Cox audit ↑	0.678	0.670	0.652	n/a	n/a	**0.685**	0.691
tdAUC, RSF audit ↑	0.672	0.647	0.644	n/a	n/a	**0.704**	0.705

The backbone column is the controlled one, since it supplies our covariate sampler and the whole utility difference is attributable to $p(T^\star, C \mid X)$. There we lead the cohort median on nine of eleven metrics and none at better than four of six, so that column is a ranking and not a win, which is what the same bound predicts.

(P3) An operation the observed pair does not admit. Scale the fitted censoring clock by λ and rebuild the observable: $T_\lambda^{\text{obs}} = \min(T^\star, \lambda C)$, $E_\lambda = \mathbf{1}\{T^\star \le \lambda C\}$. Follow-up becomes a setting (Figure 3a). The event channel is untouched, so its Kaplan–Meier estimate must not move; a shift growing with $|\log \lambda|$ was the pre-registered refutation. We read shifts against the estimator's own noise, C redrawn at the same λ. Over six cohorts, three seeds and $\lambda \in \{0.33, 0.5, 0.75\}$ the median shift is 0.59 of it and does not grow. Truncation, the only lever a model of (T^{obs}, E) has, moves the curve ten times further at the same censoring rate, and above $\lambda = 1$ it has no move at all. The ratio is quoted only where truncation is defined: lengthening follow-up shrinks the shift by a further factor of sixty and shrinks the noise floor with it. AIDS is the exception, at 88 to 97% censoring on some twenty holdout events.

Three columns do not move. Concordance ties, 0.6691 against 0.6735 for the best baseline, fourth of ten inside a spread of 0.0044 and evidence in neither direction at either unit ($p \ge 0.266$, $q \ge 0.128$, Table 6). The marginal Kaplan–Meier distance ties too, four of six cohorts, $p = 0.27$. The normalised Wasserstein distance we lose to the tabular backbone in all six cohorts, cohort median 0.184 against 0.023, the only unanimous loss in the study.

5.3 Ablations: the censoring channel and the Σ correction

Neither axis above isolates a component of our own, so we ablate two directly and under one protocol: same architecture, hyperparameters, E-step and seeds, over all sixty cohort-by-seed cells, none coming back bitwise identical to its control.

Removing the censoring channel. The ablated arm inherits the configuration tuned for the full one, so the design is set up to produce a tie, which is also what the separability argument predicts in the population limit. The full arm wins anyway: dropping C from the loss costs IBS in five of six cohorts at 0.08 to 1.40 between-seed standard deviations, and costs C_{IPCW}, Brier resolution and reliability in five of six and tdAUC in three (Table 9, Appendix F.3). Nothing resolves at the unit of §4.1, and the cohort-level t moves with the scale, from $p = 0.48$ on raw IBS to 0.06 standardised, so we report the direction and claim no win. The attribution limit follows: the IBS lead over the baselines belongs to the fitted model as a whole, and the channel shows itself alone where the latent is observable (§5.1).

Removing the Σ correction. Its closed-form divergence follows the predicted rate across the probe grid, $s_{\mathrm{imp}} = 0.05$ to 0.8, in all six cohorts (Fig. 6), so the mechanism is the one the theory describes. It moves one real-data quantity, W_1, by 14% in the cohort median. Downstream it separates no metric, at either unit (Table 1, and Table 9(b) by cohort), and the expansion says why: the selected operating points straddle the zero-crossing of §5.1. The analysis that derives the correction is also what licenses skipping it (Appendix F).

Side by side the two ablations separate the advertised improvement from the component that carries the claim (Table 9). On IBS the standardised effects are 0.00 to 0.35 between-seed standard deviations for Σ, whose macro means change sign from metric to metric, against 0.08 to 1.40 for the channel.

6 Discussion and conclusion

Survival analysis has always treated the censoring mechanism as something to correct for. Modelling both latent times makes it something to sample from, and once it can be sampled, follow-up stops being a property of the dataset and becomes a quantity one can set. Asked for follow-up the study never ran, the model returns a cohort; a model of (T^{obs}, E) has no operation to answer with, its one lever being truncation, which only shortens. Power calculations and planned extensions are questions of that shape, and what such a cohort is worth is the obvious thing to measure next.

On six real cohorts the same model leads nine baselines on macro integrated Brier score, time-dependent AUC and Brier resolution, and composed with a tabular backbone it is best on every downstream-utility median we score, across two audit models and two metrics, beating four competing generators in all six cohorts. Seven comparisons resolve at the unit we argue for. Where the separability argument says no gain is available, none appears, which is the same argument paying twice.

Two things are open. We lose the normalised Wasserstein distance on event times in all six cohorts, an unbounded latent that no term in the objective asks to match in shape. And all of it rests on $T^{\star} \perp C \mid X$, which dependent censoring breaks and no test on observed data can check (Tsiatis, 1975).

Reproducibility statement

The code that produces every reported number is included as supplementary material, together with the fixed per-cohort splits and the selected configuration for each cohort, one shell entry point per experiment, and the self-tests for the two ablation wrappers. All six cohorts are public. The stratified dev/holdout split is fixed per cohort and shared across every method compared; the ten seeds are listed in Appendix C together with the architecture, the Optuna search space, the optimiser schedule, the sampler settings and the training hardware, a single NVIDIA A100 80GB per run. §4.1 and Appendix E state the unit of analysis, the cluster-bootstrap procedure and the resolution floor that follows from $m = 6$, so that every reported interval can be recomputed from per-cohort means.

Full proofs of every theorem and lemma, with their assumptions collected in Appendix A, are in Appendix B, where each result's scope is stated with it and the steps we do not complete are marked as such (Appendix D).

ETHICS STATEMENT

This work generates synthetic patient cohorts, and it does so with the censoring mechanism deliberately reproduced, since that is what makes such a cohort usable for survival analysis. The same property makes it more disclosive than a generator that reproduces only marginal times: a synthetic record whose dropout pattern matches a real one carries more information about that individual. We make no privacy claim and ran no membership-inference or re-identification analysis; both should be carried out before synthetic cohorts produced by this model are shared or used in place of real data. All six benchmark cohorts are public, de-identified and used under their original release terms, and no new human-subjects data was collected. We note also that the identifiability of $p(C \mid X)$ rests on censoring-at-random (§1), which is untestable from observed data; a model deployed where that assumption fails would misstate risk in ways this evaluation cannot detect.

AI USE STATEMENT

In this work, generative AI tools were used to assist with writing and editing the manuscript text. They were also used for retrieval and discovery, to find candidate related work, which the authors then read and assessed. They were not used to generate research ideas, the theoretical results, or any reported number. A Paper Assistant Tool (PAT) was used to review the manuscript before submission. We take responsibility for the final content of this work, including text, claims and artefacts produced with the aid of generative AI.

REFERENCES

Takuya Akiba, Shotaro Sano, Toshihiko Yanase, Takeru Ohta, and Masanori Koyama. Optuna: A next-generation hyperparameter optimization framework. *arXiv preprint arXiv:1907.10902*, July 2019.

Barry C. Arnold and S. James Press. Compatible conditional distributions. *Journal of the American Statistical Association*, 84(405):152–156, 1989.

Mohd Ashhad and Ricardo Henao. Generating accurate synthetic survival data by conditioning on outcomes. *arXiv preprint arXiv:2405.17333*, May 2024.

Yoav Benjamini and Yosef Hochberg. Controlling the false discovery rate: A practical and powerful approach to multiple testing. *Journal of the Royal Statistical Society. Series B (Methodological)*, 57(1):289–300, 1995. ISSN 00359246. URL http://www.jstor.org/stable/2346101.

Herm Jan Brascamp and Elliott H Lieb. On extensions of the brunn-minkowski and prékopa-leindler theorems, including inequalities for log concave functions, and with an application to the diffusion equation. *Journal of Functional Analysis*, 22(4):366–389, 1976. ISSN 0022-1236. doi: https://doi.org/10.1016/0022-1236(76)90004-5. URL https://www.sciencedirect.com/science/article/pii/0022123676900045.

Marie Brockschmidt, Maresa Schröder, and Stefan Feuerriegel. SurvDiff: A diffusion model for generating synthetic data in survival analysis. *arXiv preprint arXiv:2509.22352*, September 2025.

Wenlin Chen, Mingtian Zhang, Brooks Paige, José Miguel Hernández-Lobato, and David Barber. Diffusive gibbs sampling. *arXiv preprint arXiv:2402.03008*, February 2024.

Florentin Coeurdoux, Nicolas Dobigeon, and Pierre Chainais. Plug-and-Play split gibbs sampler: embedding deep generative priors in bayesian inference. *arXiv preprint arXiv:2304.11134*, April 2023.

David R Cox. Regression models and life-tables. *Journal of the Royal Statistical Society: Series B (Methodological)*, 34(2):187–202, 1972.

Christina Curtis, Sohrab P Shah, Suet-Feung Chin, Gulisa Turashvili, Oscar M Rueda, Mark J Dunning, Doug Speed, Andy G Lynch, Shamith Samarajiwa, Yinyin Yuan, Stefan Gräf, Gavin Ha, Gholamreza Haffari, Ali Bashashati, Roslin Russell, Steven McKinney, METABRIC Group, Anita Langerød, Andrew Green, Elena Provenzano, Gordon Wishart, Sarah Pinder, Peter Watson, Florian Markowetz, Leigh Murphy, Ian Ellis, Arnie Purushotham, Anne-Lise Børresen-Dale, James D Brenton, Simon Tavaré, Carlos Caldas, and Samuel Aparicio. The genomic and transcriptomic architecture of 2,000 breast tumours reveals novel subgroups. *Nature*, 486(7403): 346–352, April 2012.

Angela Dispenzieri, Jerry A Katzmann, Robert A Kyle, Dirk R Larson, Terry M Therneau, Colin L Colby, Raynell J Clark, Graham P Mead, Shaji Kumar, L Joseph Melton, 3rd, and S Vincent Rajkumar. Use of nonclonal serum immunoglobulin free light chains to predict overall survival in the general population. *Mayo Clin. Proc.*, 87(6):517–523, June 2012.

Bradley Efron. Tweedie's formula and selection bias. *J. Am. Stat. Assoc.*, 106(496):1602–1614, 2011.

J A Foekens, H A Peters, M P Look, H Portengen, M Schmitt, M D Kramer, N Brünner, F Jänicke, M E Meijer-van Gelder, S C Henzen-Logmans, W L van Putten, and J G Klijn. The urokinase system of plasminogen activation and prognosis in 2780 breast cancer patients. *Cancer Res.*, 60 (3):636–643, February 2000.

Ali Hossein Foomani Gharari, Michael Cooper, Russell Greiner, and Rahul G. Krishnan. Copula-based deep survival models for dependent censoring. In *Proceedings of the Thirty-Ninth Conference on Uncertainty in Artificial Intelligence*, volume 216 of *Proceedings of Machine Learning Research*, pp. 669–680, 2023.

Ian J. Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio. Generative adversarial networks, 2014. URL `https://arxiv.org/abs/1406.2661`.

E Graf, C Schmoor, W Sauerbrei, and M Schumacher. Assessment and comparison of prognostic classification schemes for survival data. *Stat. Med.*, 18(17-18):2529–2545, 1999.

Scott M. Hammer, Kathleen E. Squires, Michael D. Hughes, Janet M. Grimes, Lisa M. Demeter, Judith S. Currier, Joseph J. Eron, Judith E. Feinberg, Henry H. Balfour, Lawrence R. Deyton, Jeffrey A. Chodakewitz, and Margaret A. Fischl. A controlled trial of two nucleoside analogues plus indinavir in persons with human immunodeficiency virus infection and cd4 cell counts of 200 per cubic millimeter or less. aids clinical trials group 320 study team. *The New England journal of medicine*, 337(11):725–733, 1997. URL `https://api.semanticscholar.org/CorpusID:40754467`.

J. Ho, A. Jain, and P. Abbeel. Denoising diffusion probabilistic models. In *Advances in Neural Information Processing Systems*, volume 33, 2020.

Torsten Hothorn, Peter Bühlmann, Sandrine Dudoit, Annette Molinaro, and Mark J van der Laan. Survival ensembles. *Biostatistics*, 7(3):355–373, July 2006.

Hemant Ishwaran, Udaya B Kogalur, Eugene H Blackstone, and Michael S Lauer. Random survival forests. *Ann. Appl. Stat.*, 2(3):841–860, September 2008.

Jared Katzman, Uri Shaham, Jonathan Bates, Alexander Cloninger, Tingting Jiang, and Yuval Kluger. DeepSurv: Personalized treatment recommender system using a cox proportional hazards deep neural network. *arXiv preprint arXiv:1606.00931*, June 2016.

Diederik P Kingma and Max Welling. Auto-encoding variational Bayes. In *International Conference on Learning Representations*, 2014.

W A Knaus, F E Harrell, Jr, J Lynn, L Goldman, R S Phillips, A F Connors, Jr, N V Dawson, W J Fulkerson, Jr, R M Califf, N Desbiens, P Layde, R K Oye, P E Bellamy, R B Hakim, and D P Wagner. The SUPPORT prognostic model. objective estimates of survival for seriously ill hospitalized adults. study to understand prognoses and preferences for outcomes and risks of treatments. *Ann. Intern. Med.*, 122(3):191–203, February 1995.

Akim Kotelnikov, Dmitry Baranchuk, Ivan Rubachev, and Artem Babenko. TabDDPM: Modelling tabular data with diffusion models. *arXiv preprint arXiv:2209.15421*, September 2022.

Dmitrii Kriukov, Ekaterina Kuzmina, Evgeniy Efimov, Dmitry V Dylov, and Ekaterina E Khrameeva. Epistemic uncertainty challenges aging clock reliability in predicting rejuvenation effects. *Aging Cell*, 23(11):e14283, 2024.

Dmitrii Kriukov, Evgeniy Efimov, Ekaterina Kuzmina, Anastasiia Dudkovskaia, Ekaterina E Khrameeva, and Dmitry V Dylov. Computagebench: Epigenetic aging clocks benchmark. In *Proceedings of the 31st ACM SIGKDD Conference on Knowledge Discovery and Data Mining V. 2*, pp. 5560–5570, 2025.

Håvard Kvamme, Ørnulf Borgan, and Ida Scheel. Time-to-event prediction with neural networks and cox regression. *arXiv preprint arXiv:1907.00825*, July 2019.

Changhee Lee, William R. Zame, Jinsung Yoon, and Mihaela van der Schaar. Deephit: a deep learning approach to survival analysis with competing risks. In *Proceedings of the Thirty-Second AAAI Conference on Artificial Intelligence and Thirtieth Innovative Applications of Artificial Intelligence Conference and Eighth AAAI Symposium on Educational Advances in Artificial Intelligence*, AAAI'18/IAAI'18/EAAI'18. AAAI Press, 2018. ISBN 978-1-57735-800-8.

Jun S. Liu, Wing Hung Wong, and Augustine Kong. Covariance structure of the Gibbs sampler with applications to the comparisons of estimators and augmentation schemes. *Biometrika*, 81 (1):27–40, 1994.

A. Yu. Mitrophanov. Sensitivity and convergence of uniformly ergodic Markov chains. *Journal of Applied Probability*, 42(4):1003–1014, 2005.

Radford M. Neal and Geoffrey E. Hinton. *A view of the EM algorithm that justifies incremental, sparse, and other variants*, pp. 355–368. MIT Press, Cambridge, MA, USA, 1999. ISBN 0262600323.

Alexander Norcliffe, Bogdan Cebere, Fergus Imrie, Pietro Lio, and Mihaela van der Schaar. SurvivalGAN: Generating time-to-event data for survival analysis. *arXiv preprint arXiv:2302.12749*, February 2023.

Norberto Pantoja-Galicia, Olivia I Okereke, Deborah Blacker, and Rebecca A Betensky. Concordance measures and time-dependent ROC methods. *Biostat. Epidemiol.*, 5(2):232–249, May 2021.

George Papamakarios, Eric Nalisnick, Danilo Jimenez Rezende, Shakir Mohamed, and Balaji Lakshminarayanan. Normalizing flows for probabilistic modeling and inference. *Journal of Machine Learning Research*, 22(57):1–64, 2021.

Marc-Oliver Pohle. The murphy decomposition and the calibration-resolution principle: A new perspective on forecast evaluation. *arXiv preprint arXiv:2005.01835*, May 2020.

Danilo Jimenez Rezende and Shakir Mohamed. Variational inference with normalizing flows. In *Proceedings of the 32nd International Conference on Machine Learning*, pp. 1530–1538, 2015.

Gareth O. Roberts and Adrian F. M. Smith. Simple conditions for the convergence of the Gibbs sampler and Metropolis–Hastings algorithms. *Stochastic Processes and their Applications*, 49 (2):207–216, 1994.

Daniel Rudolf and Nikolaus Schweizer. Perturbation theory for Markov chains via Wasserstein distance. *Bernoulli*, 24(4A):2610–2639, 2018.

M Schumacher, G Bastert, H Bojar, K Hübner, M Olschewski, W Sauerbrei, C Schmoor, C Beyerle, R L Neumann, and H F Rauschecker. Randomized 2 x 2 trial evaluating hormonal treatment and the duration of chemotherapy in node-positive breast cancer patients. german breast cancer study group. *J. Clin. Oncol.*, 12(10):2086–2093, October 1994.

Juntong Shi, Minkai Xu, Harper Hua, Hengrui Zhang, Stefano Ermon, and Jure Leskovec. TabDiff: A mixed-type diffusion model for tabular data generation. *arXiv preprint arXiv:2410.20626*, October 2024.

Jiaming Song, Arash Vahdat, Morteza Mardani, and Jan Kautz. Pseudoinverse-guided diffusion models for inverse problems. In *International Conference on Learning Representations*, 2023. URL https://openreview.net/forum?id=9_gsMA8MRKQ.

Yang Song and Stefano Ermon. Generative modeling by estimating gradients of the data distribution. *arXiv preprint arXiv:1907.05600*, July 2019.

Yang Song, Jascha Sohl-Dickstein, Diederik P Kingma, Abhishek Kumar, Stefano Ermon, and Ben Poole. Score-based generative modeling through stochastic differential equations. *arXiv preprint arXiv:2011.13456*, November 2020.

Yang Song, Conor Durkan, Iain Murray, and Stefano Ermon. Maximum likelihood training of score-based diffusion models. *arXiv preprint arXiv:2101.09258*, January 2021.

Viktor Stojkoski, Trifce Sandev, Lasko Basnarkov, Ljupco Kocarev, and Ralf Metzler. Generalised geometric brownian motion: Theory and applications to option pricing. *Entropy (Basel)*, 22(12): E1432, December 2020.

Martin A Tanner and Wing Hung Wong. The calculation of posterior distributions by data augmentation. *J. Am. Stat. Assoc.*, 82(398):528–540, June 1987.

Luke Tierney. Markov chains for exploring posterior distributions. *Ann. Stat.*, 22(4):1701–1728, December 1994.

Anastasios Tsiatis. A nonidentifiability aspect of the problem of competing risks. *Proceedings of the National Academy of Sciences*, 72(1):20–22, 1975.

Hajime Uno, Tianxi Cai, Michael J Pencina, Ralph B D'Agostino, and L J Wei. On the C-statistics for evaluating overall adequacy of risk prediction procedures with censored survival data. *Statistics in Medicine*, 30(10):1105–1117, 2011.

Pascal Vincent. A connection between score matching and denoising autoencoders. *Neural Comput.*, 23(7):1661–1674, July 2011.

Lei Xu, Maria Skoularidou, Alfredo Cuesta-Infante, and Kalyan Veeramachaneni. Modeling tabular data using conditional GAN. *arXiv preprint arXiv:1907.00503*, July 2019.

Weijia Zhang, Chun Kai Ling, and Xuanhui Zhang. Deep copula-based survival analysis for dependent censoring with identifiability guarantees. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 38, pp. 20613–20621, 2024.

APPENDIX A ASSUMPTIONS

$V_{\text{prior}}, g_3, g_4, \Sigma$ and $u^\star$ are as in Definition 1. Theorem 1 assumes (A1)–(A5) and Theorem 2 adds (A6); (A0) is what separates the two channels in §3.2 rather than a hypothesis of either proof.

- **(A0)** *(CAR.)* $T^\star \perp C \mid X$.
- **(A1)** $z_0 \mapsto \log p_\theta(z_0 \mid x)$ is C^4 near the posterior mode.
- **(A2)** *(Non-degenerate mode.)* The mode $z_0^\dagger$ of q_s is unique and $\Sigma > 0$, that is $1+\sigma_{\text{post}}^2 V_{\text{prior}} > 0$. The expansions use the curvature of q_s, which is $V_{\text{prior}} + \sigma_{\text{post}}^{-2}$, so $V_{\text{prior}} < 0$ is admissible. The condition holds with probability at least $1 - \alpha_2$ over the states the chain visits, and Appendix D.3 measures $\alpha_2 \in [0, 0.040]$ over six cohorts and three seeds. The proofs charge that exception as they charge (A5)'s, through $\|\cdot\|_{\text{TV}} \le 1$.
- **(A3)** $\bar{g}_3 := \sup_{z_0,x} |\partial_{z_0}^3 \log p_\theta| < \infty$. (A1) gives finiteness at the mode only; Corollary 1 needs it uniformly in the state.
- **(A4)** $L_4(x) := \sup |\partial_{z_0}^4 \log p_\theta| < \infty$.
- **(A5)** *(Interior mode.)* For every x and every $z_0 \ge \log T^{\text{obs}}$, $u^\star \ge u_- > 0$ with probability $\ge 1 - \alpha$ over $Z_s \sim p_{s_{\text{imp}}|0}(\cdot \mid z_0)$. The probability is conditional on z_0 because Theorem 2(i) takes a supremum over it. On that event the score error is bounded, $\sup |\delta_\theta| < \infty$, which is what makes the $\bar{\mathcal{K}}$ of Theorem 2(i) finite.
- **(A6)** *(Uniform curvature.)* $|\partial_{z_0}^2 \log p_\theta(z_0 \mid x)| \le \bar{V} < \infty$ for all $z_0 \ge \log T^{\text{obs}}$ and all x. Two-sided because the slack carries V_{prior}^2; (A2) bounds the negative side below by $-\sigma_{\text{post}}^{-2}$.

(A7) appears in Appendix B.7 and no theorem uses it. Two hypotheses are assumed nowhere above: a lower curvature bound $V_{\text{prior}} \ge \underline{V} > 0$ and a contraction condition on $\tilde{P}$ (Remark 4).

APPENDIX B PROOFS

Throughout, z_0 denotes the latent log-time of the channel under analysis, $a := \log T^{\text{obs}}$ the truncation boundary, ϕ and Φ the standard normal density and CDF, $\bar{\Phi} := 1 - \Phi$, and $h := \phi/\bar{\Phi}$ the normal hazard. We write $m := \phi/\Phi$ for the Mills ratio of Definition 1 and $\sigma_{\text{post}}^2(s) = \sigma^2 s$ for the forward variance of equation 2.

B.1 THE GBM LOSS IS AN EXACT-CHANNEL NLL BOUND

Lemma 1 (GBM DSM as exact-channel NLL bound)**.** *Under the GBM forward kernel, the ε-parameterisation, log-uniform sampling on $[s_{\min}, s_{\max}]$ with $s_{\min} > 0$, and square-integrable score $\mathbb{E}[\|\varepsilon_\theta\|^2] < \infty$ at each s, the bound equation 4 holds with $L = \log(s_{\max}/s_{\min})$ and C independent of θ. It is prior-agnostic and transfers to $-\log p_\theta(z_0 \mid X)$ as $s_{\min} \to 0$ up to an $O(s_{\min})$ reconstruction term.*

Proof. The argument has three steps: the path-space bound, the exact absorption of the GBM weight, and the transfer to the unsmoothed likelihood.

Step 1 (path-space bound). For a forward SDE $\mathrm{d}z_s = f(z_s, s)\,\mathrm{d}s + g(s)\,\mathrm{d}W_s$ with marginals p_s, the Girsanov change of measure between the true and the learned reverse process gives, for any $s_{\min} > 0$,

$$\mathbb{E}_{z_0}\big[-\log p_\theta^{s_{\min}}(z_0 \mid x)\big] \;\le\; \tfrac{1}{2}\int_{s_{\min}}^{s_{\max}} g(s)^2\, \mathbb{E}_{z_s}\big\|s_\theta(z_s, s, x) - \nabla_{z_s} \log p_s(z_s \mid x)\big\|^2\,\mathrm{d}s + C, \quad (12)$$

with C collecting the terminal cross-entropy at $s_{\max}$ and the θ-independent entropy of the forward path measure (Song et al., 2021). Denoising score matching replaces the intractable $\nabla \log p_s$ by the conditional score at the cost of a θ-independent constant (Vincent, 2011), so the integrand may be taken to be the conditional DSM residual.

Step 2 (the GBM weight is exactly log-uniform). The kernel equation 2 is the law of a constant-coefficient Brownian motion in log-space, $\mathrm{d}z_s = \mu_{\text{drift}}\,\mathrm{d}s + \sigma\,\mathrm{d}W_s$, so $g(s) \equiv \sigma$ and $\sigma_{\text{post}}^2(s) =$

$\sigma^2 s$. Under the ε-parameterisation $s_\theta = -\varepsilon_\theta/\sigma_{\text{post}}$ the residual in equation 12 carries a factor $\sigma_{\text{post}}^{-2}$, so the weight is

$$\frac{g(s)^2}{\sigma_{\text{post}}^2(s)} \;=\; \frac{\sigma^2}{\sigma^2 s} \;=\; \frac{1}{s}. \tag{13}$$

The log-uniform law on $[s_{\min}, s_{\max}]$ has density $1/(sL)$ with $L = \log(s_{\max}/s_{\min})$, so for any integrable f

$$\int_{s_{\min}}^{s_{\max}} \frac{f(s)}{s}\,\mathrm{d}s \;=\; L \cdot \mathbb{E}_{s\sim\text{logunif}}\big[f(s)\big],$$

an identity, not an approximation: the sampling density cancels the weight pointwise. Applying it to equation 12 with $f(s) = \mathbb{E}\|\varepsilon_\theta - \varepsilon\|^2$ turns the weighted integral into $L \cdot \mathcal{L}_{\text{DSM}}^{\text{exact}}(\theta)$ with $\mathcal{L}_{\text{DSM}}^{\text{exact}} = \frac{1}{2}\mathbb{E}[(\varepsilon_\theta - \varepsilon)^2]$ *unweighted*. The integrand is non-negative and jointly measurable in (s, z_s), so Tonelli licenses the interchange of $\int \mathrm{d}s$ and $\mathbb{E}$ with no integrability hypothesis; what $\mathbb{E}\|\varepsilon_\theta\|^2 < \infty$ at each s buys is that the resulting bound, and the θ-independent constant separating DSM from explicit score matching, are finite.

Step 3 (removing the smoothing). This step needs the square-integrability hypothesis uniformly on $(0, s_0]$ rather than at each s separately, and it is the one step of the theorem that is asymptotic rather than exact. $p_\theta^{s_{\min}}$ is the model marginal at noise level $s_{\min}$, i.e. $p_\theta * \mathcal{N}(\mu_{\text{drift}} s_{\min}, \sigma^2 s_{\min})$. If $\sup_{s\le s_0}\mathbb{E}\|\varepsilon_\theta\|^2 < \infty$ for some $s_0 > 0$, then $\mathbb{E}|z_{s_{\min}} - z_0|^2 = O(s_{\min})$ and the reconstruction gap $\mathbb{E}[-\log p_\theta(z_0 \mid x)] - \mathbb{E}[-\log p_\theta^{s_{\min}}(z_0 \mid x)]$ is $O(s_{\min})$ by the same estimate used for the terminal term of a discretised diffusion ELBO. Hence the bound transfers as $s_{\min} \to 0$ up to $O(s_{\min})$. The statement is prior-agnostic because no step used any property of $p_\theta(\cdot \mid x)$ beyond square-integrability of its score. □

B.2 Theorem 1: the proposal slack at q^{work}

Write $\xi := \sigma_{\text{post}}^2 V_{\text{prior}}$ and recall $\Sigma = \sigma_{\text{post}}^2/(1+\xi)$. Let $r := \mathcal{N}(z_0^\dagger, \Sigma)|_{\ge a}$ denote the Laplace approximation to q_s at its mode. The proof decomposes

$$\mathrm{KL}(q^{\text{work}}\|q_s) \;=\; \underbrace{\mathrm{KL}(q^{\text{work}}\|r)}_{\text{(a)+(b)}} \;+\; \underbrace{\mathbb{E}_{q^{\text{work}}}\big[\log(r/q_s)\big]}_{\text{(c)}}, \tag{14}$$

which is the chain rule for the log-ratio and is exact.

Proof. All three terms are evaluated on the interior-mode event $\{u^\star \ge u_-\}$, of probability at least $1-\alpha$ by (A5).

Step 0 (Laplace). By (A1) the map $z_0 \mapsto \log p_\theta(z_0 \mid x)$ is C^4 near $z_0^\dagger$. The Gaussian factor of $q_s \propto p_\theta(z_0 \mid x)\, p_{s|0}(z_s \mid z_0)\mathbf{1}\{z_0 \ge a\}$ contributes exactly $-\frac{1}{2}\sigma_{\text{post}}^{-2}(z_0 - \cdot)^2$, so the curvature at the mode is $V_{\text{prior}} + \sigma_{\text{post}}^{-2} = \Sigma^{-1}$, and by (A4) the remainder is $O(\Sigma^{3/2} g_3) + O(\Sigma^2 g_4)$, uniformly on that event.

Step (a): variance mismatch. For two normals of common location ℓ truncated to $[a, \infty)$, with $\alpha_p := (a-\ell)/\sigma_p$, $\alpha_q := (a-\ell)/\sigma_q$ and $\mathbb{E}[\eta^2 \mid \eta \ge \alpha] = 1 + \alpha h(\alpha)$ for $\eta \sim \mathcal{N}(0,1)$,

$$\mathrm{KL} = \log\frac{\sigma_q}{\sigma_p} + \log\frac{\bar\Phi(\alpha_q)}{\bar\Phi(\alpha_p)} + \frac{\sigma_p^2\big(1 + \alpha_p h(\alpha_p)\big)}{2}\Big(\frac{1}{\sigma_q^2} - \frac{1}{\sigma_p^2}\Big). \tag{15}$$

Set $\sigma_p^2 = \sigma_{\text{post}}^2$ and $\sigma_q^2 = \Sigma = \sigma_{\text{post}}^2/(1+\xi)$ and expand in ξ. The linear term vanishes, equation 15 being a KL divergence, hence non-negative, with a zero at $\xi = 0$. Writing $u^\star = (z_0^\dagger - a)/\sqrt{\Sigma}$, so that $\alpha_p \to -u^\star$ and $\bar\Phi(\alpha_p) \to \Phi(u^\star)$, and $m = \phi/\Phi$, the quadratic coefficient is

$$C_{\text{var}}(u^\star) \;=\; \tfrac{1}{4} - \tfrac{u^\star m(u^\star)}{8}\big[1 + u^\star\big(m(u^\star) + u^\star\big)\big].$$

Since $\xi = \sigma_{\text{post}}^2 V_{\text{prior}}$ the quadratic term is $C_{\text{var}}\sigma_{\text{post}}^4 V_{\text{prior}}^2$, and the ξ^3, ξ^4 terms give the V_{prior}^3, V_{prior}^4 parts of C_1, C_2.

Step (b): mean mismatch. q^{work} is centred at the Tweedie mean $\hat{z}_\theta$ of equation 3 as the *network* computes it, r at $z_0^\dagger$. Split at the exact posterior mean:

$$\hat{z}_\theta - z_0^\dagger \;=\; \underbrace{\big(\mathbb{E}_{q_s}[z_0 \mid z_s, x] - z_0^\dagger\big)}_{\text{Laplace}} \;+\; \underbrace{\big(\hat{z}_\theta - \mathbb{E}_{q_s}[z_0 \mid z_s, x]\big)}_{=-\sigma_{\mathrm{post}}\delta_\theta \text{ by Tweedie}}.$$

On the interior-mode event the constrained and unconstrained modes coincide and the Laplace expansion of q_s about its mode gives $\mathbb{E}_{q_s}[z_0] - z_0^\dagger = \frac{1}{2}g_3\Sigma^2 + O(\Sigma^{5/2})$, which is the first moment of the Hermite expansion Step (c) uses. A location shift Δ between truncated normals of common scale σ_{post} contributes $\Delta^2/(2\sigma_{\mathrm{post}}^2)$ to the KL at leading order, so

$$\frac{\Delta^2}{2\sigma_{\mathrm{post}}^2} \;=\; \underbrace{O(\sigma_{\mathrm{post}}^6 g_3^2)}_{\to C_1} \;+\; \underbrace{\tfrac{1}{2}\delta_\theta^2 + O(\sigma_{\mathrm{post}}^3 |g_3|\,\delta_\theta)}_{\text{score error}}.$$

The second group is $O(1)$ in σ_{post}: Tweedie's two powers of σ_{post} cancel against the $\sigma_{\mathrm{post}}^{-2}$ of the Gaussian KL. Setting $\hat{z}_\theta = \mathbb{E}_{q_s}[z_0 \mid z_s, x]$ would remove it; $\mathcal{K}$ stands in the statement instead (Remark 4).

Step (c): shape mismatch. With $a_3 = g_3\Sigma^{3/2}/6$ and η the Σ-standardised coordinate, $q_s \propto \phi_\Sigma\,[1 + a_3\mathrm{He}_3(\eta) + O(a_3^2)]$, and orthogonality of the Hermite polynomials against the Gaussian measure gives

$$\mathbb{E}_{q^{\mathrm{work}}}\big[\log(r/q_s)\big] \;=\; \frac{g_3^2\Sigma^3}{12} \;+\; O(\sigma_{\mathrm{post}}^8 g_4^2 + \sigma_{\mathrm{post}}^{12} g_3^4),$$

i.e. $O(\sigma_{\mathrm{post}}^6 g_3^2)$ at leading order, inside C_1, with the g_4^2 and g_3^4 cross terms in C_2. The support $[a, \infty)$ costs two corrections. The scalar $1/12$ is the $u^\star \to \infty$ value, since $\int_{-u^\star}^{\infty} \mathrm{He}_3(\eta)\phi(\eta)\,\mathrm{d}\eta = ((u^\star)^2 - 1)\phi(u^\star) \neq 0$; at finite $u^\star$ it is a bounded $u^\star$-dependent coefficient. Against the first moment the expectation is under q^{work}, of scale σ_{post}, and not under r, of scale Σ, and $\sigma_{\mathrm{post}}^2 - \Sigma = \sigma_{\mathrm{post}}^4 V_{\mathrm{prior}} + O(\sigma_{\mathrm{post}}^6 V_{\mathrm{prior}}^2)$, so skew and variance mismatch couple at $O(\sigma_{\mathrm{post}}^5 |g_3| V_{\mathrm{prior}})$: subdominant to $C_{\mathrm{var}}\sigma_{\mathrm{post}}^4 V_{\mathrm{prior}}^2$ when $\sigma_{\mathrm{post}}|g_3| \ll V_{\mathrm{prior}}$, one order slower than the $O(\sigma_{\mathrm{post}}^6)$ terms, hence in C_1. Theorem 3 removes it, $q^{\mathrm{work},\Sigma}$ and r sharing the scale Σ.

Collecting. At leading order (b) carries no V_{prior} and (c) carries it only through the cross term, linearly. So as $\xi \to 0$ at fixed g_3 term (a) dominates and the bound of the theorem follows. □

Remark 1 (Two limits of Theorem 1, both measured). *Scope.* Steps (a)–(c) hold on $\{u^\star \geq u_-\}$; off that event the mode sits at or below the boundary and the theorem says nothing. The bound is not uniform in z_s, so a downstream use must average over z_s or pay α. Theorem 2 averages, which is why it is stated in total variation; a supremum over z_s, which a W_2 route needs, is not available here.

Regime. The statement is an expansion in $\xi = \sigma_{\mathrm{post}}^2 V_{\mathrm{prior}}$ asserted as $\xi \to 0$. On the fitted models ξ spans 0.02 to 1.8 over the grid and 0.05 to 0.35 at the selected operating points (§5.3), where the $O(\xi^3)$ and $O(\xi^4)$ remainders are tens of percent of the leading term. The rate holds to 0.23 in log-log slope and to 0.14 in four cohorts (Fig. 6); we do not claim the constant.

Remark 2 (A location term the expansion of Step (b) does not carry). Step (b) expands the mean of the *unconstrained* posterior about its mode, where the gap is $\frac{1}{2}g_3\Sigma^2$. Under truncation the constrained mean sits $\sqrt{\Sigma}\,m(u^\star)$ above that mode, one order larger in σ_{post}, and a location gap enters the divergence divided by σ_{post}^2, so part of it carries no power of σ_{post}.

On the interior-mode event q^{work} and r are normals of common scale truncated at the same a, so their divergence is exact. With d the standardised gap between the centres,

$$\begin{aligned} \mathrm{KL} \;&=\; \frac{d^2}{2} + d\,h(\alpha_p) + \log\frac{\bar{\Phi}(\alpha_q)}{\bar{\Phi}(\alpha_p)} \;=\; \mathcal{K}(u^\star, d), \\ \mathcal{K}(u, d) \;&:=\; \frac{d^2}{2} + d\,m(u+d) + \log\frac{\Phi(u)}{\Phi(u+d)}, \end{aligned} \tag{16}$$

with $d = m(u^\star) - \delta_\theta$ to leading order, since the constrained mean stands $m(u^\star)$ scale units above the mode and Tweedie places the network a further $-\delta_\theta$ away. $\mathcal{K}$ is free of σ_{post}. It reduces to $\frac{1}{2}\delta_\theta^2$,

the score-error cost alone, as $u^\star \to \infty$, and to zero when the score is exact and the boundary is far. At a fixed state $u^\star$ grows like $s_{\mathrm{imp}}^{-1/2}$, so with an exact score $\mathcal{K}$ vanishes superexponentially along the schedule and equation 10 holds as stated; with $\delta_\theta \neq 0$ it settles at $\frac{1}{2}\delta_\theta^2$ instead. (A5) in standardised units also admits states where $u^\star$ stays bounded as s_{imp} falls, and there $\mathcal{K}$ is the leading contribution and equation 10 is not tight.

The fitted models sit in the second regime. Inverting the truncation factor of Appendix D.2 at each cohort's selected s_{imp} gives $u^\star$ from 0.65 on FLCHAIN to 1.64 on GBSG2, where $\mathcal{K}(u^\star, m(u^\star))$ is 1.3 to 13.6 times $C_{\mathrm{var}}\sigma_{\mathrm{post}}^4 V_{\mathrm{prior}}^2$. The contribution Theorem 3 removes is therefore the smaller of the two at every cohort, which is the reading §5.3 reports for the correction. Those $u^\star$ come from a per-state average inverted through a nonlinear map, so they locate the regime and not the constant.

B.3 Lemma 2: global behaviour of C_{var}

Lemma 2 (Global bound on C_{var})**.** *The function $C_{\mathrm{var}} : \mathbb{R} \to (0, \frac{1}{2})$ defined by equation 11 is continuous and bounded with $C_{\mathrm{var}}(0) = \frac{1}{4}$, $C_{\mathrm{var}}(u^\star) \to \frac{1}{4}$ as $u^\star \to +\infty$, $C_{\mathrm{var}}(u^\star) \to \frac{1}{2}$ as $u^\star \to -\infty$, and $\sup_{u^\star \in \mathbb{R}} C_{\mathrm{var}}(u^\star) = \frac{1}{2}$.*

Proof. Continuity is immediate: $m = \phi/\Phi$ is continuous and $\Phi > 0$ on $\mathbb{R}$. At $u^\star = 0$ the bracket multiplies $u^\star = 0$, leaving $C_{\mathrm{var}}(0) = \frac{1}{4}$. As $u^\star \to +\infty$, $m(u^\star) \to 0$ faster than any polynomial, so the correction term vanishes and $C_{\mathrm{var}} \to \frac{1}{4}$. As $u^\star \to -\infty$ put $t := -u^\star \to +\infty$ and use the lower-tail Mills expansion $\Phi(-t) = \phi(t)\big(t^{-1} - t^{-3} + 3t^{-5} - \cdots\big)$, whence

$$m(u^\star) = t + t^{-1} - 2t^{-3} + O(t^{-5}), \qquad u^\star m = -t^2 - 1 + O(t^{-2}),$$
$$1 + u^\star\big(m + u^\star\big) = 2t^{-2} + O(t^{-4}),$$

so the correction term tends to $(-t^2)(2t^{-2})/8 = -\frac{1}{4}$ and $C_{\mathrm{var}} \to \frac{1}{4} + \frac{1}{4} = \frac{1}{2}$. Non-negativity holds because C_{var} is the leading Taylor coefficient of the non-negative quantity equation 15; that argument gives $C_{\mathrm{var}} \geq 0$ and no more; numerically the infimum is 0.1591, near $u^\star = 1.377$. The supremum $\frac{1}{2}$ is approached but not attained, and only that upper bound is used below. □

B.4 Theorem 2: consistency of the imputation chain

The Gibbs joint is $\pi(z_0, z_s) \propto p_\theta(z_0 \mid x)\mathbf{1}\{z_0 \geq a\}\, p_{s|0}(z_s \mid z_0)$, with conditionals $z_s \mid z_0 \sim p_{s|0}$ and $z_0 \mid z_s \sim q_s$. Marginalising z_s gives the ideal kernel $P(z_0, \cdot) = \int p_{s|0}(z_s \mid z_0) q_s(\cdot \mid z_s)\, \mathrm{d}z_s$ and, with q^{work} in place of q_s, the working kernel $\tilde{P}$.

Proof. $q^\star$ is invariant for P by marginalisation of π (Tanner & Wong, 1987), and is a probability measure because $p_\theta(\cdot \mid x)$ is a density with positive mass on $[a, \infty)$.

Part (i): the algorithmic error. Write $A := P$, $B := \tilde{P}$ and use the telescoping identity

$$A^n - B^n = \sum_{k=0}^{n-1} A^k\,(A - B)\,B^{\,n-1-k}, \tag{17}$$

valid in the algebra of bounded operators on signed measures. With $\mu_k := \mu A^k$, the triangle inequality and the contraction $\|\lambda B^m\|_{\mathrm{TV}} \leq \delta(B)^m \|\lambda\|_{\mathrm{TV}}$ give

$$\|\mu A^n - \mu B^n\|_{\mathrm{TV}} \;\leq\; \sum_{k=0}^{n-1} \delta(B)^{\,n-1-k} \sup_{z_0} \big\|A(z_0, \cdot) - B(z_0, \cdot)\big\|_{\mathrm{TV}},$$

with δ the Dobrushin coefficient. The mirror identity $B^n - A^n = \sum_k B^k(B - A)A^{\,n-1-k}$ would put the *ideal* kernel under the outer operator and so need Lipschitz constants we do not bound. The prefactor $\delta(B) \leq 1$ is the only step available, for the reason Remark 3 gives.

For the one-step distance, P and $\tilde{P}$ share the mixing weight $p_{s|0}(\cdot \mid z_0)$ and total variation is jointly convex, so

$$\big\|P(z_0, \cdot) - \tilde{P}(z_0, \cdot)\big\|_{\mathrm{TV}} \;\leq\; \mathbb{E}_{z_s|z_0}\big\|q_s(\cdot \mid z_s) - q^{\mathrm{work}}(\cdot \mid z_s)\big\|_{\mathrm{TV}} \;=:\; \bar{\varepsilon}_{\mathrm{TV}}(z_0),$$

so that $\bar{\varepsilon}_{\mathrm{TV}} = \sup_{z_0} \bar{\varepsilon}_{\mathrm{TV}}(z_0)$ is a supremum over z_0 of an *average* over z_s and never of a supremum over it, which is what makes the theorem compatible with Theorem 1 (Remark 1). On the interior-mode event, Pinsker turns the per-z_s bound of that theorem into

$$\left\|q^{\mathrm{work}} - q_s\right\|_{\mathrm{TV}} \;\le\; \sqrt{\tfrac{1}{2}\mathrm{KL}} \;\le\; \sqrt{\tfrac{\bar{C}_{\mathrm{var}}}{2}}\,\sigma^2_{\mathrm{post}}\bar{V} \;+\; \sqrt{\tfrac{\bar{\mathcal{K}}}{2}} \;+\; O(\sigma^3_{\mathrm{post}}),$$

with $\bar{C}_{\mathrm{var}} = \sup_u C_{\mathrm{var}}(u) = \frac{1}{2}$ from Lemma 2, the state entering through $u^\star$, and $\bar{V}$ from (A6). Two steps split the root, and the first is not enough on its own. Subadditivity carries the leading term and the $O(\sigma^6_{\mathrm{post}})$ and $O(\sigma^8_{\mathrm{post}})$ remainders of Theorem 1 at $O(\sigma^3_{\mathrm{post}})$ and $O(\sigma^4_{\mathrm{post}})$, but on the $O(\sigma^5_{\mathrm{post}}|g_3|V_{\mathrm{prior}})$ cross term it leaves $O(\sigma^{5/2}_{\mathrm{post}})$, which is the larger. Concavity of the root at the leading term gives $\sigma^5_{\mathrm{post}}|g_3|V_{\mathrm{prior}}/(2\sqrt{C_{\mathrm{var}}}\,\sigma^2_{\mathrm{post}}|V_{\mathrm{prior}}|) = O(\sigma^3_{\mathrm{post}}|g_3|)$ instead, where V_{prior} cancels, so the stated order needs no lower bound on the curvature. Taking the supremum over the interior-mode event replaces $\mathcal{K}$ by $\bar{\mathcal{K}}$. Off that event, $\|\cdot\|_{\mathrm{TV}} \le 1$ costs at most α, since (A5) bounds the exceptional event conditionally on z_0. Averaging the two cases gives the stated bound.

Part (ii): the ergodic error. Both conditionals of π have densities positive on their supports, so the data-augmentation chain is $q^\star$-irreducible (Roberts & Smith, 1994). Aperiodicity needs nothing from the denoiser: $P(z_0, \cdot)$ is a mixture of densities positive on all of $[a, \infty)$, so from every z_0 it charges every set of positive Lebesgue measure, while a periodic decomposition would need the kernel to avoid all of its classes but one. Writing $\mathcal{A}f(z_s) := \mathbb{E}_{q_s(\cdot|z_s)}[f]$ and $\mathcal{A}^* g(z_0) := \mathbb{E}_{p_{s|0}(\cdot|z_0)}[g]$, integrating $\langle \mathcal{A}f, g\rangle$ against π makes the two adjoint, so on $L^2(q^\star)$

$$P = \mathcal{A}^*\mathcal{A}, \qquad P = P^*, \qquad \mathrm{spec}(P) \subset [0, 1]$$

(Liu et al., 1994): P is reversible for $q^\star$, and a period-two mode at -1 is excluded whatever $\partial_{z_s}\hat{z}_\theta$ does. A $q^\star$-irreducible chain with invariant probability $q^\star$ is positive recurrent, and Harris recurrent as soon as $P(z_0, \cdot) \ll q^\star$ for *every* z_0 (Tierney, 1994, Cor. 2), which holds pointwise here because $P(z_0, \cdot)$ mixes the densities $q_s(\cdot \mid z_s)$, all positive exactly on $[a, \infty) = \mathrm{supp}\, q^\star$. Hence $\|\nu_0 P^n - q^\star\|_{\mathrm{TV}} \to 0$ from *every* initial law, where irreducibility and aperiodicity alone would give it only for $q^\star$-almost every start, and the E-step starts from a point it chooses. No curvature assumption enters, and no rate is obtained.

Part (iii). By the triangle inequality,

$$\|\nu_0 \tilde{P}^n - q^\star\|_{\mathrm{TV}} \;\le\; n\,\bar{\varepsilon}_{\mathrm{TV}} \;+\; \|\nu_0 P^n - q^\star\|_{\mathrm{TV}},$$

and both terms vanish along the schedule by assumption. With $\bar{\varepsilon}_{\mathrm{TV}} = O(\sigma^2_{\mathrm{post}}\bar{V}) + \alpha$ and $\sigma^2_{\mathrm{post}} = \sigma^2 s_{\mathrm{imp}}$, the first asks for $s_{\mathrm{imp}} = o(1/n_{\mathrm{gibbs}})$ and $\alpha = o(1/n_{\mathrm{gibbs}})$. The second of these is not automatic: α is a constant of (A5), and it shrinks along the schedule for a subject with an interior mode, where $u^\star$ grows like $s_{\mathrm{imp}}^{-1/2}$, but not for one whose constrained mode sits at the boundary. Nor is the second hypothesis: part (ii) holds at fixed s_{imp} and gives no rate, so a rate for P uniform on the schedule, from a Doeblin minorisation or a drift condition, would be needed and we do not have one. The two requirements pull against each other, since part (i) caps the sweeps at $o(1/s_{\mathrm{imp}})$ while the mixing time grows as s_{imp} falls; Appendices D.2 and D.4 measure both sides on these models. □

Remark 3 (What the prefactor n costs, and why it stays). The perturbation bounds that charge a constant in place of n (Mitrophanov, 2005) ask the unperturbed chain to be uniformly ergodic, that is to admit a Doeblin minorisation, and on these checkpoints it does not. Let $M(z) := \mathbb{E}_\xi\big[\hat{z}_\theta\big(z + \mu_{\mathrm{drift}} s_{\mathrm{imp}} + \sigma_{\mathrm{post}}\xi\big)\big]$ be the one-sweep mean map of Appendix D.4. Where the truncation at a is inactive, $\tilde{P}(z_0, \cdot)$ has mean $M(z_0)$ and standard deviation $O(\sigma_{\mathrm{post}})$. Both hold by construction on this architecture: the network reads z only through a sinusoidal embedding (Appendix C.3), so ε_θ is bounded in z and $\hat{z}_\theta(z_s) = z_s - \mu_{\mathrm{drift}} s_{\mathrm{imp}} - \sigma_{\mathrm{post}}\varepsilon_\theta(z_s)$ is the identity plus a bounded Lipschitz term. With $\Delta := |M(z_0) - M(y_0)|$ and $G := [\frac{1}{2}(M(z_0) + M(y_0)), \infty)$, Chebyshev gives $\tilde{P}(z_0, G) \to 1$ and $\tilde{P}(y_0, G) \to 0$ as $\Delta/\sigma_{\mathrm{post}} \to \infty$, which any unbounded M reaches along $|z_0 - y_0| \to \infty$. Throughout, $\|\mu - \nu\|_{\mathrm{TV}} := \sup_A |\mu(A) - \nu(A)|$, the normalisation Pinsker is stated in, so $\delta \le 1$ holds for every kernel and

$$\delta(\tilde{P}) \;=\; \sup_{z_0, y_0} \left\|\tilde{P}(z_0, \cdot) - \tilde{P}(y_0, \cdot)\right\|_{\mathrm{TV}} \;\ge\; \sup_{z_0, y_0} \left|\tilde{P}(z_0, G) - \tilde{P}(y_0, G)\right| \;=\; 1$$

pins it from below, giving $\delta(\tilde{P}) = 1$. A Doeblin constant would need $\hat{z}_\theta$ to saturate on $[a, \infty)$, which the identity term forbids, so nothing here is left to measure. That closes the uniform route and only that one; the weighted routes stay open in principle, and Appendix D closes each on its own hypothesis.

Remark 4 (Hypotheses this proof does *not* use). The natural route assumes a lower curvature bound $V_{\text{prior}} \geq \underline{V} > 0$ and score optimality $\hat{z}_\theta = \mathbb{E}_{q_s}[z_0]$, and turns them, through the Stein identity $\partial_{z_s} \hat{z}_\theta = \mathrm{Var}_{q_s}[z_0]/\sigma^2_{\text{post}}$ and Brascamp–Lieb (Brascamp & Lieb, 1976), into a contraction $\tilde{\rho} < 1$ and a geometric rate; Appendices D.2 and D.3 refute that conjunction and assign the failure to the prior. Parts (i)–(iii) assume neither, but (i) still carries the score error: step (b) of Theorem 1 prices it inside $\mathcal{K}$, which reaches $\bar{\varepsilon}_{\text{TV}}$ through Pinsker as an additive $\sqrt{\mathcal{K}/2}$ and does not shrink with s_{imp} where $u^\star$ stays bounded, the regime Remark 2 puts the fitted models in.

B.5 Theorem 3: the Σ-corrected slack

Proof. By Definition 2 the corrected proposal has variance Σ, which is exactly the Laplace variance of r. In the decomposition equation 14 the pair compared in Step (a) therefore has $\sigma_p^2 = \sigma_q^2$, so $\xi = 0$ in equation 15 and the term vanishes identically, not merely to leading order. Steps (b) and (c) are untouched: neither used the proposal variance beyond its appearance in Σ, and both retain their orders $O(\sigma^6_{\text{post}} g_3^2)$ and $O(\sigma^8_{\text{post}} g_4^2 + \sigma^{12}_{\text{post}} g_3^4)$. Collecting gives the statement. □

B.6 Corollary 1

Proof. Substitute the slack of Theorem 3 for that of Theorem 1 in Part (i) of the proof above. Pinsker then gives $\bar{\varepsilon}^{\Sigma}_{\text{TV}} = O(\sigma^3_{\text{post}} \bar{g}_3) + \sqrt{\mathcal{K}/2} + \alpha$, one order in σ_{post} better than the uncorrected $O(\sigma^2_{\text{post}} \bar{V}) + \alpha$, and the admissible schedule in Part (iii) relaxes from $n_{\text{gibbs}} \sigma^2_{\text{post}} \to 0$ to $n_{\text{gibbs}} \sigma^3_{\text{post}} \to 0$ accordingly. □

Remark 5 (What part (i) hands the M-step, and what it does not). Theorem 2 bounds the law of the imputed draw in total variation, which controls $|\mathbb{E}_{\nu_n} f - \mathbb{E}_{q^\star} f|$ only for bounded f. The M-step integrates the per-sample denoising loss as a function of the imputed z_0, and that is unbounded on $[a, \infty)$. One moment condition closes the gap: if $\mathbb{E}_{\nu_n} |f|^{1+\delta}$ and $\mathbb{E}_{q^\star} |f|^{1+\delta}$ are finite for some $\delta > 0$, Hölder gives

$$\left|\mathbb{E}_{\nu_n} f - \mathbb{E}_{q^\star} f\right| \;\leq\; \left(2\|\nu_n - q^\star\|_{\text{TV}}\right)^{\delta/(1+\delta)} \left(\mathbb{E}_{\nu_n} |f|^{1+\delta} + \mathbb{E}_{q^\star} |f|^{1+\delta}\right)^{1/(1+\delta)},$$

so the theorem's rate carries over raised to $\delta/(1+\delta)$. We record the condition rather than assume it.

B.7 Closing the Neal–Hinton bound, and the step that remains open

The E-step iterates $\tilde{P}$: at step k draw $z_s^{(k)} \sim p_{s|0}(\,\cdot \mid z_0^{(k-1)})$, then $z_0^{(k)} \sim q^{\text{work}}(\,\cdot \mid z_s^{(k)})$. The law of the imputed value is the chain marginal ν_k of Theorem 2, an average over the trajectory and not q^{work} at a fixed z_s, so the quantity dropped from equation 6 is $\mathrm{KL}(\nu_k \| q^\star)$; the per-step slack of Theorem 1 enters only through $\bar{\varepsilon}_{\text{TV}}$ in part (i). Theorem 2(iii) supplies $\|\nu_k - q^\star\|_{\text{TV}} \to 0$, and one step remains:

$$\|\nu_k - q^\star\|_{\text{TV}} \to 0 \quad \overset{?}{\Longrightarrow} \quad \mathrm{KL}(\nu_k \| q^\star) \to 0,$$

which total variation does not give. The usual route, $\mathrm{KL}(\nu_k \| q^\star) \leq W_2(\nu_k, q^\star) \sqrt{\mathcal{I}(\nu_k \| q^\star)}$ by HWI, asks for three things we do not have: a W_2 statement, for which Theorem 1 would have to be uniform in z_s and is not (Remark 1); a log-Sobolev inequality for $q^\star$, the hypothesis Theorem 2 refuses when it uses Pinsker in place of Talagrand; and $\mathcal{I}(\nu_k \| q^\star)$ bounded along the chain.

One half closes in relative entropy. Let $\tilde{\Pi}, \Pi$ be the path laws of $(z_0^{(0)}, \ldots, z_0^{(n)})$ under $\tilde{P}$ and P from a common ν_0. The chain rule on path space, data processing under projection onto the last coordinate, and joint convexity of KL against the shared mixing weight $p_{s|0}(\cdot \mid z_0)$ give

$$\begin{aligned} \mathrm{KL}\big(\nu_0 \tilde{P}^{\,n} \,\big\|\, \nu_0 P^{\,n}\big) \;\leq\; \mathrm{KL}(\tilde{\Pi} \| \Pi) \;&=\; \sum_{k=1}^{n} \mathbb{E}_{\tilde{\Pi}} \,\mathrm{KL}\big(\tilde{P}(z_0^{(k-1)}, \cdot) \,\big\|\, P(z_0^{(k-1)}, \cdot)\big) \\ &\leq\; n\, \bar{\varepsilon}_{\text{KL}}, \qquad \bar{\varepsilon}_{\text{KL}} := \sup_{z_0} \mathbb{E}_{z_s | z_0} \mathrm{KL}\big(q^{\text{work}} \,\big\|\, q_s\big). \end{aligned} \tag{18}$$

Theorem 1 bounds $\bar{\varepsilon}_{\mathrm{KL}}$ directly: no Pinsker, no compact, no (A7), and it is the same supremum of an average that part (i) uses. Two things it does not settle. It holds on the interior-mode event only, where the (A5) exception costs α in total variation and has no such ceiling in relative entropy; and it compares ν_n with $\nu_0 P^n$, not with $q^\star$, the remaining step being part (ii), which Harris recurrence delivers in total variation alone. The obstruction therefore sits entirely in the ergodic half.

A second route, at the price of one hypothesis and a weaker conclusion. Fix $b < \infty$ and write μ for $q^\star$ and $\tilde{\nu}_k$ for ν_k, each restricted to $[a, b]$ and renormalised. With $p_k := \nu_k([a, b])$ and $q := q^\star([a, b]) > 0$, restriction costs only a constant, $\|\tilde{\nu}_k - \mu\|_{\mathrm{TV}} \le 2p_k^{-1}\|\nu_k - q^\star\|_{\mathrm{TV}}$ and $p_k \ge q - \|\nu_k - q^\star\|_{\mathrm{TV}}$. Suppose

(A7) *(bounded ratio on a compact.)* There is $M < \infty$ with $\left|\mathrm{d}\tilde{\nu}_k/\mathrm{d}\mu - 1\right| \le M$ on $[a, b]$, uniformly in k.

Then $(\mathrm{d}\tilde{\nu}_k/\mathrm{d}\mu - 1)^2 \le M\left|\mathrm{d}\tilde{\nu}_k/\mathrm{d}\mu - 1\right|$ pointwise, and $\mathrm{KL} \le \chi^2$ by $\log x \le x - 1$, so

$$\begin{aligned} \mathrm{KL}(\tilde{\nu}_k\|\mu) \;\le\; \chi^2(\tilde{\nu}_k\|\mu) \;&=\; \int\Big(\tfrac{\mathrm{d}\tilde{\nu}_k}{\mathrm{d}\mu} - 1\Big)^2 \mathrm{d}\mu \;\le\; M\int\Big|\tfrac{\mathrm{d}\tilde{\nu}_k}{\mathrm{d}\mu} - 1\Big|\,\mathrm{d}\mu \\ &=\; 2M\,\|\tilde{\nu}_k - \mu\|_{\mathrm{TV}} \;\le\; \frac{4M}{p_k}\,\|\nu_k - q^\star\|_{\mathrm{TV}}. \end{aligned} \tag{19}$$

Every step of equation 19 is elementary. Three things it does not give.

It does not convert a floor into a zero. The right-hand side is bounded by Theorem 2 and not sent to zero by it, since part (iii) drives the proposal's shape error down and leaves the score error of Remark 4 and the α of a boundary mode:

$$\mathrm{KL}(\tilde{\nu}_k\|\mu) \;\lesssim\; \frac{4M}{q}\Big[n_{\mathrm{gibbs}}\big(\sqrt{\bar{\mathcal{K}}/2} + \alpha\big) + \big\|\nu_0 P^{\,n_{\mathrm{gibbs}}} - q^\star\big\|_{\mathrm{TV}}\Big].$$

This vanishes only where $\mathcal{K}$ does, which on the fitted models it does not (Remark 2), and under $n_{\mathrm{gibbs}}\alpha \to 0$, which the proof of part (iii) leaves open. The route removes the W_2, log-Sobolev and Fisher-information obstacles and inherits the ones Theorem 2 already had.

M moves with the schedule. ν_k is a mixture of truncated normals of width at most σ_{post}, so $M \lesssim C/(\sigma_{\mathrm{post}} \inf_{[a,b]} q^\star)$ at fixed s_{imp}: uniform in k, and growing like $\sigma_{\mathrm{post}}^{-1}$ along the schedule. The route therefore asks for $n_{\mathrm{gibbs}}\sigma_{\mathrm{post}} \to 0$ where part (iii) supplies only $n_{\mathrm{gibbs}}\sigma_{\mathrm{post}}^2 \to 0$, and $n_{\mathrm{gibbs}} = \sigma_{\mathrm{post}}^{-3/2}$ satisfies the second and fails the first.

The conclusion is about the restricted law. Convergence is to $q^\star$ restricted to $[a, b]$ and not to $q^\star$, and the latent is unbounded, so that is a weakening rather than a formality.

We state the route and do not take it. Nothing in the experiments rests on it: the training objective is the Monte-Carlo EM gradient of equation 7, a valid upper bound for any proposal supported on the constraint set, and the argument here concerns only how tight that bound becomes as n_{gibbs} grows.

APPENDIX C TRAINING PROCEDURE, ALGORITHM, AND BASELINES

C.1 TRAINING PROCEDURE AND SAMPLING

We now turn from the theoretical analysis to the practical training loop, in the same ε-parameterisation used throughout. The total training loss combines the strict DSM upper bound on the exact channel (equation 4) with the variational EM bound on the censored channel (Theorem 1 or 3):

$$\mathcal{L}_{\mathrm{EM}}(\theta) \;=\; \mathbb{E}_{z_0 \sim q^{(\cdot)}}\big[\mathcal{L}_{\mathrm{DSM}}^{\mathrm{exact}}(\theta; z_0, x)\big], \tag{20}$$

where the expectation is taken under either q^{work} or $q^{\mathrm{work},\Sigma}$. The full variational bound (6)–(7) is $L \cdot \mathcal{L}_{\mathrm{EM}}(\theta) - H(q^{(\cdot)}) + C$. With the proposal fixed at the EMA parameters in the E-step, L, $H(q^{(\cdot)})$, and C are θ-independent, so $\nabla_\theta \mathcal{L}_{\mathrm{EM}}$ is the gradient of the full per-step bound (standard MCEM). Since $q^{(\cdot)} \neq q^\star$, the fixed point is displaced from the MLE by the residual KL.

E-step. For each sample, K independent chains run n_{gibbs} Gibbs sweeps: noise z_0 to z_s, apply the frozen EMA network to obtain $\hat{z}_0$, then resample the censored channel from $q^{\mathrm{work}} =$

Algorithm 1 One training step of the proposed model.

Notation: $\sigma_{\text{post}}(s) := \sigma\sqrt{s}$, $b(s) := (\mu - \frac{1}{2}\sigma^2)\, s$; latent channel $c_i := E_i$ ($E_i{=}0 \Rightarrow T^\star$, $E_i{=}1 \Rightarrow C$); observed boundary $\ell_i := \log T_i^{\text{obs}}$, repeated across the K chains so that it is indexed by the BK chain index below.

E-step
$\mathcal{D} \leftarrow \varnothing$ {set of imputed clean pairs}
for $i = 1, \dots, B$ **and** $k = 1, \dots, K$ **do**
  $z_0 \leftarrow (\ell_i, \ell_i)$; $z_{0,c_i} \leftarrow \max\big(\ell_{\pi(i,k)},\, \ell_i\big)$ {latent channel reseeded from permuted anchor; $\pi \in \mathrm{Sym}(BK)$}
  for $g = 1, \dots, n_{\text{gibbs}}$ **do**
    $z_s \leftarrow z_0 + b(s_{\text{imp}}) + \sigma_{\text{post}}(s_{\text{imp}})\, \varepsilon$; $\varepsilon \sim \mathcal{N}(0, I_2)$
    $\hat{z}_0 \leftarrow z_s - b(s_{\text{imp}}) - \sigma_{\text{post}}(s_{\text{imp}})\, \bar{\varepsilon}_\theta(z_s,\, s_{\text{imp}},\, x_i)$
    $\sigma_{\text{prop}} \leftarrow \begin{cases} \sigma_{\text{post}} & \text{OFF (FATUM)} \\ \sigma_{\text{post}} \big/ \sqrt{1 + \sigma_{\text{post}}^2 \widehat{V}(z_s, x_i)} & \text{GRAD (FATUM}_\Sigma\text{)} \end{cases}$ {$\widehat{V}$ from the denoiser Jacobian, equation 21}
    $z_{0,c_i} \sim \mathcal{N}\big(\hat{z}_{0,c_i},\, \sigma_{\text{prop}}^2\big)\big|_{\geq \ell_i}$ {survival-side inverse CDF; observed channel stays ℓ_i}
  end for
  $\mathcal{D} \leftarrow \mathcal{D} \cup \{(z_0,\, x_i)\}$
end for
M-step:
$\mathcal{L} \leftarrow 0$
for each $(z_0,\, x) \in \mathcal{D}$ **do**
  $s \sim \mathrm{LogUniform}[s_{\min}, s_{\max}]$
  $z_s \leftarrow z_0 + b(s) + \sigma_{\text{post}}(s)\, \varepsilon$; $\varepsilon \sim \mathcal{N}(0, I_2)$
  $\mathcal{L} \leftarrow \mathcal{L} + \frac{1}{2} \big\|\varepsilon_\theta(z_s,\, s,\, x) - \varepsilon\big\|_2^2$
end for
$\theta \leftarrow \theta - \eta\, \nabla_\theta\big(\mathcal{L}/|\mathcal{D}|\big)$
update EMA $\bar{\varepsilon}_\theta$.

$\mathcal{N}(\hat{z}_0,\, \sigma_{\text{prop}}^2)\big|_{z_0 \geq \log T^{\text{obs}}}$ via inverse CDF (observed channel stays at $\log T^{\text{obs}}$). Scale: $\sigma_{\text{prop}} = \sigma_{\text{post}}(s_{\text{imp}})$ for q^{work}, $\sqrt{\Sigma}$ for $q^{\text{work},\Sigma}$. Each chain's final state contributes independently to the M-step.

M-step. Standard DSM on the imputed pairs with log-uniform s; loss $\frac{1}{2}\|\varepsilon_\theta(z_s, s, x) - \varepsilon\|^2$ averaged over both channels, all imputations, and the batch, which is the unweighted $\mathcal{L}_{\text{DSM}}^{\text{exact}}$ of Lemma 1.

Sampling. At inference we draw conditional samples from $p(T^\star, C \mid X)$ via predictor–corrector SDE and estimate $\hat{S}(t \mid x) = N_{\text{mc}}^{-1} \sum_{k=1}^{N_{\text{mc}}} \mathbf{1}\{T_k^\star > t\}$ over N_{mc} Monte-Carlo trajectories per subject.

C.2 Baselines

Discriminative Nine baselines: Cox proportional hazards, Weibull AFT and log-normal AFT (`lifelines`); random survival forests (RSF) and gradient-boosted survival analysis (GBSA, `scikit-survival`); DeepSurv, DeepHit, CoxTime and CoxCC (`pycox`). Each is tuned individually by Optuna TPE (Akiba et al., 2019) over 100 trials against dev-val IBS (Graf et al., 1999), and all are scored through one `EvalSurv` pipeline (Kvamme et al., 2019) for IBS and tdAUC. Table 1 reports all nine, and the macro means average cohorts rather than cells.

Generative Our proposed pipeline (TabDiff+FATUM) uses the decomposed factorisation $p(X)\, p(T^\star, C \mid X)$: TabDiff generates covariates and FATUM generates $(T^\star, C)$ conditioned on the synthetic X. End-to-end generative baselines are TabDiff, CTGAN, TVAE, and SurvivalGAN (the latter three via `synthcity`). For each split we generate $n_{\text{train}}^{\text{syn}} = n_{\text{train}}$ synthetic rows for

TSTR and $n^{\text{syn}}_{\text{holdout}} = n_{\text{holdout}}$ rows for marginal-fidelity evaluation. Generative baselines use their published best hyperparameters without dataset-specific tuning. FATUM tunes only the conditional $p(T^\star, C \mid X)$ component via Optuna; the TabDiff covariate model and all end-to-end baselines retain their published configurations, consistent with standard practice where end-to-end tuning against held-out real data would conflate generation quality with overfitting. Under TSTR, CoxPH and RSF audit models are trained on synthetic data and evaluated on the real holdout with the discriminative metric battery. The IPCW weights, the integration window and the admissible holdout all come from the real train fold, so they are identical for every generator and the IBS stays a proper scoring rule on the real holdout. A holdout subject whose time exceeds the longest real training time carries no weight and leaves the audit set; that subset is fixed per cohort and drops 0.00% to 0.72% of the holdout, 0.20% on average. A generator whose synthetic times stop short of the evaluation grid pays for it through flat extrapolation of its survival curve; the audit set does not shrink.

C.3 Architecture, search space, optimiser and sampler

The curvature estimate $\widehat{V}$ FATUM$_\Sigma$ needs the local prior curvature, which has no closed form for a learned score, so it is read off the denoiser Jacobian. Stein's identity for the truncated posterior, $\partial_{z_s}\widehat{z}_\theta = \mathrm{Var}_{q_s}[z_0]/\sigma^2_{\text{post}}$, together with the Laplace variance $\mathrm{Var}_{q_s}[z_0] \approx \Sigma = \sigma^2_{\text{post}}/(1+\sigma^2_{\text{post}}V_{\text{prior}})$, inverts to

$$\widehat{V}(z_s, x) \;=\; \frac{1 - J(z_s, x)}{\sigma^2_{\text{post}}(s_{\text{imp}})\, J(z_s, x)}, \qquad J(z_s, x) := \mathrm{clip}\big(\partial_{z_s}\widehat{z}_\theta(z_s, x),\, 10^{-4},\, 1\big), \tag{21}$$

so that $\widehat{\Sigma}(z_s, x) = \sigma^2_{\text{post}} J$ of equation 9, the E-step proposal scale is $\sigma_{\text{post}}\sqrt{J}$, and the square root is always real. $\partial_{z_s}\widehat{z}_\theta$ is the exact per-channel diagonal of the denoiser Jacobian by reverse-mode autodiff (two backward passes per sweep), not a finite difference. Only the upper clip binds in practice; $J \le 1$ forces $\widehat{V} \ge 0$, and where it binds it sets $\widehat{V} = 0$ and $\widehat{\Sigma} = \sigma^2_{\text{post}}$, so FATUM$_\Sigma$ proposes there exactly what FATUM proposes, where the unclipped formula would *widen* the proposal. The Jacobian is recorded before clipping, so the clip cannot hide how often it fires: 0.3 to 8.9% of the forward-noised states (Appendix F) and 3.2 to 66.4% of the states the chain visits (Appendix D.2). Neither figure measures V_{prior}; both measure $\partial_{z_s}\widehat{z}_\theta$, and the step between them is Stein's identity, which holds at the exact posterior mean and not at the network's (Remark 4). Appendix D.3 closes that step by reading the curvature off the score directly.

Preprocessing Continuous covariates are z-scored, categoricals one-hot encoded, and both time channels log-standardised so the latent z_0 has approximately unit scale.

Architecture The noise-prediction network $\varepsilon_\theta(z_s, s, x)$ is a residual MLP with sinusoidal embeddings for s, a linear projection for x, and SiLU residual blocks. An EMA copy of the weights (decay 0.999) (Song & Ermon, 2019) drives both E-step imputation and evaluation. The forward process is fixed at $\mu = 0$, $\sigma = 1$, $s_{\min} = 10^{-2}$, $s_{\max} = 5$. Depending on the dataset, the selected architecture has between 15k and 120k parameters.

Hyperparameter search Per-dataset hyperparameters are tuned by Optuna TPE on an $85/15$ train/val split of the dev fold for 100 trials, each scored by the mean over seeds of dev-val IBS plus 0.5 KM-MSE; non-finite trials are pruned, not recorded. Three of the searched parameters are the ones the theory names. The imputation noise level s_{imp} is log-uniform on $[0.01, 1]$ and sets the proposal spread (§3.3). The number of Monte-Carlo imputations $K_{\text{imp}} \in \{1, 2, 4, 8, 10\}$ per censored sample cuts the M-step gradient variance by roughly $1/K$ at K times its cost. The number of Gibbs sweeps $n_{\text{gibbs}} \in \{2, 4, 8, 10, 20\}$ controls how far the imputation chain advances towards $\tilde{q}$ before each M-step gradient (Theorem 2); $n_{\text{gibbs}} = 1$ is excluded as the biased step the theorem does not cover.

The rest the theory says nothing about: learning rate, log-uniform on $[10^{-6}, 10^{-3}]$; batch size in $\{64, 128, 256, 512\}$; hidden widths, three residual stacks from $[96, 64, 32]$ to $[128, 96, 96, 64]$; the epoch budget and early-stopping patience; and three structural switches, whether the trunk is channel-factorised, whether the two channels' tails are decoupled, and whether batches are balanced by event status. Chain initialisation is anchored off the censoring boundary, which moves the E-step's starting point and not its stationary law.

The selected operating points are $(s_{\text{imp}}, n_{\text{gibbs}}) = (0.771, 20)$ for AIDS, $(0.256, 2)$ for FLCHAIN, $(0.156, 10)$ for SUPPORT, $(0.058, 20)$ for Rotterdam, $(0.036, 4)$ for METABRIC and $(0.033, 20)$ for GBSG2. Three sit above the crossing at $s_{\text{imp}} = 0.05$ to 0.10 where the Σ correction starts to pay, two below it and Rotterdam inside it, which is the spread §5.3 reads.

Two arms are reported. **FATUM** sets `v_prior=off` and uses the uncorrected proposal q^{work}; this is the arm every headline number refers to. **FATUM$_\Sigma$** sets `v_prior=grad` and uses $q^{\text{work},\Sigma}$ with $\widehat{V}$ read from the denoiser Jacobian by automatic differentiation equation 21, so it introduces no new tunable parameter and runs as a single-knob override on the first arm's configuration.

Compute Every run, ours and every baseline, trains and evaluates on a single NVIDIA A100 80GB.

Optimizer and schedule AdamW with weight decay 10^{-4} and cosine-annealing learning-rate schedule over up to 1000 epochs with early stopping on dev-val IBS.

Sampling Survival curves are estimated from $N_{\text{mc}} = 500$ Monte Carlo reverse-SDE trajectories per subject, using a predictor–corrector sampler with $N_{\text{steps}} = 200$ discretization steps.

Seed sensitivity We repeat the holdout evaluation under ten random seeds[2] and report means with 95% Student-t confidence intervals.

Statistical testing The unit of analysis is the cohort ($m = 6$), for the reason given in §4.1: the ten seeds inside a cohort share one split, so they resample the training procedure rather than the data. Each cohort contributes its mean paired difference; we report the paired Student-t interval and p-value over those m means, the number of cohorts favouring our model, and the exact two-sided sign-flip p, whose floor at $m = 6$ is $2^{1-m} = 0.031$. A two-stage cluster bootstrap (resampling cohorts, then seeds within a cohort) is computed alongside and is narrower than the t interval on every one of the forty discriminative comparisons, so we quote the t interval as the conservative one and say where the two disagree.

Because the sign-flip floor is 0.031, a Benjamini–Hochberg correction over a family of ten cannot star an isolated comparison at $q < 0.05$ (§4.1), so we do not report BH-adjusted stars at cohort level. The pooled cell-level Wilcoxon signed-rank test over all sixty (cohort, seed) cells, with BH adjustment across the ten comparisons within a metric (Benjamini & Hochberg, 1995), is reported in Table 6 for comparison only, beside the intra-cohort correlation and the effective sample size that explain the difference between the two columns. The same rule is applied to comparisons that go against us.

C.4 Benchmark cohorts

Table 3 lists the six cohorts with their sizes, covariate dimensions, censoring rates and follow-up ranges. Censoring spans 32% to 92% and n spans 686 to $8{,}873$, which is the spread every per-cohort number in this paper is read against.

Appendix D Routes to Theorem 2 that measurement closes

D.1 What Theorem 2 assumes, and what it does not

Part (i) uses no contraction, no log-concavity and no property of $\hat{z}_\theta$ beyond measurability; the prefactor n is the trivial Dobrushin bound $\delta(\tilde{P}) \leq 1$, and the statement needs no invariant law to exist. Two standard perturbation arguments would replace that n by a constant, and one property of the network denies both. The network reads z only through a sinusoidal embedding (Appendix C.3), so at fixed (s, x) the map $z \mapsto \varepsilon_\theta$ is bounded and $\hat{z}_\theta(z_s) = z_s + b(z_s)$ with $\sup|b| \leq B < \infty$, whatever the trained weights are.

[2]Seeds drawn once and fixed for all experiments: 2960, 2764, 3648, 4848, 4897, 6098, 6930, 7105, 8644, 9062.

Table 3: Benchmark cohort characteristics. n and p are total patients and covariate dimension after preprocessing (one-hot encoding where applicable); $[\min T, \max T]$ is the range of observed times in each cohort's own units, and the last column the share of records with $E = 0$.

Cohort	n	p	$[\min T,\ \max T]$	Censored (%)
AIDS	1151	19	[1, 364]	91.7
METABRIC	1904	9	[0.01, 355]	42.1
GBSG2	686	9	[8, 2659]	56.4
SUPPORT	8873	14	[3, 2029]	32.0
FLCHAIN	7874	23	[0.01, 5215]	72.5
Rotterdam	2232	7	[0.26, 87.4]	43.2

First, $\hat{z}_\theta$ is then no contraction on $[a, \infty)$. If $|\partial_{z_s} \hat{z}_\theta| \le \rho < 1$ everywhere, then for $z > y$, $\rho(z-y) \ge \hat{z}_\theta(z) - \hat{z}_\theta(y) \ge (z-y) - 2B$, so $(1-\rho)(z-y) \le 2B$, which fails once $z - y > 2B/(1-\rho)$; the secants tend to 1, so $\sup_{z_s} |\partial_{z_s} \hat{z}_\theta| \ge 1$. A Wasserstein perturbation bound (Rudolf & Schweizer, 2018) runs on that contraction, which a lower curvature bound $V_{\text{prior}} \ge \underline{V} > 0$ would otherwise supply through the Stein identity and Brascamp–Lieb (Brascamp & Lieb, 1976); its variant that asks for a Lyapunov condition in place of a global contraction returns to the infimum of the pull below. Either form would also need the slack in W rather than in total variation, where Remark 1 has only the latter. Appendix D.3 reads the curvature directly, to say which of the two objects the failure belongs to.

Second, the identity term leaves M unbounded, which is $\delta(\tilde{P}) = 1$ and rules out the Doeblin minorisation a uniform perturbation bound needs (Mitrophanov, 2005) (Remark 3). What stays open is the infimum of the pull along the ray, which decides whether $\tilde{P}$ has an invariant law at all and which Appendix D.4 probes without settling. Theorem 2 is stated so as to need none of this.

D.2 The chain's realised behaviour

Theorem 2 bounds the E-step's error by n_{gibbs} times a per-sweep slack, using nothing about the chain beyond non-expansion of a Markov kernel in total variation. That is deliberately weak, and the coupling argument behind it predicts something sharper, which we check directly: run the proof's two chains forward with common noise and a common uniform, and compare the realised one-sweep ratio against the per-step derivative of the truncated one-sweep mean map, the map M of Appendix D.4 composed with the truncation at a,

$$M'_a \;=\; \left|\partial_{z_s} \hat{z}_\theta\right| \left(1 - h'(\alpha_q)\right),$$

for h the inverse Mills ratio of Appendix B.2 and α_q its standardised truncation point, which is not the exception probability α of (A5). Above the data the factor is one and $M_a = M$.

Across six cohorts, four values of s_{imp} and three seeds, at the smallest offset of the d_0 sweep (0.03 in standardised units, against 0.1, 0.3 and 1), so that the measurement estimates a derivative rather than a secant, the realised ratio falls below M'_a in all 72 cells, at 0.862 ± 0.049 of it with correlation 0.986 (Fig. 4). The analysis therefore bounds the realised behaviour closely, and it also shows how much the theorem gives away: the realised factor at the operating points is 0.44–0.70, so error accumulates geometrically rather than to the $n_{\text{gibbs}} \in \{2, \dots, 20\}$ the bound charges, and is conservative by a measured factor of 1.1 to 7.3.

Two caveats come from the same run. The ratio is formed from the mean coupling gap across subjects, so it speaks to average behaviour. And at the *supremum* the one-step contraction hypothesis is outright false: the unclamped $\sup_{z_s} |\partial_{z_s} \hat{z}_\theta|$, differentiated exactly at the states the chain visits, is 1.02–1.41 per cell and 1.03–1.38 seed-averaged, its product with the truncation factor 1.02–1.38 (1.02–1.31), and the violating set is macroscopic rather than marginal, 3.2% to 66.4% of those states per cell (3.8%–58.1%), so neither restricting the domain nor reweighting the metric repairs it. The Jacobian is not the curvature: $\partial_{z_s} \hat{z}_\theta \ge 1$ is equivalent to $V_{\text{prior}} \le 0$ only through the Stein identity, which holds at the exact posterior mean and not at $\hat{z}_\theta$. Appendix D.3 reads the curvature directly and settles which of the two fails. These are the two routes one would try first, and closing them is why Theorem 2 runs through non-expansion, which needs no contraction constant at all.

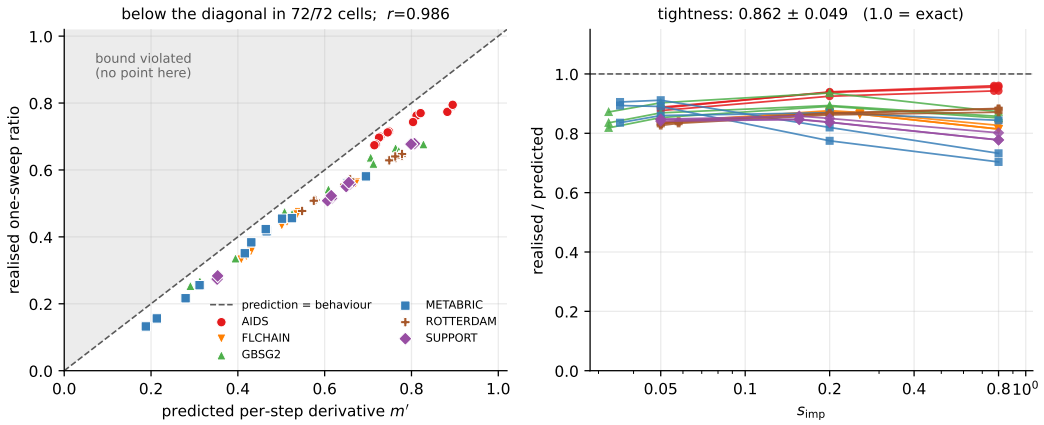


Figure 4: **The per-step derivative the proof computes bounds the contraction the chain realises, and tightly.** Left: realised one-sweep ratio against the predicted M'_a, one point per cohort, s_{imp} and seed, all 72; the shaded region, where the bound would fail, contains no point. Right: their ratio against s_{imp}, one line per cohort and seed. The ratio is read at the smallest offset of the d_0 sweep, so it estimates a derivative and not a secant.

The same runs settle the double limit in Theorem 2(iii), where part (i) caps the sweeps at $o(1/s_{\mathrm{imp}})$ while the chain mixes more slowly as s_{imp} falls. Within every cohort the realised one-sweep factor does rise along the grid, from 0.21 to 0.46 on METABRIC and from 0.69 to 0.79 on AIDS between $s_{\mathrm{imp}} = 0.8$ and $s_{\mathrm{imp}} = 0.05$, but it rises slowly: regressing the implied mixing time $1/(1-\rho)$ on s_{imp} gives s_{imp}^{-b} with $b = 0.12$ to 0.29 over the six cohorts, against the $b < 1$ that part (i) leaves room for (`mixing_exponent.py`). Schedules meeting both conditions therefore exist on these models, which is a measurement and not the general statement part (iii) would need. We tune s_{imp} on dev-val IBS as a bias–variance tradeoff. The product $n_{\mathrm{gibbs}}(\sigma^2_{\mathrm{post}}\bar{V} + \alpha)$ that part (iii) asks to vanish needs a supremum of $|\partial^2_{z_0} \log p_\theta|$ over every state and an (A5) exception probability, and the run identifies neither, so the per-sweep factor above stands in for it. The range over which the bound says anything is set by $\bar{\mathcal{K}}$, which does not shrink with s_{imp} in the regime Remark 2 puts the fitted models in: at the $u^\star$ reported there, $n\bar{\varepsilon}_{\mathrm{TV}}$ reaches the trivial bound of one at an n of order ten, and the burn-in below puts the chain at stationarity by 1 to 5 sweeps.

Finally, the bound is most sensitive to n_{gibbs}, and the same runs give it: started as the algorithm starts it, the chain becomes indistinguishable from stationarity in 1 to 5 sweeps against configured n_{gibbs} of 2 to 20, so every configuration is adequate and four are generous.

D.3 The prior curvature, measured directly

Appendix D.2 records $\partial_{z_s} \hat{z}_\theta \geq 1$ and turns it into $V_{\mathrm{prior}} \leq 0$ through the Stein identity, which holds at the exact posterior mean and not at the network's. The curvature is available without that step. $V_{\mathrm{prior}} = -\partial^2_{z_0} \log p_\theta$ is $-\partial_{z_0} s_\theta(z_0, s)$ at small s, one backward pass, read at the same states the chain visits (`prior_curvature.py`).

Read there over three seeds, the two shares agree in five cohorts.

	AIDS	FLCHAIN	GBSG2	METABRIC	ROTTERDAM	SUPPORT
$V_{\mathrm{prior}} \leq 0$	10.6	13.6	5.1	5.2	15.8	6.5
$\partial_{z_s} \hat{z}_\theta \geq 1$	1.6	11.3	5.5	5.5	17.0	7.0

Table 4: Share of visited states (%, seed means over three seeds) on which the directly measured curvature is non-positive, against the share on which the denoiser Jacobian expands.

Table 5: **Is the working kernel pulled back above the data?** Each cell is $m(z) - z$ for the worst subject in that cohort, at z standard deviations above the cohort's 99th percentile, from 512 draws per point and one seed per cohort. Negative means every subject is pulled back. Gray marks the points where the worst subject is not, by more than twice its own Monte-Carlo error. κ is the slope of m at the far end.

Cohort	2	4	8	16	32	64	κ	pulled
AIDS	$+0.059$	-0.011	-0.097	-0.256	-0.321	-0.236	1.006	100%
METABRIC	-0.155	-0.133	-0.173	-0.121	-0.014	-0.167	0.994	100%
GBSG2	-0.046	-0.052	-0.041	-0.023	-0.056	-0.072	1.000	100%
SUPPORT	-0.171	-0.131	-0.173	-0.266	-0.059	-0.074	1.002	100%
FLCHAIN	-0.212	-0.189	-0.097	-0.073	-0.023	$+0.119$	1.006	94%
Rotterdam	-0.053	$+0.008$	$+0.112$	$+0.160$	$+0.115$	$+0.127$	1.000	4%

On AIDS the direct reading finds the event more often than the Jacobian does, the opposite of a Jacobian that overstates expansion, and its seed spread is the widest of the six, 4.9% to 16.6%. The event where the Jacobian expands and the direct curvature says it should not covers 0.3% to 3.2% of states as a seed mean, and no cell exceeds 3.5%. The learned prior is therefore log-convex on those states as a matter of fact, and what the measurement of Remark 4 refutes is a property of the prior rather than of the score.

The probe smooths what it reads, since p_s is p_θ convolved with a Gaussian of width $\sigma\sqrt{s}$ and convolution moves a density toward log concavity. That cuts one way: $V_{\text{prior}} \le 0$ at the probe implies the same unsmoothed, while $V_{\text{prior}} > 0$ does not. The shares above are read at $s = 0.005$ and hold to $s = 0.1$, so they are lower bounds on how often the event occurs.

(A2) asks for less, $1 + \sigma_{\text{post}}^2 V_{\text{prior}} > 0$, and its margin is $\sigma_{\text{post}}^{-2}$, which shrinks as s_{imp} grows. Four cohorts meet it at every visited state on every seed. FLCHAIN misses on 0.1% to 0.2% at the finest probe and on none at any coarser one. AIDS misses on 1.1% to 4.0%, on all three seeds, and it is the cohort tuned to $s_{\text{imp}} = 0.77$ against 0.033 to 0.26 elsewhere, so its floor sits at -1.3 where the others sit between -3.9 and -30.7. Those shares are the α_2 of (A2). The assumption is thus tied to a tuned quantity rather than to the architecture, and the cohort that buys the widest proposal is the one with the least room.

D.4 A DRIFT CONDITION FOR THE WORKING KERNEL, MEASURED

A Foster drift would give what the contraction route cannot, an invariant law for $\tilde{P}$ with no contraction hypothesis. Take the affine $W(z) = 1 + (z - a)$, whose sublevel sets are compact in $[a, \infty)$; Foster's criterion then asks for a negative pull outside a compact, $M(z) - z \le -\epsilon$, where $M(z) := \mathbb{E}_\xi[\hat{z}_\theta(z + \mu_{\text{drift}} s_{\text{imp}} + \sigma_{\text{post}}\xi)]$ is the one-sweep mean map, written M to keep it clear of the Mills ratio. Above the data the truncation at a is inactive, so $\mathbb{E}[z' \mid z] = M(z)$ and the condition is a property of the denoiser alone.

The quantity to measure is the pull, not the drift ratio. Comparing $\mathbb{E}[V(z')]/V(z)$ with one as a supremum over subjects would need the supremum of several thousand Monte-Carlo estimates, biased upward by roughly $\sqrt{2 \log B}$ standard errors, which at these sample sizes exceeds the margin $|\cdot - 1| \approx 0.01$–0.06 that separates a drift from its failure. The pull is per subject, so it needs no supremum over noisy estimates, only enough draws to resolve it. We measure it with 512 draws per probe point, at 2 to 64 standard deviations above each cohort's 99th percentile, on one trained model per cohort, the arm the tables are built from (`tail_slope.py`). One model per cohort is enough to separate the cohorts that pass from the ones that fail, and not enough to put an interval on the crossing point.

Table 5 gives the worst subject at each probe point. Four cohorts pull every subject back at every point probed and two do not. Foster's criterion asks for more than that, and the grid is where the gap sits. The network reads z only through a sinusoidal embedding (Appendix C.3), so ε_θ is bounded in z and $M(z) - z = -\sigma_{\text{post}}\mathbb{E}[\varepsilon_\theta]$ is bounded and quasi-periodic rather than monotone, its shortest period 2π in standardised units against gaps of 16 and 32 between the last three probe points. Six points bound the pull where they sit and not its supremum along the ray, so a grid that settled the

criterion would have to be finer than the period rather than wider in range. We read the table as a statement about the probed points and claim no invariant law from it. The half-line would cost nothing, since the criterion asks the sublevel set $\{W \leq d\}$ to be petite and not the state space to be compact. That set is the body of the distribution with a margin, of diameter 40 to 50 standard deviations at the selected s_{imp}, so a one-step minorisation across it costs $\exp(-O(1/s_{\mathrm{imp}}))$: petite, with a constant too small to carry a rate. FLCHAIN crosses only at 64 standard deviations ($+0.119 \pm 0.020$, 6% of subjects), where no training data exists; Rotterdam crosses from 8 on ($+0.112 \pm 0.011$) with only 4% of subjects pulled back, which is not a probing artefact. The statement is therefore per cohort, and a drift that moves with regularisation or s_{imp} would make the theorem a property of the checkpoint. Theorem 2 claims no invariant law for that reason.

No rate follows under this W. The additive form $M(z) = z - c(z)$ is structural, since a bounded ε_θ leaves the identity term of $\hat{z}_\theta$ alone, and the measured $\kappa = 0.99$ to 1.01 records that rather than an asymptotic slope. An affine W turns an additive drift into Foster's criterion and not into $PW \leq \lambda W + b$. An exponential W would give a geometric rate governed by c/σ_{post}. The forward drift cancels in the one-sweep mean map, leaving $M(z) - z = -\sigma_{\mathrm{post}}\mathbb{E}[\varepsilon_\theta] = \sigma^2_{\mathrm{post}}\,\mathbb{E}[s_\theta]$ by the ε-parameterisation, so c/σ_{post} falls like $\sqrt{s_{\mathrm{imp}}}$ and the route is weakest in the limit Theorem 2(iii) certifies.

A denser follow-up, thirteen offsets on three seeds, puts the exponential-W constant at $\lambda^\star = 0.929$ to 0.998 where the drift holds, an implied $1/(1-\lambda^\star)$ of 14 to 514 sweeps against the 1.8 to 3.3 the chain realises, and the tail Jacobian at 1.036 to 1.592 against the 1.02–1.41 of the body (`run_tail_dense.sh`). $\lambda^\star$ is a plug-in that ignores the truncation at a and reads a measured supremum as a global Lipschitz constant, so it places an order of magnitude and certifies no rate. Two cohorts change verdict with the seed.

Both constants are uniform over the starting point and the algorithm is not. $\lambda^\star$ and any Doeblin constant certify mixing for a chain started anywhere on $[a, \infty)$, while the E-step starts at the prior anchor, from which the burn-in of Appendix D.2 reaches the Monte-Carlo noise floor in 1 to 5 sweeps. The factor of 20 to 100 between the two measures that uniformity; it is not slack in either bound.

APPENDIX E BOTH UNITS OF ANALYSIS, SIDE BY SIDE

§4.1 fixes the cohort as the unit of analysis and states that the pooled cell-level numbers appear here. Table 6 gives both, with the intra-cohort correlation and effective sample size that account for the difference, and Figure 5 plots the two intervals against each other. Nine of ten IBS comparisons clear zero when all sixty cells are treated as replications; two do at the cohort level. We draw no conclusion from the pooled column and include it so the design effect is visible rather than asserted. At $m = 6$ the effect size detectable at 80% power is 1.42 between-cohort standard deviations, against observed effect sizes of 0.24 to 1.61, so the design resolves only the largest differences in the panel. Table 7 is the raw material for both columns: holdout IBS, one column per cohort, for every method compared.

Effect sizes instead of raw differences. The primary test runs on each metric's own scale, where IBS spans 0.06 on AIDS to 0.20 on SUPPORT and the between-seed spread spans a factor of twenty, so a cohort with noisy seeds can dominate the $m = 6$ mean. Rerunning every comparison on the standardised effect $\Delta/\mathrm{SD}_{\mathrm{seed}}$, the units §5.3 already reports ablations in, leaves five comparisons below 0.05 rather than four, and not the same set: tdAUC over CoxTime ($0.100 \to 0.034$) and over RSF ($0.179 \to 0.041$) and resolution over RSF ($0.065 \to 0.039$) enter, while tdAUC over DeepHit ($0.033 \to 0.122$) and resolution over DeepSurv ($0.036 \to 0.070$) leave; IBS over DeepHit and over CoxTime resolve on both scales. Choosing the scale after seeing this would be a fork, so the raw scale stays primary and the standardised one is reported as a sensitivity check (`effect_size_sensitivity.py`).

Table 6: **The same comparisons at both units of analysis.** Pooled columns treat all sixty (cohort, seed) cells as replications; cohort columns take the cohort as the unit, each contributing its mean paired difference. ICC is the intra-cohort correlation of those differences and $n_{\text{eff}} = n/(1 + (k - 1)\text{ICC})$ is what the pooled $n = 60$ was worth. Δ is always baseline $-$ ours, so it favours our model when it is positive for IBS (lower is better) and when it is negative for tdAUC, concordance and resolution (higher is better); the *cohorts* column always counts cohorts won by us, so read the two together. Sign-flip p cannot fall below $2^{1-6} = 0.031$, so no isolated comparison can be starred here; the column is given for completeness.

Metric	Baseline	Δ	cohorts	p_{cell}	p_{cohort}	ICC	n_{eff}
IBS ↓	DeepHit	+0.0122	6/6	$< 10^{-3}$	0.011	0.67	8.5
	CoxTime	+0.0054	6/6	$< 10^{-3}$	0.025	0.28	16.9
	GBSA	+0.0102	6/6	$< 10^{-3}$	0.078	0.78	7.5
	RSF	+0.0061	5/6	$< 10^{-3}$	0.080	0.64	8.8
	LogNAFT	+0.0048	4/6	$< 10^{-3}$	0.221	0.80	7.3
	CoxPH	+0.0040	3/6	$< 10^{-3}$	0.239	0.74	7.8
	WeibAFT	+0.0041	4/6	$< 10^{-3}$	0.240	0.75	7.7
	CoxCC	+0.0032	5/6	$< 10^{-3}$	0.270	0.56	9.9
	DeepSurv	+0.0016	5/6	0.008	0.582	0.50	10.9
mean tdAUC ↑	DeepHit	-0.0348	6/6	$< 10^{-3}$	0.033	0.46	11.7
	CoxTime	-0.0216	5/6	$< 10^{-3}$	0.100	0.43	12.4
	DeepSurv	-0.0133	4/6	$< 10^{-3}$	0.125	0.27	17.6
	GBSA	-0.0357	4/6	$< 10^{-3}$	0.154	0.64	8.8
	RSF	-0.0096	5/6	$< 10^{-3}$	0.179	0.32	15.6
	CoxCC	-0.0239	4/6	$< 10^{-3}$	0.189	0.55	10.1
	LogNAFT	-0.0111	4/6	0.006	0.374	0.68	8.4
	CoxPH	-0.0105	3/6	0.009	0.395	0.73	8.0
	WeibAFT	-0.0093	3/6	0.016	0.472	0.72	8.0
C_{IPCW} ↑	CoxTime	-0.0154	4/6	0.019	0.266	0.59	9.6
	GBSA	-0.0204	3/6	0.043	0.369	0.78	7.5
	WeibAFT	+0.0043	2/6	0.457	0.611	0.63	9.0
	CoxCC	-0.0072	2/6	0.566	0.622	0.61	9.2
	RSF	-0.0053	4/6	0.041	0.642	0.71	8.1
	CoxPH	+0.0035	2/6	0.484	0.655	0.58	9.7
	LogNAFT	+0.0020	4/6	0.648	0.778	0.47	11.5
	DeepHit	-0.0015	2/6	0.051	0.841	0.33	15.2
	DeepSurv	-0.0010	2/6	0.211	0.919	0.43	12.4
RES ↑	DeepSurv	-0.0042	5/6	$< 10^{-3}$	0.036	0.28	16.9
	RSF	-0.0040	5/6	$< 10^{-3}$	0.065	0.46	11.6
	DeepHit	-0.0038	5/6	$< 10^{-3}$	0.084	0.43	12.3
	CoxTime	-0.0051	5/6	$< 10^{-3}$	0.089	0.60	9.4
	GBSA	-0.0031	4/6	$< 10^{-3}$	0.115	0.27	17.7
	CoxCC	-0.0035	4/6	$< 10^{-3}$	0.160	0.43	12.2
	CoxPH	-0.0022	4/6	0.005	0.475	0.65	8.7
	WeibAFT	-0.0018	4/6	0.015	0.550	0.71	8.1
	LogNAFT	-0.0005	4/6	0.080	0.900	0.75	7.7

APPENDIX F THE TWO ABLATIONS: WHAT WE CAN AND CANNOT CLAIM

F.1 THE CORRECTED PROPOSAL AND ITS RATE

§3.3 states the slack of the working proposal and the size of the correction; the correction itself is defined here, with the consistency rate that follows from it.

Definition 2 (Σ-corrected proposal)**.** The Σ-corrected working proposal replaces the forward variance σ^2_{post} by the Laplace variance Σ of Definition 1:

$$q^{\text{work},\Sigma}(z_0 \mid x, z_s, s_{\text{imp}}) \;:=\; \mathcal{N}\big(\hat{z},\, \Sigma(z_s, x)\big)\big|_{z_0 \geq \log T^{\text{obs}}}. \tag{22}$$

In implementation Σ is replaced by $\widehat{\Sigma}$ of equation 9 throughout.

Table 7: **Holdout IBS, one column per cohort**, the unit the cohort-level tests use. Mean over ten seeds, with the between-seed standard deviation on our arm; **bold** is the best competing model in a column. FATUM$_\Sigma$ (gray) is our own ablation.

Method	AIDS	METABRIC	GBSG2	SUPPORT	FLCHAIN	Rotterdam
FATUM	0.061 $\pm$0.004	**0.157** $\pm$**0.002**	0.174 $\pm$0.011	**0.192** $\pm$**0.001**	0.095 $\pm$0.001	**0.184** $\pm$**0.002**
FATUM$_\Sigma$	0.061	0.157	0.174	0.191	0.095	0.184
CoxPH	**0.060**	0.169	0.169	0.204	**0.094**	0.190
CoxCC	0.066	0.170	0.168	0.193	0.096	0.188
CoxTime	0.066	0.168	0.184	0.192	0.097	0.188
WeibAFT	0.060	0.167	0.168	0.206	0.095	0.191
LogNAFT	0.061	0.169	0.168	0.208	0.097	0.188
RSF	0.061	0.171	0.172	0.202	0.107	0.187
GBSA	0.070	0.188	0.184	0.193	0.097	0.191
DeepSurv	0.062	0.169	**0.166**	0.193	0.097	0.186
DeepHit	0.086	0.170	0.182	0.203	0.098	0.197

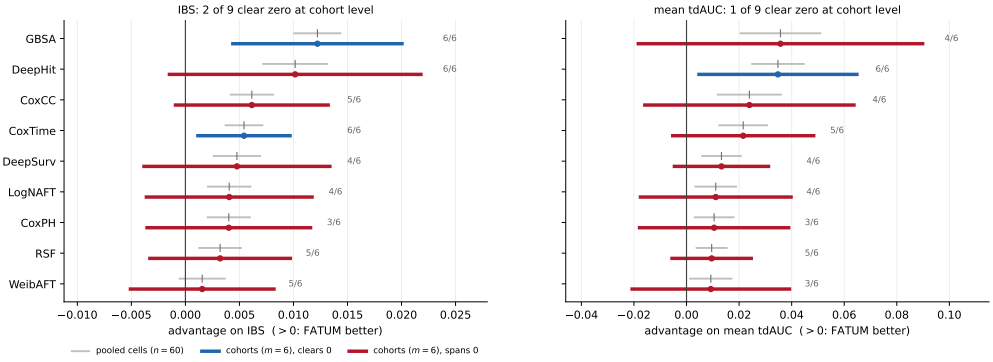


Figure 5: **What the unit of analysis decides.** Advantage of FATUM over each baseline on IBS, with the pooled cell-level interval (thin grey, $n = 60$) beside the cohort-level one (thick, $m = 6$). The pooled interval clears zero for nine of ten comparisons, the cohort-level one for two (blue); counts on the right are cohorts won.

Theorem 3 (Σ-corrected proposal slack)**.** *Under the conditions of Theorem 1,*

$$\mathrm{KL}(q^{\mathrm{work},\Sigma} \,\|\, q_s) \;\le\; \mathcal{K} + O(\sigma_{\mathrm{post}}^6\, g_3^2) + O(\sigma_{\mathrm{post}}^8 g_4^2 + \sigma_{\mathrm{post}}^{12} g_3^4),$$

with $\mathcal{K}$ the location term of Theorem 1, carried unchanged because the correction moves the proposal's scale and not its centre. It is the leading contribution wherever $u^\star$ is of order one (Remark 2).

The $O(\sigma_{\mathrm{post}}^5 |g_3| V_{\mathrm{prior}})$ boundary cross term of Theorem 1 is *absent* here, and for a structural reason rather than a bookkeeping one: it arises from evaluating a half-line Hermite residual under a proposal whose scale differs from the target's, and $q^{\mathrm{work},\Sigma}$ shares the scale Σ with r by construction. The correction therefore removes two terms, the leading $C_{\mathrm{var}} \sigma_{\mathrm{post}}^4 V_{\mathrm{prior}}^2$ and this one, and the second removal is the one that improves the *order* of the remainder rather than only its coefficient.

Proof in Appendix B.5.

Remark (proposal variance). Contribution (a) vanishes *exactly* only when the proposal scale equals the true posterior curvature V_{prior}. The implemented arm substitutes the Jacobian estimate $\widehat{V}$ of equation 9, a plug-in in the spirit of ΠGDM's fixed-precision approximation (Song et al., 2023), which replaces the leading coefficient $C_{\mathrm{var}} V_{\mathrm{prior}}^2$ by $C_{\mathrm{var}}(V_{\mathrm{prior}} - \widehat{V})^2$. The term is reduced, and removed only where $\widehat{V} = V_{\mathrm{prior}}$; the parameterisation cannot widen the proposal, so on the 0.3% to 8.9% of forward-noised states where it would have to, it clamps instead.

Table 8: **The four comparisons that resolve at the cohort level.** Advantage is oriented so that positive favours our model on every row, with the paired Student-t interval over the $m = 6$ cohort means beside it. The cluster bootstrap agrees on all four and is wider nowhere. Paragraph (P1) of §5.2 states them in the text.

Metric	vs.	Advantage	p	Cohorts
IBS	DeepHit	$+0.0122$ $[+0.0042, +0.0202]$	0.011	$6/6$
IBS	CoxTime	$+0.0054$ $[+0.0010, +0.0098]$	0.025	$6/6$
tdAUC	DeepHit	$+0.035$ $[+0.004, +0.066]$	0.033	$6/6$
Resolution	DeepSurv	$+0.0042$ $[+0.0004, +0.0080]$	0.036	$5/6$

Corollary 1 (Σ-corrected consistency rate)**.** *Under the conditions of Theorems 3 and 2, the Σ-corrected proposal replaces the slack of Theorem 2(i) by*

$$\bar{\varepsilon}_{\mathrm{TV}}^{\Sigma} \;=\; O\big(\sigma_{\mathrm{post}}^{3}\,\bar{g}_3\big) + \alpha,$$

one order in σ_{post} better than the uncorrected $O(\sigma_{\mathrm{post}}^2 \bar{V}) + \alpha$, both read modulo the score-error floor $\sqrt{\mathcal{K}/2}$, which is $\approx |\delta_\theta|/2$ only where the boundary is far, and that neither correction nor schedule can move (Remark 4), and correspondingly relaxes the admissible schedule in Theorem 2(iii) from $n_{\mathrm{gibbs}}\sigma_{\mathrm{post}}^2 \to 0$ to $n_{\mathrm{gibbs}}\sigma_{\mathrm{post}}^3 \to 0$.

Proof in Appendix B.6.

F.2 Why the Σ null is informative

The GRAD arm instantiates Theorem 3 with the curvature read from the denoiser Jacobian, so nothing in it is tuned. The trunk is factorised, so the cross-channel Jacobian is zero and the corrected variance is the within-channel second moment: the correction runs where its derivation is exact, which is what makes a null informative.

Three measurements. The rate is confirmed, the closed-form KL tracking the predicted exponent over four decades in all six cohorts, 1.37–1.76 measured against 1.42–1.88 predicted (Fig. 6); at the operating points the divergence it removes is 9×10^{-4} to 5×10^{-2} nats. The benefit is confirmed where it is observable, growing linearly in s_{imp} on all six metric-by-channel combinations (Fig. 7). On the benchmark cohorts the discriminative effect is absent, below the between-seed standard deviation in every cohort, which is what §5.3 predicts of a change confined to the latent proposal. The one real-data signature is tail-sensitive: the normalised Wasserstein distance falls by 14% in the cohort median and 44% in the mean, with every downstream metric unmoved.

Two things bound what the null can mean. The plug-in runs $\widehat{\Sigma} = \sigma_{\mathrm{post}}^2 J$ with $J = \mathrm{clip}(\partial_{z_s}\hat{z}_\theta, 10^{-4}, 1)$, so it can only narrow the proposal, and the upper clamp binds on 3.4% to 58.9% of the states the chain visits at the selected s_{imp}, setting $\widehat{V} = 0$ so that the corrected arm proposes what the uncorrected one proposes. Part of the null is arithmetic. And the measured curvature drifts with the probe scale as s_{imp}^a, $a \in [-0.29, -0.06]$, which is why the observed exponent is $2 + 2a$ and not 2.

F.3 Both ablations, cohort by cohort

§5.3 reports the two ablations as macro effects; the breakdown is here, on one protocol and in one table. The cohort that prefers the ablated arm in panel (a), GBSG2, has the widest between-seed spread of the six, so its sign is the least determined: 0.25 between-seed standard deviations against 1.40 on FLCHAIN, where the full model leads. Panel (b) is the Σ correction on the same scale, an order of magnitude smaller and without a consistent sign, which is what §5.1 predicts for operating points that straddle the zero-crossing.

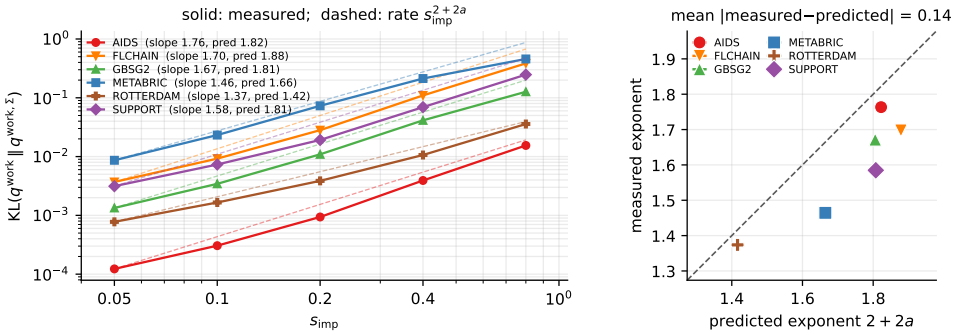


Figure 6: **The variance-mismatch term scales at the predicted rate in all six cohorts.** Left: median $\mathrm{KL}(q^{\mathrm{work}} \| q^{\mathrm{work},\Sigma})$ against s_{imp}, with each cohort's dashed guide at the exponent $2 + 2a$ that Theorem 1 predicts once the measured drift $V_{\mathrm{prior}} \propto s_{\mathrm{imp}}^{a}$ is accounted for. Right: measured exponent against predicted. The median over subjects is monotone in all six cohorts where the mean is not, the truncation term dominating a handful of subjects at large s_{imp}.

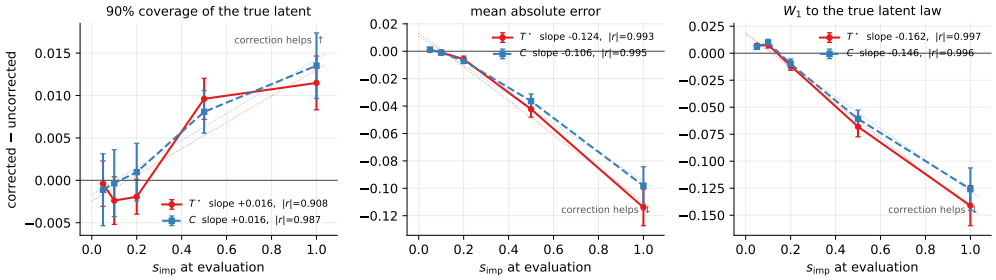


Figure 7: **The Σ-correction's benefit grows linearly in s_{imp}, as the analysis predicts.** Corrected minus uncorrected arm on the simulation with known latents, paired on channel, seed, coupling mode and s_{imp} (300 pairs, 30 per grid point); bars are 95% paired intervals, dotted lines least-squares fits. The leading bias $O(\sigma_{\mathrm{post}}^2 V_{\mathrm{prior}})$ is linear in s_{imp}, and the gap widens accordingly on all six combinations of metric and channel. The least-squares fits give $|r| = 0.987$–0.997 on five of them and 0.908 on $T^\star$ coverage, whose gap is an order of magnitude smaller in absolute size than either error metric's.

APPENDIX G GENERATED COHORTS, COHORT BY COHORT

Table 2 summarises the six cohorts by their median. The two figures below give them one at a time: which cohorts favour the cascade for every metric and generator, and how each generator's Kaplan–Meier curve sits against the holdout.

Table 9: **The two ablations, cohort by cohort.** In both panels ΔIBS is the *variant* minus FATUM, the arm reported everywhere else, and seed SD is FATUM's between-seed standard deviation. In (a) the variant is the ablation, so positive favours FATUM; in (b) the variant *adds* the Σ correction, so negative favours FATUM$_\Sigma$. The *Better* column names the winner directly and is the column to read. Ten seeds per cohort, one shared split, sixty cells per panel, none identical to its control.

Cohort	ΔIBS	seed SD	$\lvert\Delta\rvert$/SD	Better
(a) without the censoring channel				
FLCHAIN	$+0.00082$	0.00058	1.40	full
AIDS	$+0.00339$	0.00431	0.79	full
Rotterdam	$+0.00131$	0.00172	0.76	full
SUPPORT	$+0.00088$	0.00116	0.76	full
GBSG2	-0.00278	0.01114	0.25	without C
METABRIC	$+0.00016$	0.00184	0.08	full
(b) with the Σ correction added, i.e. FATUM$_\Sigma$ against FATUM				
FLCHAIN	-0.00006	0.00058	0.10	FATUM$_\Sigma$
AIDS	$+0.00007$	0.00431	0.02	tie
Rotterdam	$< 10^{-5}$	0.00172	0.00	tie
SUPPORT	-0.00040	0.00116	0.35	FATUM$_\Sigma$
GBSG2	-0.00009	0.01114	0.01	tie
METABRIC	$< 10^{-5}$	0.00184	0.00	tie

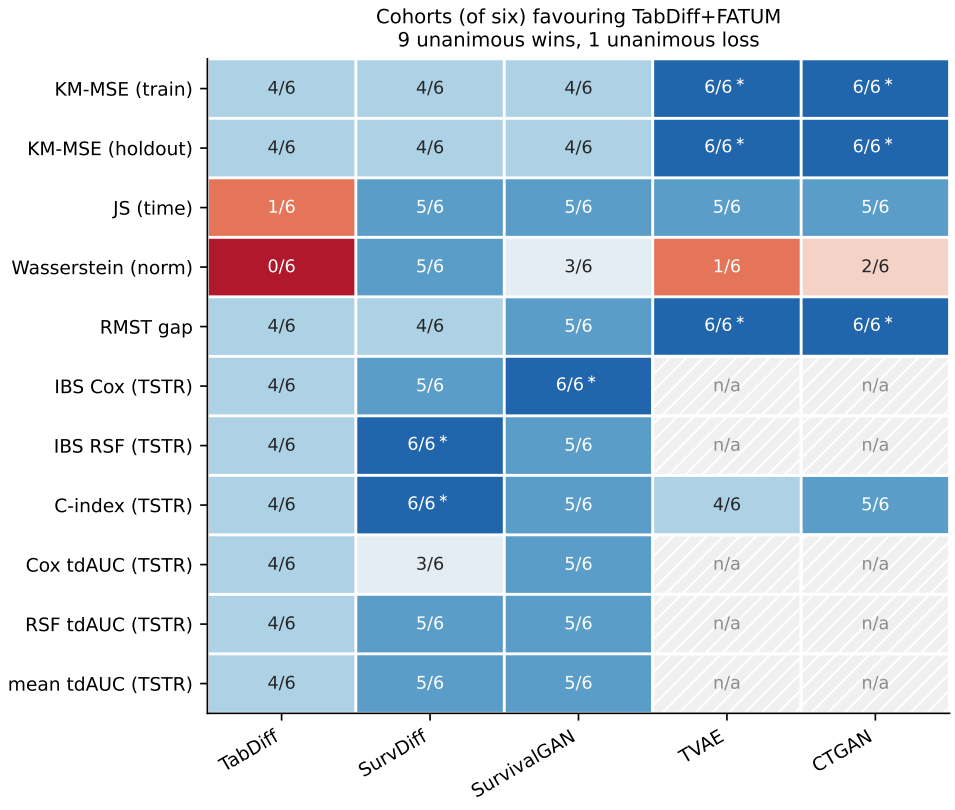


Figure 8: **Cohorts favouring the cascade, by metric and generator.** Seeds are averaged within a cohort before counting, and asterisks mark the unanimous cells, which is the only statistic the 0.031 floor of §4.1 leaves. Red marks the cells we lose, and they are all marginal time fidelity, worst against the backbone that supplies our covariate sampler and resamples times inside the training support. Hatched cells are unavailable rather than tied: TVAE and CTGAN emit non-positive times in $57/60$ and $54/60$ cells, up to 17.9% of a cell, after which their censored draws stop short of the evaluation grid and the IPCW-weighted quantities cannot be formed. No other generator produces $t \leq 0$.

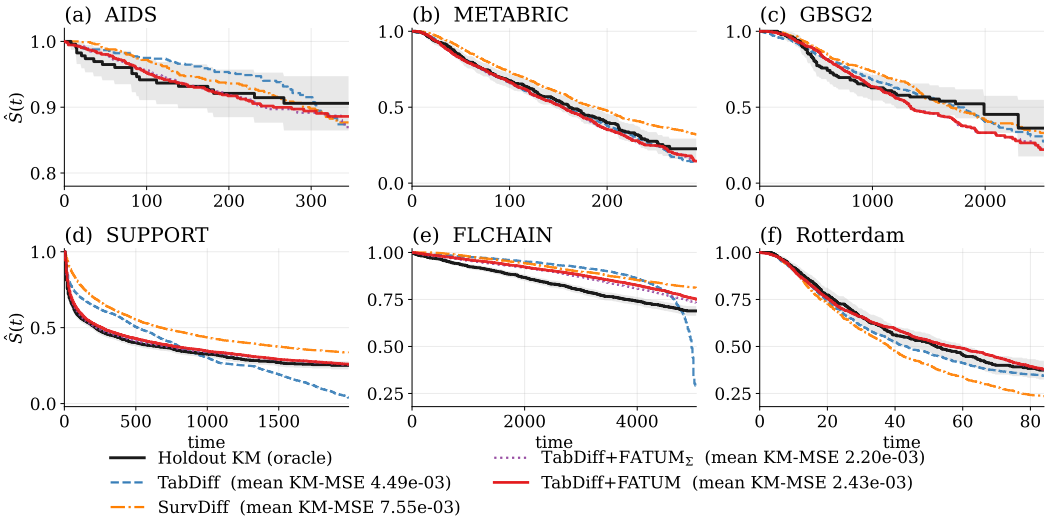


Figure 9: **Synthetic cohort KM fidelity (seed 6098).** Holdout KM (black) against TabDiff (blue, dashed), SurvDiff (orange, dash-dot), TabDiff+FATUM (red) and its Σ-corrected ablation (purple, dotted), for all six cohorts. Legends give the mean KM-MSE; y-limits are per cohort.