# EMBER: LEARNING HIDDEN LESION DYNAMICS FOR MULTIPLE SCLEROSIS PROGRESSION PREDICTION FROM INTERMITTENT BRAIN MRI

**Anonymous authors**
Paper under double-blind review

## ABSTRACT

Routine brain MRI records when lesions are visible, but not when they arose or which activity preceded them. In multiple sclerosis this limits progression prediction, because lesions detected at the same visit may reflect different event orders or a shared hidden disease process. EMBER learns from these alternative explanations and predicts progression independent of relapse activity. A joint model infers uncertain lesion timing and hidden common activity, allows regional relationships to have different temporal profiles, and uses longitudinal tissue changes to constrain the inferred histories; progression risk is then averaged over the histories that the available scans support. To keep that risk stable when imaging is incomplete, a posterior-refinement objective trains the predictor to update coherently as scan availability changes, so that after an intermediate scan is removed the prediction matches the posterior average over compatible finer observations even though an individual finer observation may raise or lower risk. Coarsening identities and a Brier-risk decomposition characterize this relationship and its expected information cost. Public MRI from 195 participants tests regional inference, lesion observations, and change forecasting, while a two-center cohort of 240 adults tests clinical progression. EMBER predicts 24-month relapse-independent progression from routine T1-weighted and FLAIR MRI with held-out-site AUROC 0.835 and Brier score 0.116.

## 1 INTRODUCTION

Multiple sclerosis (MS) can accumulate disability without an overt clinical relapse. Progression independent of relapse activity (PIRA) contributes substantially to disability accumulation even during the relapsing phase (Kappos et al., 2020). MRI provides complementary evidence: new focal lesions reflect inflammatory activity, while tissue loss carries information about progression that may occur without new lesions (Cagol et al., 2022). A useful model must combine these signals over time, including patients whose recent lesion count is zero.

MRI reveals only fragments of disease evolution: if two lesions are absent at month 0 and visible at month 12, either could have appeared first, or both could reflect shared hidden activity. Figure 1 illustrates how EMBER carries multiple compatible lesion histories into PIRA prediction.

We introduce EMBER (**E**vent **M**arginalization for **B**rain-imaging **E**volution and **R**isk), a partially observed Hawkes model (Hawkes, 1971) that jointly models lesion timing, hidden activity, and tissue evolution for prognosis. Lesion intervals and tissue measurements constrain the possible histories, while regional relationships can have different temporal profiles. Averaging history-specific risks with posterior weights yields the PIRA prediction.

A second question follows naturally: how should risk change when a scan is unavailable? An intermediate MRI can raise or lower an individual's estimated risk by revealing which histories are more plausible. Without that scan, the prediction should average over the possible observations that could have been obtained. We call this relationship *refinement coherence*. It allows new imaging to revise risk while keeping predictions made with different amounts of information consistent.



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.

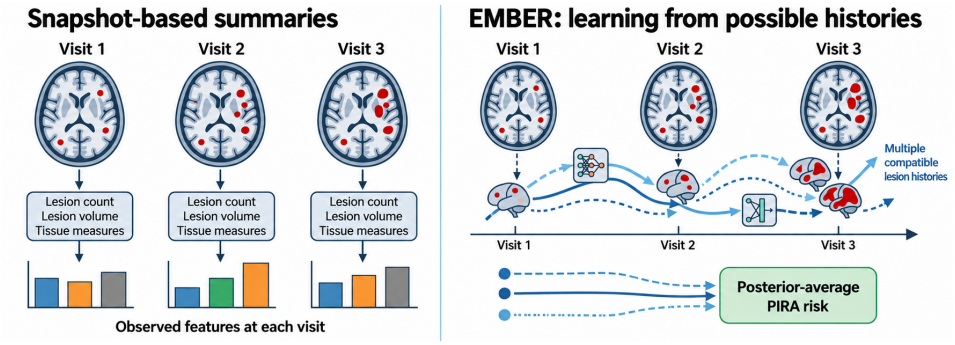



Figure 1: **From MRI snapshots to posterior-average PIRA risk.** Left: lesion counts, volumes, and tissue measures summarize individual visits. Right: the same scans support multiple lesion histories. EMBER infers these histories and averages their history-specific PIRA risks using posterior weights.

Two uncertainties shape this problem. The more immediate concerns lesion chronology: we jointly infer unresolved lesion timing and hidden common activity, allow temporal effects to differ across regional relationships, and use tissue evolution to constrain the inferred histories, so that alternative explanations of the observed MRI reach progression prediction through posterior risk averaging. The other concerns scan availability. Here we introduce a posterior-refinement objective that trains the prediction from retained scans to match the posterior average over compatible finer observations, and show that this relationship is learnable with an amortized predictor and finitely many sampled histories. Coarsening identities and a Brier-risk decomposition specify its target and characterize the expected predictive information lost when a scan is removed.

We first ask whether the inferred histories explain imaging that the model was not shown, using hidden-region and hidden-visit prediction over intervals that contain incident lesions. We then remove matched scans and ablate the refinement objective to see how risk responds to reduced imaging information, and use component and mechanism controls to connect these results to uncertain timing, hidden activity, and tissue evolution. Public MRI from 195 participants and a two-center cohort of 240 patients link this evidence to lesion forecasting and PIRA prediction at an independent center. Our code is available at https://anonymous.4open.science/r/EMBER-0F2F/README.md.

## 2 Related Work

**Lesion and tissue modeling in MS.** Standard MRI protocols support longitudinal lesion assessment with three-dimensional FLAIR and structural imaging (Wattjes et al., 2021). MRI-based disease subtypes can characterize heterogeneity beyond conventional clinical phenotypes (Eshaghi et al., 2021), and atrophy is associated with relapse-independent progression (Cagol et al., 2022). Our task is individual landmark prognosis from a short sequence of routine scans. Lesion counts, lesion growth, and regional tissue loss enter a shared temporal model rather than being treated as interchangeable measures of activity.

**Partially observed event processes.** Interval-censored Hawkes estimation has been studied through mean-behavior processes (Rizoiu et al., 2022), Monte Carlo estimation for binned observations (Shlomovich et al., 2022; 2021), and Bayesian inference from aggregated data (Zhou & Papadogeorgou, 2025). These works establish that interval observations can be modeled without inventing precise event times. Hidden-structure discovery uses rank constraints and, more recently, path-based conditions for complex latent networks (Huang et al., 2022; Jin & Huang, 2026). Building on these approaches, EMBER jointly models unknown event chronology, hidden common activity, heterogeneous temporal effects, and longitudinal imaging emissions. Posterior averaging then links these possible histories to progression risk, and a separate training objective constrains how that risk changes with observation resolution.

**Irregular-time representation learning.** GRU-D (Che et al., 2018), latent ODE models (Rubanova et al., 2019), and neural controlled differential equations (Kidger et al., 2020) process irregular measurements. Transformer Hawkes processes model flexible event dependencies (Zuo et al., 2020).

A missing measurement and an interval-censored event are different objects: the latter constrains an unobserved event to have occurred within a known interval. EMBER preserves this constraint and supervises how risk changes when several such intervals are merged.

## 3 CLINICAL TASK AND OBSERVATION MODEL

For patient $p$, let $0 = v_{p0} < \cdots < v_{pK} = T$ denote MRI visits in a 24-month look-back window. The final scan is the prediction landmark. Registration and longitudinal lesion matching identify each incident lesion $e$, its anatomical compartment $i_e$, and the interval $I_e = (a_e, b_e]$ between its last negative and first positive scan. A lesion receives one birth event. Its subsequent enlargement is a longitudinal measurement, not another birth. Lesions already present at the first scan enter the initial state.

We use $d = 12$ mutually exclusive bilateral compartments. Observations $X_p(v_k)$ include regional lesion volumes, matched-lesion growth, interval lesion-change volume, brain parenchymal fraction, and normalized thalamic volumes. Covariates $c_p(t)$ include age, disease duration, recorded relapses, disability, and treatment history available by time $t$. Imaging-derived quantities are measured at their actual scan dates; a volume first measured at $b_e$ is never used as a covariate at an earlier putative birth time.

The outcome $Y_p$ is the first PIRA onset within 24 months after $T$, with disability worsening sustained for 12 months. Outcome ascertainment therefore extends to month $T + 36$. EDSS baselines are updated after qualifying relapses, prior progression, and confirmed improvements, following the standardized framework of Müller et al. (2025). The predictor receives no post-landmark scans, treatment changes, or disability measurements. It estimates prognosis under the observed care setting.

Our primary analysis uses scheduled surveillance MRI, conditioning on scan times as the acquisition design. The event list includes incident lesions that remain visible and matchable through the landmark; transient or ambiguous lesions contribute visit-level burden only. Deleting an intermediate scan widens the affected birth intervals, while the retained examinations continue to define the admissible histories. The event target is persistent MRI-visible focal-lesion appearance at clinical imaging resolution.

## 4 THE EMBER FRAMEWORK

Figure 2 connects the inference and prediction pipeline with refinement training. We first infer lesion histories compatible with the available MRI and average their progression risks. During training, retained scans also define possible finer observations whose average prediction supplies the refinement target.

### 4.1 HIDDEN LESION DYNAMICS WITH HETEROGENEOUS MEMORY

Let $Z = (N_1, \ldots, N_d, L_1, \ldots, L_q)$ combine regional new-lesion counting processes and hidden activity processes. The latter capture common temporal influences beyond measured clinical covariates. We model all components by

$$\lambda_a(t) = \mu_a\{c_p(t)\} + \sum_{b=1}^{d+q} \sum_{\ell=1}^{J} C_{ab}^{(\ell)} \int_0^{t^-} \kappa_\ell(t-u)\, dZ_b(u) + \xi_a(t), \tag{1}$$

where $C^{(\ell)} \geq 0$, $\kappa_\ell(s) = \beta_\ell e^{-\beta_\ell s}\mathbf{1}(s > 0)$, and $\xi_a(t)$ is the decaying contribution of the initial state. Background rates are positive functions of observed covariates. The three basis time constants are 3, 9, and 24 months; edge-specific weights allow different response shapes. The full block matrix includes observed-to-latent and latent-to-latent interactions.

We constrain the maximum column sum of $\bar{C} = \sum_\ell C^{(\ell)}$ to be below $\rho = 0.85$. Thus the spectral radius is smaller than one. With bounded background rates, this gives a stable cluster representation and finite expected event counts (Hawkes & Oakes, 1974). A group penalty jointly shrinks incoming and outgoing blocks for each latent component; validation selects the retained dimension from $q \in \{0, 1, 2\}$ by held-out imaging likelihood. Latent components therefore summarize common

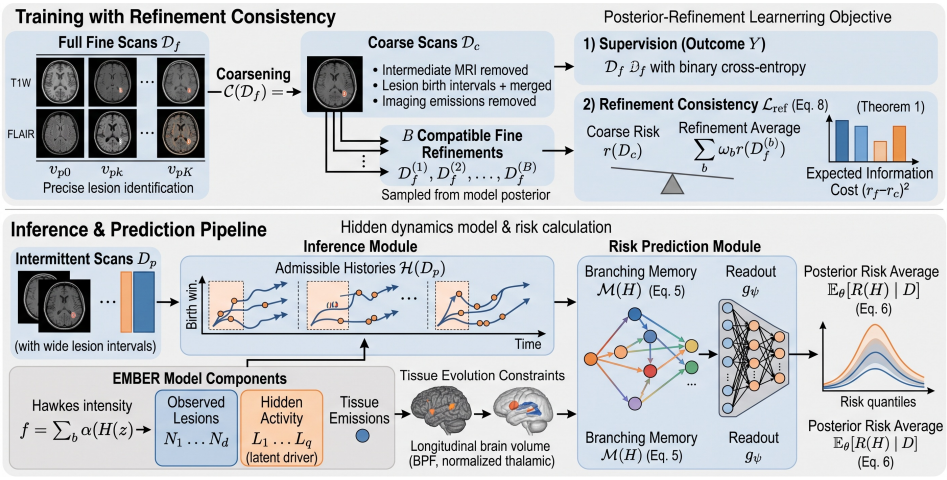



Figure 2: **EMBER training and prediction framework.** Top: removing intermediate scans yields coarse observations, and compatible finer ones are sampled from the posterior; outcome supervision trains the risk readout, and posterior-refinement learning aligns coarse risk with the weighted average of finer risks. Bottom: regional lesion events, hidden activity, and tissue measurements constrain admissible histories, which branching memory summarizes into risks averaged under the posterior. Eqs. (1), (5), (6), and (8) define the intensity, branching responsibilities, risk, and refinement objective.

temporal influences beyond the measured covariates, and enter the model when that scoring supports them.

The exponential traces $S_{b\ell}(t) = \int_0^{t^-} e^{-\beta_\ell(t-u)} dZ_b(u)$ evolve exactly between events. An imaging emission model links a small neural mapping of these traces, initial lesion burden, elapsed time, and observed covariates to the continuous MRI features $X_p(v_k)$. Signed tissue features use Gaussian emissions; nonnegative volumes use censored log-Gaussian emissions. Acquisition-specific scales are fitted on development scans.

### 4.2 Marginalizing the chronology that MRI does not observe

Write $D_p = (\{I_e, i_e\}_e, X_p, \mathcal{V}_p, c_p)$ and let $H_p$ contain lesion birth times, latent event paths, initial traces, and tissue-change effects. Its admissible set $\mathcal{H}(D_p)$ enforces $t_e \in I_e$ and the recorded event identities. The observed likelihood is

$$p_\theta(D_p \mid \mathcal{V}_p, c_p) = \int_{\mathcal{H}(D_p)} p_\theta(H_p, X_p \mid \mathcal{V}_p, c_p)\, dH_p. \tag{2}$$

Integration includes summation over latent event counts. Conditional on a complete path, the log density has the usual event-intensity and compensator terms for *all* observed and latent components, plus imaging emissions and the initial-state density.

An interval-overlap feature makes posterior proposals sensitive to unresolved order. For source interval $I = [a, b]$ and target interval $J = [c, d]$, define

$$G_\beta(I, J) = \frac{1}{|I||J|} \int_a^b \int_c^d \mathbf{1}(v > u) \beta e^{-\beta(v-u)}\, dv\, du. \tag{3}$$

With $F_\beta(x) = x_+ - (1 - e^{-\beta x_+})/\beta$, this equals

$$G_\beta(I, J) = \frac{F_\beta(d-a) - F_\beta(d-b) - F_\beta(c-a) + F_\beta(c-b)}{|I||J|}. \tag{4}$$

For a common interval of width $\Delta$, it is $[\Delta - (1 - e^{-\beta\Delta})/\beta]/\Delta^2$. A tied midpoint representation would discard this positive expected interaction. Equation (3) initializes a proposal under uniform timing, and final posterior weights are then assigned by the complete path likelihood.

### 4.3 From possible lesion histories to progression risk

Each possible history gives a different account of the activity preceding a lesion. Branching memory summarizes these accounts for progression prediction. For a complete history (Veen & Schoenberg, 2008), the responsibility of earlier observed lesion $f$ for lesion $e$ is

$$\pi_{ef}(H) = \frac{\sum_\ell C^{(\ell)}_{i_e i_f} \kappa_\ell(t_e - t_f)}{\lambda_{i_e}(t_e)}, \qquad \pi_{eL}(H) = \frac{\sum_{r,\ell} C^{(\ell)}_{i_e, d+r} \beta_\ell S_{d+r,\ell}(t_e)}{\lambda_{i_e}(t_e)}. \tag{5}$$

Background and initial-state responsibilities complete the normalization. We summarize regional responsibility flows into $M_{ij\ell}(H)$ using lag-basis weights, and separately pool latent-parent mass, regional terminal traces, and recent-versus-persistent activity. This branching memory distinguishes shared activity from supported regional excitation.

A two-layer readout $g_\psi$ combines the memory, landmark tissue features $X_p(T)$, and landmark clinical covariates. The history-specific risk is $R(H) = \mathrm{sigmoid}\{g_\psi(M(H), S(T), X_p(T), c_p(T))\}$. Prediction averages the nonlinear risk:

$$r_{\theta,\psi}(D_p) = \mathbb{E}_\theta[R(H_p) \mid D_p] \simeq \sum_{s=1}^{S} w_{ps} R(H_{ps}), \quad \sum_s w_{ps} = 1. \tag{6}$$

Taking the risk of an averaged history would erase posterior alternatives before applying the nonlinear clinical mapping. An interval encoder amortizes admissible proposals $q_\phi(H \mid D)$, with importance weights proportional to $p_\theta(H, D)/q_\phi(H \mid D)$. Quantiles of $R(H)$ quantify the risk variation attributable to unresolved histories.

### 4.4 Learning risk updates when scans are missing

Let $D_c = C(D_f)$ delete an intermediate scan, merge the adjacent birth intervals, and remove its imaging emission. Boundary scans, lesion identities, and landmark covariates are retained. The missing scan could have supported several finer observations, each with its own predicted risk. Under a fixed joint model, the coarse prediction averages over them:

$$r(D_c) = \mathbb{E}_{D_f \mid D_c}\big[r(D_f)\big]. \tag{7}$$

Our posterior-refinement objective trains an amortized predictor to respect this identity. For each coarse history, $B$ compatible refinements are drawn from the model posterior at the withheld acquisition times, including their imaging emissions. We penalize

$$\mathcal{L}_{\mathrm{ref}} = E_{D_c}\left[\widehat{r}(D_c) - \mathrm{sg}\left\{\sum_{b=1}^{B} \omega_b \widehat{r}(D_f^{(b)})\right\}\right]^2, \tag{8}$$

where $\omega_b$ are normalized posterior weights and sg stops teacher gradients. Each teacher is evaluated under the same fixed model snapshot. The weighted fine-risk mean supplies a training target for the coarse prediction, so that the amortized predictor stays consistent across observation resolutions with finitely many sampled histories.

Training alternates constrained posterior sampling and generalized EM updates (Dempster et al., 1977) for the imaging-event model, followed by risk learning with binary cross-entropy and Eq. (8). The outcome is excluded from the patient posterior used by the predictor. The latent dimension, penalties, and refinement weight are selected using development validation only. The final predictor uses 128 weighted histories; training uses 64 and $B = 8$ refinements. Appendix C specifies posterior proposals and the objective.

## 5 Identities induced by scan removal

Under a fixed joint model, conditional-expectation identities (Durrett, 2019, Sec. 4.1) relate predictions made with different amounts of imaging information.

**Proposition 1** (Observation projectivity)**.** *Assume scan coarsening is determined by the acquisition design, the observation maps preserve event identities, and $D_c = C(D_f)$. The marginal likelihood in Eq.* (2) *satisfies $p_\theta(D_c) = \int_{C(D_f)=D_c} p_\theta(D_f)\, dD_f$, with sums replacing integrals for discrete observations. This holds for heterogeneous kernels and latent interactions.*

The identity follows by partitioning the same latent-history space; assigning each resolution a different deterministic event history would break it. Stability of Eq. (1) keeps the path model well-defined at every observation resolution.

**Theorem 1** (Refinement coherence and predictive information)**.** *Let $\mathcal{F}_c \subseteq \mathcal{F}_f$ be the information in coarse and fine observations, and let $r_j = \mathbb{E}[Y \mid \mathcal{F}_j]$ for a binary outcome. Then*

$$r_c = \mathbb{E}[r_f \mid \mathcal{F}_c], \tag{9}$$

$$\mathbb{E}[(Y - r_c)^2] - \mathbb{E}[(Y - r_f)^2] = \mathbb{E}[(r_f - r_c)^2] \geq 0. \tag{10}$$

*For $R(H)$ in Eq.* (6)*, posterior history-risk variance obeys*

$$\mathrm{Var}(R \mid \mathcal{F}_c) = \mathbb{E}[\mathrm{Var}(R \mid \mathcal{F}_f) \mid \mathcal{F}_c] + \mathrm{Var}(r_f \mid \mathcal{F}_c). \tag{11}$$

Equation (10) expresses the expected Brier improvement as the variation in risk revealed by finer observations. Equation (11) similarly describes history-risk uncertainty in conditional expectation. Proofs and numerical checks appear in Appendix B.

# 6 Experiments

## 6.1 Experimental setup

The clinical study includes 240 adults with relapsing-remitting MS. Center A provides 128 training and 32 validation participants; center B provides 80 independent test participants with 15 PIRA outcomes. Inputs are routine T1-weighted and FLAIR MRI and clinical history from the 24-month look-back. Eligibility requires at least three usable surveillance scans. The endpoint is PIRA onset within 24 months after the landmark, with subsequent 12-month confirmation. All model selection and risk calibration use center A.

The public component comprises MSLesSeg (75 patients), MSSEG-2 (100), and Ljubljana longitudinal (20) (Guarnera et al., 2025; Masson et al., 2026; Lesjak et al., 2016). Their tasks evaluate shared image measurements, hidden regional counts, and prospective lesion-change volume, respectively. Official patient partitions are retained, and all 20 Ljubljana patients are held out. Appendix D details dataset roles, available inputs, and training partitions.

Clinical comparators are clinical-plus-MRI regression, GRU-D, neural CDE, Transformer Hawkes, aggregated Hawkes, and shared-kernel latent Hawkes. They receive the same available clinical and imaging information; event models share the tissue-feature head and image front end. We report AUROC, AUPRC, and Brier score (Gneiting & Raftery, 2007). Confidence intervals use 1,200 patient-bootstrap samples (Efron, 1992), stratified by outcome for clinical performance; paired comparisons resample the same patients. Regional queries remain clustered within patients. Implementation, input parity, and calibration are specified in Appendices C and F.

## 6.2 PIRA prediction at an independent center

Ember achieves AUROC 0.835, AUPRC 0.550, and Brier score 0.116 at center B (Table 1). The AUROC difference from shared-kernel latent Hawkes is 6.4 percentage points (paired 95% CI [0.7, 13.4]). Mean predicted risk is 20.2% against an observed outcome frequency of 18.8%, and the calibration slope is 1.07. These results assess both risk ranking and probability accuracy under a change of clinical center. Figure 9 and Table 10 provide the ROC and calibration diagnostics.

Among test participants with no newly detected focal lesion during the final pre-landmark year, AUROC is 0.726 versus 0.556 for clinical-plus-MRI prediction. Earlier lesion history and tissue evolution remain available in this group, and the subgroup carries 5 confirmed outcomes. Figure 8 reports the full subgroup estimates, event counts, and intervals, and Appendix E defines the subgroup.

## 6.3 Predicting unobserved imaging information

To test whether the possible histories explain withheld observations, we hide the interior scan closest to the midpoint of each test patient's look-back, retaining both boundary scans. Registration, lesion matching, and features are recomputed from the retained images. The hidden scan supplies scoring

Table 1: **PIRA prediction at the independent test center.** Evaluation includes 80 patients and 15 outcomes, with 12-month confirmation. AUROC intervals are patient-bootstrap 95% confidence intervals.

| Model | AUROC (95% CI) | AUPRC | Brier ↓ |
|---|---|---|---|
| Clinical + MRI | 0.704 (0.565–0.827) | 0.320 | 0.143 |
| GRU-D | 0.729 (0.614–0.837) | 0.384 | 0.142 |
| Neural CDE | 0.748 (0.607–0.870) | 0.440 | 0.135 |
| Transformer Hawkes | 0.739 (0.620–0.853) | 0.393 | 0.139 |
| Aggregated Hawkes | 0.759 (0.649–0.863) | 0.413 | 0.136 |
| Shared-kernel latent | 0.771 (0.643–0.883) | 0.456 | 0.132 |
| EMBER | **0.835 (0.733–0.920)** | 0.550 | 0.116 |

Table 2: **Prediction of unobserved imaging.** (a) MSSEG-2 hidden-region prediction, 60 patients; NLL intervals are patient-bootstrap 95% intervals and Brier scores target regional lesion presence. (b) One hidden visit per center-B patient, $n = 80$. Count NLL averages 12 regions per patient; CRPS, 90% coverage, and width refer to total lesion volume. Lower NLL, MAE, Brier, and CRPS are better.

**(a) Public masked-region prediction**

| Model | Count NLL (95% CI) ↓ | Count MAE ↓ | Brier ↓ |
|---|---|---|---|
| Regional Poisson | 0.501 (0.380–0.634) | 0.301 | 0.113 |
| Aggregated Hawkes | 0.469 (0.354–0.592) | 0.295 | 0.107 |
| Shared-kernel latent | 0.457 (0.343–0.578) | 0.298 | 0.105 |
| EMBER | 0.451 (0.334–0.574) | 0.286 | 0.104 |

**(b) Hidden-visit reconstruction**

| Model | Count NLL | CRPS (mL) | Coverage | Width (mL) |
|---|---|---|---|---|
| Midpoint timing | 0.320 | 1.72 | 57.5% | 4.71 |
| Shared decay | 0.283 | 1.33 | 80.0% | 6.22 |
| EMBER | 0.270 | 1.01 | 83.8% | 5.93 |

references for 12 regional counts and one total lesion volume per patient. Table 2b reports count NLL 0.270 and volume CRPS 1.01 mL for EMBER, compared with 0.320 and 1.72 mL for midpoint timing. Nominal 90% volume intervals have 83.8% coverage, with width reported alongside coverage.

The first contribution is evaluated on all 960 regional queries and again on the subset where lesion timing is unresolved. In 631 of the 960 queries no incident lesion falls between the retained bracketing scans, so the hidden count is pinned; the remaining 329 queries from 80 patients carry the timing information. Their NLL is 0.788 for EMBER and 0.934 for midpoint timing, with a paired improvement of 0.146 nats (95% CI [0.059, 0.234]; Figure 3a,b). The gain therefore persists where lesion timing is unresolved.

Public MSSEG-2 provides a complementary regional test: predict one hidden region's new-lesion count from baseline burden and the other regions' counts. Models are fitted on its 40 development patients and evaluated on 60 test patients, each contributing 12 dependent queries. EMBER attains NLL 0.451 versus 0.501 for regional Poisson (Table 2a). The anonymized scan pairs use relative intervals. Appendix D.3 gives the conditional task and complete scoring protocol.

The separate Ljubljana forecast receives only baseline MRI and the requested follow-up interval. Its MAE is 2.48 mL versus 3.42 for aggregated Hawkes, with CRPS 1.97 mL. This evaluates transfer of an MRI-only temporal model to future change volume; Table 8 and Figure 7 report the complete experiment. Spatial attribution and region-wise gains are examined in Appendix G.

### 6.4 RISK UPDATES UNDER SCAN REMOVAL

Random scan removal tests the second contribution across acquisition resolutions. We delete intermediate scans with probabilities 0.25, 0.50, and 0.75, retaining the first and landmark scans, clinical covariates, and outcomes. Models infer new histories from the retained observations with fixed parameters. At removal probability 0.50, EMBER retains AUROC 0.806, compared with 0.699 for

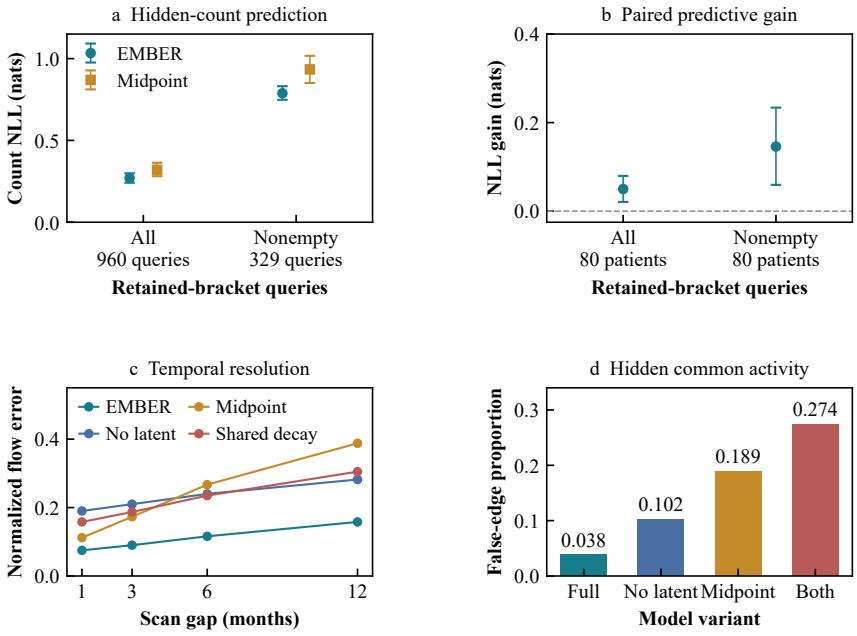



Figure 3: **Learning from unresolved lesion histories.** (a,b) Hidden-count NLL and paired gain on all queries and nonempty retained brackets, on the center-B cohort. Bars are 95% intervals from whole-patient bootstrap samples. (c,d) Mechanism controls of Appendix F.5, with branch-flow error scored against reference flows. (c) Mechanism-control error as scan gaps increase, retaining the same model variants across gaps. (d) A common driver induces false observed edges when hidden activity or timing is omitted; "Both" combines no latent activity with midpoint timing.

aggregated Hawkes and 0.687 for neural CDE (Figure 4a). The mean 90% history-risk interval width increases from .123 to .198, while the single-history predictor remains narrowly concentrated (Figure 4c).

For the fixed hidden-visit experiment, 128 compatible refinements evaluate the averaging relationship in Eq. (7). The mean absolute residual is 0.0017 with refinement learning and 0.0240 without it (Figure 4b). Deletion raises risk for 43 patients and lowers it for 37, and Appendix H traces four patients across the risk bands to show how far a single deletion can move an individual estimate. The Brier increase is 0.0285 (paired 95% CI [0.0086, 0.0512]). Figure 14 shows representative history-risk distributions, and Appendix H.5 specifies reconstruction and refinement scoring.

### 6.5 Component and Mechanism Analysis

Table 3 separates the two contributions. For the history model, midpoint timing gives AUROC 0.789, removing hidden activity gives 0.800, and sharing one decay shape gives 0.810. Omitting tissue emissions gives 0.807. Predicting from one posterior-mean history gives 0.815, supporting the role of averaging risks after the nonlinear readout.

For risk updating, removing the refinement objective gives AUROC 0.823. Forcing each coarse risk to equal its single observed fine risk reduces the gap to 0.008 but gives AUROC 0.797. The posterior-refinement target instead averages over compatible finer observations, preserving informative risk changes. These ablation gaps use eight refinements under 0.50 random removal, separately from the 128-refinement fixed-visit assessment above.

Mechanism controls of Appendix F.5, evaluated against reference branch flows, link these clinical components to the modeled observation errors (Figure 3c,d). Branch-flow error is lowest for EMBER at every scan gap and grows most slowly, from 0.075 at one month to 0.158 at twelve, whereas midpoint latent Hawkes rises from 0.112 to 0.388 and becomes the least accurate variant once the gap widens. The margin over the weakest variant therefore grows with scan sparsity, from 0.115 to 0.230,

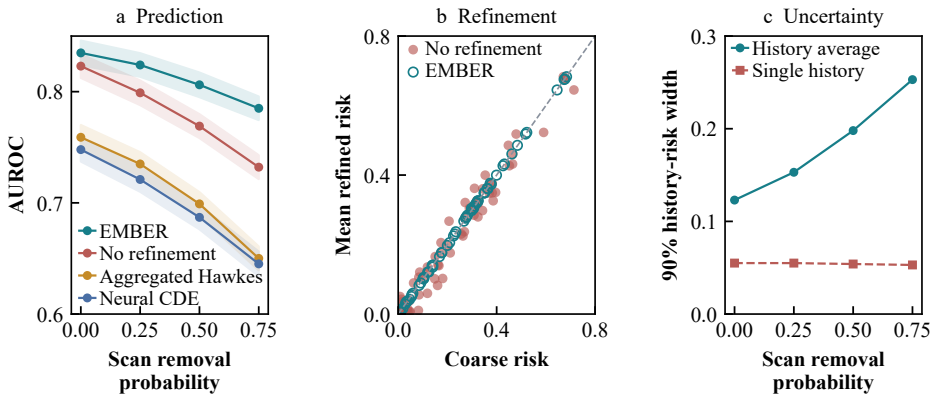



Figure 4: **Risk prediction across scan availability.** (a) AUROC under random intermediate-scan removal; bands show one standard deviation across five training seeds. (b) Coarse risk versus the posterior-weighted mean over 128 compatible refinements after one fixed hidden visit, with and without refinement learning. (c) Mean width of the 90% history-risk interval under random removal. Boundary scans and outcomes remain fixed.

Table 3: **Components supporting the two contributions.** The refinement gap is the mean absolute difference between a coarse risk and the weighted mean of compatible finer risks. The forced-identical variant illustrates why this gap is evaluated together with AUROC and Brier.

| Variant | AUROC | Brier ↓ | Refinement gap ↓ |
|---|---|---|---|
| Full EMBER | 0.835 | 0.116 | 0.012 |
| *History inference and risk averaging* | | | |
| Midpoint event times | 0.789 | 0.128 | 0.063 |
| No latent activity | 0.800 | 0.125 | 0.032 |
| One shared decay | 0.810 | 0.123 | 0.026 |
| No atrophy emissions | 0.807 | 0.124 | 0.028 |
| Posterior-mean history | 0.815 | 0.121 | 0.049 |
| *Risk updates across scan availability* | | | |
| No refinement objective | 0.823 | 0.118 | 0.039 |
| Force identical coarse/fine risks | 0.797 | 0.126 | 0.008 |

so modeling interval timing matters more as acquisitions thin out. In a pure common-driver motif, the false-edge proportion is 0.038 with the full model and 0.274 when hidden activity and interval timing are both omitted. Appendix I expands these comparisons across scan sparsity, hidden activity, and kernel dispersion; Appendix G reports regional structure and its variation between patients and centers.

## 7 Discussion and Conclusion

EMBER addresses two linked questions in progression prediction from intermittent MRI. One is how to learn regional lesion dynamics while retaining uncertainty about event timing and hidden common activity: tissue evolution constrains the possible histories, and posterior risk averaging carries them into a patient-level prediction. The other is how to keep that prediction consistent across scan availability, which posterior-refinement learning enforces. The connection is direct: the history model supplies plausible explanations of the available MRI, and the refinement objective constrains how predictions from different observation levels relate.

The experimental evidence follows this connection: hidden-region and hidden-visit prediction test the inferred histories against withheld imaging, matched scan removal and refinement ablations assess how risk responds to reduced information, and independent-center PIRA performance links these behaviors to patient-level prognosis.

## AI USE STATEMENT

AI tools were used for language polishing and manuscript review.

## REFERENCES

Thomas Brox, Olaf Ronneberger, Özgün Çiçek, Ahmed Abdulkadir, and Soeren S Lienkamp. 3d u-net: learning dense volumetric segmentation from sparse annotation. *Lecture Notes in Computer Science*, 9901:424–432, 2016.

Alessandro Cagol, Sabine Schaedelin, Muhamed Barakovic, Pascal Benkert, Ramona-Alexandra Todea, Reza Rahmanzadeh, Riccardo Galbusera, Po-Jui Lu, Matthias Weigel, Lester Melie-Garcia, et al. Association of brain atrophy with disease progression independent of relapse activity in patients with relapsing multiple sclerosis. *JAMA neurology*, 79(7):682–692, 2022.

Zhengping Che, Sanjay Purushotham, Kyunghyun Cho, David Sontag, and Yan Liu. Recurrent neural networks for multivariate time series with missing values. *Scientific reports*, 8(1):6085, 2018.

Arthur P Dempster, Nan M Laird, and Donald B Rubin. Maximum likelihood from incomplete data via the em algorithm. *Journal of the royal statistical society: series B (methodological)*, 39(1): 1–22, 1977.

Rick Durrett. *Probability: theory and examples*, volume 49. Cambridge university press, 2019.

Bradley Efron. Bootstrap methods: another look at the jackknife. In *Breakthroughs in statistics: Methodology and distribution*, pp. 569–593. Springer, 1992.

Arman Eshaghi, Alexandra L Young, Peter A Wijeratne, Ferran Prados, Douglas L Arnold, Sridar Narayanan, Charles RG Guttmann, Frederik Barkhof, Daniel C Alexander, Alan J Thompson, et al. Identifying multiple sclerosis subtypes using unsupervised machine learning and mri data. *Nature communications*, 12(1):2078, 2021.

Tilmann Gneiting and Adrian E Raftery. Strictly proper scoring rules, prediction, and estimation. *Journal of the American statistical Association*, 102(477):359–378, 2007.

Peter J Green. Reversible jump markov chain monte carlo computation and bayesian model determination. *Biometrika*, 82(4):711–732, 1995.

Francesco Guarnera, Alessia Rondinella, Elena Crispino, Giulia Russo, Clara Di Lorenzo, Davide Maimone, Francesco Pappalardo, and Sebastiano Battiato. Mslesseg: baseline and benchmarking of a new multiple sclerosis lesion segmentation dataset. *Scientific Data*, 12(1):920, 2025.

Alan G Hawkes. Spectra of some self-exciting and mutually exciting point processes. *Biometrika*, 58 (1):83–90, 1971.

Alan G Hawkes and David Oakes. A cluster process representation of a self-exciting process. *Journal of applied probability*, 11(3):493–503, 1974.

Sebastian Hitziger, Wen Xin Ling, Thomas Fritz, Tiziano D'Albis, Andreas Lemke, and Joana Grilo. Triplanar u-net with lesion-wise voting for the segmentation of new lesions on longitudinal mri studies. *Frontiers in Neuroscience*, 16:964250, 2022.

Biwei Huang, Charles Jia Han Low, Feng Xie, Clark Glymour, and Kun Zhang. Latent hierarchical causal structure discovery with rank constraints. *Advances in neural information processing systems*, 35:5549–5561, 2022.

Songyao Jin and Biwei Huang. Causal structure learning in hawkes processes with complex latent confounder networks. In *International Conference on Learning Representations*, volume 2026, pp. 58550–58582, 2026.

Ludwig Kappos, Jerry S Wolinsky, Gavin Giovannoni, Douglas L Arnold, Qing Wang, Corrado Bernasconi, Fabian Model, Harold Koendgen, Marianna Manfrini, Shibeshih Belachew, et al. Contribution of relapse-independent progression vs relapse-associated worsening to overall confirmed disability accumulation in typical relapsing multiple sclerosis in a pooled analysis of 2 randomized clinical trials. *JAMA neurology*, 77(9):1132–1140, 2020.

Patrick Kidger, James Morrill, James Foster, and Terry Lyons. Neural controlled differential equations for irregular time series. *Advances in neural information processing systems*, 33:6696–6707, 2020.

Diederik P Kingma and Jimmy Ba. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*, 2014.

Žiga Lesjak, Franjo Pernuš, Boštjan Likar, and Žiga Špiclin. Validation of white-matter lesion change detection methods on a novel publicly available mri image database. *Neuroinformatics*, 14(4): 403–420, 2016.

Arthur Masson, Benoit Combès, Anne Kerbrat, François-Daniel Ardellier, Sylvain Rabaste, Pauline Cloarec, Quentin Wdowik, Gilles Edan, Frédéric Cervenansky, Romain Casey, et al. Performances of experts and automated methods on new multiple sclerosis lesions detection: insights from the msseg2 challenge. *Scientific Reports*, 2026.

Jannis Müller, Sifat Sharmin, Johannes Lorscheider, Serkan Ozakbas, Rana Karabudak, Dana Horakova, Bianca Weinstock-Guttman, Vahid Shaygannejad, Masoud Etemadifar, Raed Alroughani, et al. Standardized definition of progression independent of relapse activity (pira) in relapsing-remitting multiple sclerosis. *JAMA neurology*, 82(6):614–625, 2025.

Marian-Andrei Rizoiu, Alexander Soen, Shidi Li, Pio Calderon, Leanne J Dong, Aditya Krishna Menon, and Lexing Xie. Interval-censored hawkes processes. *Journal of Machine Learning Research*, 23(338):1–84, 2022.

Yulia Rubanova, Ricky TQ Chen, and David K Duvenaud. Latent ordinary differential equations for irregularly-sampled time series. *Advances in neural information processing systems*, 32, 2019.

Leigh Shlomovich, Edward AK Cohen, and Niall Adams. A parameter estimation method for multivariate aggregated hawkes processes. *arXiv preprint arXiv:2108.12357*, 2021.

Leigh Shlomovich, Edward AK Cohen, Niall Adams, and Lekha Patel. Parameter estimation of binned hawkes processes. *Journal of Computational and Graphical Statistics*, 31(4):990–1000, 2022.

Alejandro Veen and Frederic P Schoenberg. Estimation of space–time branching process models in seismology using an em–type algorithm. *Journal of the American Statistical Association*, 103 (482):614–624, 2008.

Mike P Wattjes, Olga Ciccarelli, Daniel S Reich, Brenda Banwell, Nicola De Stefano, Christian Enzinger, Franz Fazekas, Massimo Filippi, Jette Frederiksen, Claudio Gasperini, et al. 2021 magnims–cmsc–naims consensus recommendations on the use of mri in patients with multiple sclerosis. *The Lancet Neurology*, 20(8):653–670, 2021.

Mingrui Xia, Jinhui Wang, and Yong He. Brainnet viewer: a network visualization tool for human brain connectomics. *PloS one*, 8(7):e68910, 2013.

Lingxiao Zhou and Georgia Papadogeorgou. Bayesian inference for aggregated hawkes processes. *Bayesian analysis*, 1(1):1–29, 2025.

Simiao Zuo, Haoming Jiang, Zichong Li, Tuo Zhao, and Hongyuan Zha. Transformer hawkes process. In *International conference on machine learning*, pp. 11692–11702. PMLR, 2020.

# A NOTATION FOR THE OBSERVATION MODEL

| Symbol | Meaning |
|---|---|
| $v_{pk}$, $T$ | Patient MRI visits and final pre-outcome landmark |
| $I_e = (a_e, b_e]$ | Last-negative to first-positive interval for incident lesion $e$ |
| $N_i$, $L_r$ | Regional incident-lesion and hidden activity processes |
| $Z$, $H$ | Full event process and complete latent history |
| $C^{(\ell)}$, $\beta_\ell$ | Nonnegative branching weights and decay rates |
| $S_{b\ell}(t)$ | Exponentially decaying event-history trace |
| $X(v_k)$, $c(t)$ | Scan-date imaging features and available clinical covariates |
| $D_f$, $D_c$ | Fine and coarsened observations of the same patient history |
| $M(H)$ | Regional and latent branching-memory features |
| $R(H)$, $r(D)$ | History-specific risk and posterior-averaged risk |

Table 4: Principal notation. Clinical and Ljubljana time is in months; MSSEG-2 reconstruction uses a unit scan interval.

The event model represents the occurrence of new MRI-visible lesions. A nonzero observed block coefficient describes a conditional excitation relationship after accounting for the modeled latent activity and measured covariates, and the prognostic features are computed as expectations over the posterior distribution of histories, carrying unresolved timing and latent activity into the risk readout.

Persistent lesion identity provides an advantage over anonymous count data: the spatial mark associates an incident lesion with a specific birth interval and subsequent volume trajectory. The spatial compartment is fixed at first detection and is not selected using the outcome. Baseline lesions have unknown pre-window ages and are represented by initial traces and baseline burden rather than entering the incident event list.

The regional memory for kernel basis $\ell$ is explicitly

$$M_{ij\ell}(H) = \sum_{e:i_e=i} \sum_{f:i_f=j,\ t_f<t_e} \frac{C_{ij}^{(\ell)} \kappa_\ell(t_e - t_f)}{\lambda_i(t_e)}.$$

Summing over $\ell$ gives the expected number of observed-parent assignments from compartment $j$ to $i$ for that history. Analogous pooled terms collect hidden-parent and background responsibilities. The risk head receives these flows divided by $1+$ the number of incident lesions, together with the count itself, terminal traces, and retained boundary covariates. This normalization keeps event burden and flow composition separately available.

The imaging emission state includes incident-lesion traces, latent traces, baseline regional lesion burden, and a patient-specific slow tissue-change component. Slow components have a Gaussian prior conditioned on baseline clinical variables. Their posterior is informed by the observed longitudinal volumes. This allows a history with no incident lesions to remain informative: a zero event count is an observed constraint on the intensity, while tissue evolution constrains a different part of the disease state.

# B ANALYTICAL RESULTS

## B.1 THE EXACT INTERVAL-OVERLAP KERNEL

**Proposition 2** (Interval kernel)**.** *Let $U \sim \mathrm{Unif}[a, b]$ and $V \sim \mathrm{Unif}[c, d]$ independently, with positive interval widths and $\beta > 0$. Then*

$$\mathbb{E}\left[\beta e^{-\beta(V-U)} \mathbf{1}(V > U)\right] = \frac{F_\beta(d-a) - F_\beta(d-b) - F_\beta(c-a) + F_\beta(c-b)}{(b-a)(d-c)},$$

*where $F_\beta(x) = x_+ - (1 - e^{-\beta x_+})/\beta$.*

*Proof.* For $x > 0$, $F'_\beta(x) = 1 - e^{-\beta x}$ and $F''_\beta(x) = \beta e^{-\beta x}$. For $x < 0$ both derivatives vanish; the first derivative is continuous at zero. Hence $F''_\beta(x) = \kappa_\beta(x)$ almost everywhere. Integrating first in $v$

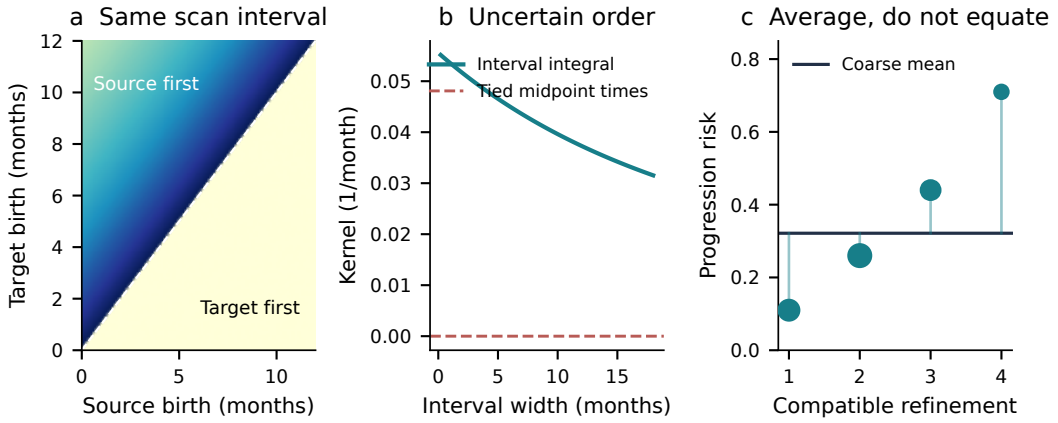



Figure 5: **Two distinctions induced by intermittent observation.** (a) Two births in the same 12-month interval admit both orders. (b) The exact uniform-proposal kernel integral remains positive although tied midpoint times yield zero. (c) A coarse risk is a weighted average over possible refinements; each realized refinement can legitimately have a different risk.

and then in $u$ gives

$$\int_a^b \int_c^d \kappa_\beta(v-u)\,dv\,du = \int_a^b \{F'_\beta(d-u) - F'_\beta(c-u)\}\,du$$
$$= F_\beta(d-a) - F_\beta(d-b) - F_\beta(c-a) + F_\beta(c-b).$$

Dividing by the product of interval widths proves the result. For identical intervals, $F_\beta(0) = F_\beta(-\Delta) = 0$, so the numerator is $F_\beta(\Delta)$. For $c \geq b$, the numerator simplifies to

$$\frac{(e^{\beta b} - e^{\beta a})(e^{-\beta c} - e^{-\beta d})}{\beta}.$$

The general expression avoids case-by-case handling of overlapping intervals. Numerically we evaluate $F_\beta(x) = x_+ + \mathrm{expm1}(-\beta x_+)/\beta$ to reduce cancellation. □

The uniform distribution in this proposition is a proposal construction. Equation (2) is evaluated with weighted complete histories, which preserves the dependence structure of $U, V$ under the Hawkes model and avoids substituting $\mathbb{E}[\kappa(V-U)]$ into each log-intensity term.

### B.2 Stability with latent feedback

**Proposition 3** (Finite event burden under column contraction)**.** *Suppose all kernels are nonnegative and integrate to one, all background rates are predictably bounded by a finite constant vector, the initial history has finite expected residual excitation, and $\|\bar{C}\|_1 \leq \rho < 1$. The joint observed-latent process is nonexplosive and has finite expected event counts on any finite observation window. With constant exogenous background and a stationary initialization, it admits the usual stationary subcritical Hawkes construction.*

*Proof.* An event of type $b$ has expected first-generation offspring $\sum_a \bar{C}_{ab} \leq \rho$. Expected total offspring in generation $k$ is bounded by $\rho^k$, including observed and latent types. The expected cluster size is at most $\sum_{k\geq 0} \rho^k = (1-\rho)^{-1}$. Bounded immigration produces a finite expected number of initiating events on a finite window, and the initial-history contribution has finite expectation by assumption. The resulting cluster process therefore has finite expected total size; an infinite number of events on a finite window would contradict this bound. With predictable varying backgrounds, domination by the bounded-background process gives the same conclusion. The constant-background stationary case follows from the subcritical cluster construction (Hawkes & Oakes, 1974). □

This argument applies to the entire block matrix. Constraining only the observed-observed block would be insufficient when a hidden feedback loop is supercritical. The implementation rescales the full nonnegative kernel sum when its maximum column sum exceeds 0.85.

### B.3 Proof of Proposition 1

Let $O_f(H)$ map a complete history to fine observations and let $O_c(H) = C\{O_f(H)\}$. For any measurable coarse-observation set $A$,

$$\begin{aligned}\Pr_\theta\{D_c \in A\} &= \int \mathbf{1}\{C(O_f(H)) \in A\}\, dP_\theta(H) \\ &= \int \mathbf{1}\{C(d_f) \in A\}\, dP_\theta^f(d_f).\end{aligned}$$

The second line is the pushforward definition of the fine-observation distribution. Expressing it in densities yields the proposition. Observations with continuous imaging features require integration over the omitted features; discrete count refinements require summation. Both operations are justified by nonnegativity. The argument does not require common decay functions, equal latent path lengths, or a DAG at the process level.

For MRI, deleting an internal visit combines the adjacent birth intervals and removes that visit's volume measurements. The equality relates the two acquisition designs through the same underlying law, which is the setting evaluated throughout this work.

### B.4 Proof of Theorem 1

The tower property gives $\mathbb{E}[r_f \mid \mathcal{F}_c] = \mathbb{E}[\mathbb{E}(Y \mid \mathcal{F}_f) \mid \mathcal{F}_c] = r_c$. For the Brier identity, decompose

$$Y - r_c = (Y - r_f) + (r_f - r_c).$$

The cross term has zero expectation because $r_f - r_c$ is $\mathcal{F}_f$-measurable and $\mathbb{E}[Y - r_f \mid \mathcal{F}_f] = 0$. Taking squared expectations establishes Eq. (10). Thus fine observations improve expected squared-error risk by exactly the variability of their conditional-mean update.

For the history-risk statement, the joint model specifies $\Pr(Y = 1 \mid H) = R(H)$, with the retained landmark covariates included in $H$. Observations do not further change $Y$ conditional on $H$, so $r_j = \mathbb{E}[R \mid \mathcal{F}_j]$. The conditional law of total variance gives

$$\mathrm{Var}(R \mid \mathcal{F}_c) = \mathbb{E}[\mathrm{Var}(R \mid \mathcal{F}_f) \mid \mathcal{F}_c] + \mathrm{Var}(\mathbb{E}[R \mid \mathcal{F}_f] \mid \mathcal{F}_c),$$

which is Eq. (11). The statement applies to the distribution over all compatible refinements.

### B.5 An approximation bound for practical prediction

**Proposition 4** (Posterior approximation and refinement gap)**.** *Let $0 \le R \le 1$. At each observation resolution, suppose a history approximation $q_j$ is within total variation $\epsilon_j$ of the exact posterior $p_j$. Then*

$$|\mathbb{E}_{q_j} R - r_j| \le \epsilon_j.$$

*If the distribution used for compatible refinements is within total variation $\delta$ of $p(D_f \mid D_c)$, the population refinement gap is bounded by*

$$\epsilon_c + \mathbb{E}_{\widetilde{p}(D_f \mid D_c)}[\epsilon_f] + \delta.$$

*Proof.* For a function bounded in $[0, 1]$, the difference between its expectations is at most total variation, using the convention $\mathrm{TV}(P, Q) = \sup_A |P(A) - Q(A)|$. Apply this to $R$ at each resolution. Insert and subtract the exact coarse risk and the average exact fine risk. The latter equals the exact coarse risk by Theorem 1. The triangle inequality gives the three error terms, yielding a deterministic approximation bound on the population risk difference. □

### B.6 Executed numerical checks

The accompanying program compares Eq. (4) with adaptive quadrature for 80 overlapping and non-overlapping interval pairs. The maximum absolute discrepancy is $4.02 \times 10^{-16}$. An enumerated finite latent-activity model verifies the tower identity to $2.22 \times 10^{-16}$ and the variance identity to $1.67 \times 10^{-16}$. Its fine and coarse Brier risks are 0.1674077 and 0.1688902; their difference, 0.00148249, equals the expected squared refinement update to numerical precision. These calculations verify the analytical identities and the numerical implementation of the estimator.

# C Inference and optimization details

## C.1 Complete-path likelihood

For an ordered full event history with labels $z_e \in \{1, \dots, d+q\}$ and event times $t_e$, the event contribution is

$$\ell_{\text{event}}(H;\theta) = \sum_e \log \lambda_{z_e}(t_e) - \sum_{a=1}^{d+q} \int_0^T \lambda_a(t)\, dt. \tag{12}$$

For an exponential term arising from an event at $u$, the compensator is $C_{ab}^{(\ell)}\{1 - e^{-\beta_\ell(T-u)}\}$. Background covariates are held at their last documented values between update times, so their contribution is integrated piecewise. Initial traces have the corresponding analytic exponential integral, so every compensator is evaluated in closed form in continuous time.

For a nonnegative volume, let $U \sim \mathcal{N}(m, \sigma^2)$ and $V = \exp\{\max(0, U)\} - 1$. The observation law has mass $\Phi(-m/\sigma)$ at zero and the transformed Gaussian density for $V > 0$, including its Jacobian. Signed normalized tissue changes retain Gaussian emissions.

The full density adds $\sum_k \log p_\theta\{X(v_k) \mid S(v_k), X(0), c(v_k)\}$ and a prior on initial traces and slow tissue-change effects. Posterior evaluation uses only the pre-landmark history. Outcome labels enter the risk loss after event-model fitting; they are not used to refine the history of a held-out patient.

## C.2 Admissible path proposals

Observed lesion identities are fixed. A proposed birth time must lie strictly after its last negative scan and no later than its first positive scan. The sampler alternates interval-truncated time moves with birth, death, and time-shift moves for latent events (Green, 1995). Initial trace values are updated on the logarithmic scale. Each proposal carries its forward and reverse density; for latent birth/death moves these include the event-count probability, component choice, and time-density factors. Sorting event times is accompanied by the corresponding ordered-event density. Moves outside an observed interval have zero target probability.

Proposals use the closed-form overlap feature to favor temporally compatible parent assignments, and a mixture with a uniform interval proposal preserves support. At least 20% of time proposals use the uniform component. Hidden-event proposals are drawn from a Poisson reference process with a conservative rate; importance weights correct to the full Hawkes path density. Equal treatment of midpoint and interval proposals without this correction would change the target model.

The amortized proposal uses a 16-dimensional bidirectional GRU over pre-landmark visit features, together with interval widths and the overlap features. A two-layer network outputs positive beta-distribution parameters for each lesion's normalized interval position. A reference-rate head proposes hidden event counts and component labels; its density includes the Poisson count mass and conditional time density. The encoder is fitted to posterior samples by inclusive cross-entropy and is subsequently updated with the refinement objective. Importance correction and local Metropolis rejuvenation retain the generative target. Coarse observations are encoded afresh after masking; a hidden encoder state computed from a removed scan is never reused.

## C.3 Training objectives and computation

The imaging-event objective is

$$-\sum_p \log p_\theta(D_p) + \lambda_A \sum_\ell \|C_{OO}^{(\ell)}\|_1 + \lambda_L \sum_{r=1}^{q} \left\| \{C_{r,:}^{(\ell)}, C_{:,r}^{(\ell)}\}_{\ell=1}^{J} \right\|_2 .$$

In the last term $r$ indexes a latent row/column in the full matrix. Duplicated entries are counted once. The risk objective is

$$-\sum_p \{Y_p \log \widehat{r}(D_p) + (1 - Y_p) \log[1 - \widehat{r}(D_p)]\} + \lambda_{\text{ref}} \mathcal{L}_{\text{ref}} + \lambda_\psi \|\psi\|_2^2.$$

Group shrinkage and dimension selection regularize hidden common activity, with the same hidden-state complexity budget available to the common-kernel latent comparator.

**Algorithm 1** EMBER fitting and prediction

**Require:** Development histories $D_p$, outcomes $Y_p$, latent candidates $q = 0{:}2$

**for** each candidate dimension $q$ **do**
  Initialize nonnegative kernel blocks, baseline rates, emissions, and initial traces.
  **for** each generalized EM iteration **do**
    Draw weighted, interval-compatible histories $H_{ps}$ from the current path posterior.
    Update parameters using weighted Eq. (12), emissions, and group penalties.
    Project the full integrated kernel matrix to maximum column sum at most $0.85$.
  **end for**
  Fit the risk head with binary cross-entropy of Eq. (6).
  Mask internal scans and draw compatible posterior refinements.
  Update the amortized predictor with Eq. (8) and outcome supervision.
**end for**
Select dimension and penalties using development validation only.
Freeze model parameters and risk calibration.
At test time, infer histories without outcomes and average history-specific risks.

Maintaining exponential traces avoids summing over every preceding event for each intensity evaluation. For a path with $E$ events and $K$ scans, dense computation costs $O\{J(d+q)^2(E+K)\}$; sparse block multiplication reduces this to the number of active coefficients, which stays small because intervals are long and most patients have few incident lesions. Posterior computation multiplies this by the number of retained paths and local updates. The risk head has two hidden layers of widths 16 and 8, with dropout 0.2.

| Item | Configuration |
|---|---|
| Time unit and kernels | Months; normalized exponential bases with time constants 3, 9, 24 |
| Anatomical dimension | 12 non-overlapping compartments, with hemispheric labels |
| Latent candidates | $q = 0, 1, 2$; selected by validation imaging likelihood |
| Stability | Maximum column sum of full integrated matrix $\leq 0.85$ |
| Histories | 64 for training, 128 for evaluation; effective sample size monitored |
| Refinement samples | 8 compatible finer observations per coarse history |
| Penalties | Fixed $\lambda_A = 10^{-3}$ and $\lambda_L = 10^{-2}$ |
| Refinement weight | $\lambda_{\text{ref}} \in \{0, 0.1, 0.3, 1.0\}$ |
| Risk readout | Two layers, 16/8 units; dropout 0.2; weight decay $10^{-2}$ |
| Training controls | Early stopping on validation Brier score; five random initializations |
| Imaging features | Log lesion burden, matched-lesion growth, parenchymal fraction, bilateral thalamic volumes |
| Clinical features | Age, sex, disease duration, landmark EDSS, pre-landmark relapse and treatment history |

Table 5: Model configuration and development-only selection grid.

# D Open-data evaluation

## D.1 Dataset roles and availability

The public component comprises 195 patients and 355 MRI examinations. These totals distinguish patients from repeat visits: MSLesSeg contributes 75 patients and 115 examinations, MSSEG-2 contributes 100 patients and 200 examinations, and Ljubljana contributes 20 patients and 40 examinations. Development and evaluation follow the official patient partitions where provided; all examinations of one participant remain together. The public cohorts are collected in Italy, France, and Slovenia, respectively, and are evaluated separately.

**MSLesSeg.** The official Figshare archive (`https://doi.org/10.6084/m9.figshare.27919209`) includes raw images, preprocessed T1/T2/FLAIR images, lesion masks, clinical metadata, and transformation files under CC-BY 4.0 (Guarnera et al., 2025). The 75 patients have one

| Source | Patients | MRI visits | Development | Test |
|---|---|---|---|---|
| MSLesSeg | 75 | 115 | 53 | 22 |
| MSSEG-2 | 100 | 200 | 40 | 60 |
| Ljubljana longitudinal | 20 | 40 | 0 | 20 |
| Public total | 195 | 355 | 93 | 102 |
| Clinical center A | 160 | 545 | 160 | 0 |
| Clinical center B | 80 | 268 | 0 | 80 |

Table 6: Patient-level data organization. Clinical development contains 128 training and 32 validation participants. Public test tasks retain their own endpoints and are not pooled into one accuracy estimate.

to four visits: 50 have one, 15 have two, five have three, and five have four. All 25 patients with repeated examinations belong to the 53-patient development partition, which contains 93 scans. The 22 test patients have one examination each. Thus the development partition pairs 53 independent patients with 93 scans, so repeated examinations enrich the image-training set beyond the patient count. This cohort validates all-lesion segmentation and burden measurement.

**MSSEG-2.** The challenge contains 100 patients with paired 3D FLAIR images and expert new-lesion masks (Masson et al., 2026). Its official development/test partition is 40/60; the test set has 32 patients with new lesions and 28 without them. Images are shared through `https://shanoir.irisa.fr/shanoir-ng/welcome` under the MSSeg2 study and its data-use agreement. The post-challenge release of test images and expert masks is documented by Hitziger et al. (2022). The two images are registered to an intermediate space, and the release is used for delineation of interval-visible new lesions from the image pair. The original study reports a mean interval of 395 days and a range of 70–1187 days, which sets the interval scale for this task.

**Ljubljana longitudinal.** This CC-BY release has 20 patients, each with two examinations comprising two-dimensional T1, T2 and FLAIR images, brain masks, and expert reference change masks (Lesjak et al., 2016). Its patient-specific intervals are 81–723 days. T1/FLAIR sampling is approximately $0.9 \times 0.9 \times 3$ mm. The raw images and elapsed days are distributed through `https://github.com/muschellij2/open_ms_data`. Baseline lesion features and elapsed time support a forecast of subsequent annotated change volume. Clinical variables and tissue-volumetric features are masked for this task.

### D.2 Shared image front end

Two compact 3D U-Net modules (Brox et al., 2016) extract all-lesion masks and interval new-lesion candidates. The first uses T1 and FLAIR, fitted on the 53-patient MSLesSeg development partition; the second uses the two registered FLAIR images, fitted on the 40-patient MSSEG-2 development partition. A FLAIR-only branch is trained with modality dropout for acquisition transfer. Training samples $96^3$-voxel patches with four resolution levels and 16 initial channels, a combined Dice/cross-entropy objective, and Adam (Kingma & Ba, 2014) at $10^{-3}$. The training budget is fixed at 20,000 updates. Four-fold patient-level development predictions select a probability threshold; each final network is refitted on its full development partition. All-lesion threshold selection maximizes mean Dice. New-lesion threshold selection maximizes lesion F1 among thresholds producing at most 0.5 false lesions per lesion-negative development patient. Public test images and annotations are reserved for evaluation.

MSLesSeg evaluation uses the provided test masks and averages Dice and total-volume absolute error across its 22 patients. MSSEG-2 uses the challenge lesion-overlap rules and its 3 $mm^3$ component cutoff. Dice and lesion F1 are averaged over the 32 lesion-positive test pairs; false detections are averaged over the 28 lesion-negative pairs, so the two metric families use distinct denominators. Table 7 reports the shared front end before clinical adjudication of ambiguous matches. Figure 6 shows MSLesSeg measurements alongside the regional-inference experiment below. The paired-image module proposes interval candidates, while temporal inference operates on the resulting observations. These extraction networks are fixed for all clinical temporal comparators.

| Public test task | $n$ | Metric | Shared front end |
|---|---|---|---|
| MSLesSeg: all lesions | 22 | Dice ↑ | $0.789 \pm 0.076$ |
| MSLesSeg: total burden | 22 | MAE (mL) ↓ | $1.04 \pm 0.93$ |
| MSSEG-2: new lesions present | 32 | Dice ↑ | $0.506 \pm 0.133$ |
| MSSEG-2: new lesions present | 32 | Lesion F1 ↑ | $0.535 \pm 0.130$ |
| MSSEG-2: no new lesion | 28 | False lesions ↓ | $0.25 \pm 0.70$ |

Table 7: **Public validation of the shared observation pipeline.** Values are patient-level means and standard deviations. Repeated rows describe different metrics on the same patients. These are measurement-module results shared by all temporal methods.

### D.3 Masked regional reconstruction on MSSEG-2

The new-lesion reference masks also provide a direct test of the latent event model. Connected components are assigned to the same 12 non-overlapping anatomical compartments used in clinical modeling, using baseline FLAIR-to-atlas registration and a fixed largest-overlap rule. Let $B$ denote baseline regional lesion burden and $K_i$ the reference new-lesion count in region $i$ over the observed scan pair. For query $j$, the input is $(B, K_{-j})$ and the target is $K_j$. Visible-region reference counts are supplied as observed measurements, so the hidden region's own count enters only as the scoring target for query $j$.

Each scan pair is represented on $[0, 1]$, with all incident events interval-censored in $(0, 1]$. Kernel time constants are $1/8$, $3/8$, and $1$ in these relative units; their weights and background rates are fitted anew on the 40 public development patients. This parameterization describes activity within a scan interval on the interval's own relative scale. Four-fold patient-level development likelihood selects latent dimension and regularization before refitting on all 40 patients. The 60 official test patients remain held out.

The conditional count law integrates the hidden region's events and the latent activity jointly:

$$p(K_j \mid B, K_{-j}) = \int \mathbf{1}\{N_j(1) = K_j\} p(H \mid B, K_{-j})\, dH.$$

Regularized regional Poisson regression uses baseline burden and visible-region counts; aggregated Hawkes conditions on visible regional intervals without hidden activity; the shared-kernel latent model uses one common response shape. All comparators receive the same baseline and visible-region measurements. The posterior mean is scored by absolute count error, and the predictive count distribution by negative log probability. Binary-presence Brier scores use $1 - p(K_j = 0 \mid B, K_{-j})$.

Every test patient is evaluated once for each of the 12 hidden regions. Scores are first averaged across the 12 queries and then across the 60 patients; 1,200 bootstrap draws resample these patient-level score vectors. The 720 queries are not treated as independent patients. Table 2 reports this evaluation, and Figure 6 displays paired patient-level count scores. EMBER lowers the mean NLL from 0.501 to 0.451. Its regional count MAE is 0.501 among the 32 lesion-positive patients and 0.039 among the 28 lesion-negative patients; the corresponding regional Poisson values are 0.460 and 0.118. Conditioning on quiet regions chiefly reduces predicted activity in lesion-negative patients, while counts in active patients retain greater uncertainty.

### D.4 A baseline-to-follow-up imaging forecast

For a public participant, the input consists of the first MRI and the known interval to the requested follow-up time. The baseline is processed on its own, with registration, lesion filling, segmentation, feature scaling, and state inference computed from the first MRI alone. The reference change mask supplies the scoring target after prediction: its nonzero-voxel volume is $V_\Delta$ in mL, with the original voxel dimensions retained.

In the imaging model, interval change volume is an emission linked to the baseline state, elapsed time, and the event traces accumulated over the interval. A censored Gaussian emission for $\log(1 + V_\Delta)$ preserves nonnegativity and a mass at zero while accommodating skewness. For an upcoming scan, new events, latent activity, and continuous lesion evolution are integrated out:

$$p(V_\Delta \mid X(0), \Delta) = \int p(V_\Delta \mid H, X(0), \Delta)\, p(H \mid X(0), \Delta)\, dH.$$

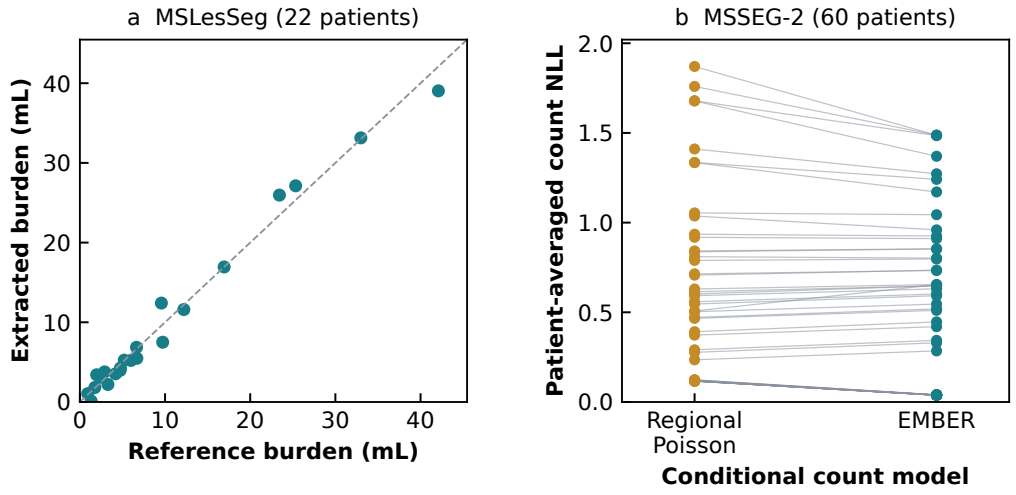



Figure 6: **Public observation tests.** (a) Total lesion burden on 22 MSLesSeg test patients. (b) Paired count NLL on the 60 MSSEG-2 test patients, averaged over 12 masked-region queries per patient. All 60 MSSEG-2 test patients contribute to the regional scores, including the 28 without new lesions.

Inference conditions on the baseline scan and the requested follow-up interval. Persistent appearance events and continuous change emissions retain their distinct meanings. A lesion can enlarge without producing another birth event.

The MRI-only version is fitted on paired scans from the 128 training participants from center A. Selection uses its 32 validation participants. All transfer models use the same baseline lesion features and interval inputs, with clinical and tissue-volumetric variables masked to match the open-data setting. The separate full-input clinical model is used for PIRA experiments in Section 6. No Ljubljana follow-up target is used for fitting, calibration or selection. Public participants are the independent evaluation units.

Table 8: **Open Ljubljana lesion-change forecasting.** Patient-level mean and standard deviation across 20 participants. Both mean absolute error (MAE) and continuous ranked probability score (CRPS) are in mL; lower is better. Each method receives only baseline MRI and the follow-up interval.

| Method | MAE (mL) ↓ | CRPS (mL) ↓ |
|---|---|---|
| Imaging regression | $4.12 \pm 4.08$ | $3.31 \pm 2.95$ |
| Neural CDE forecast | $3.61 \pm 3.43$ | $2.93 \pm 3.44$ |
| Midpoint latent Hawkes | $3.73 \pm 2.90$ | $3.00 \pm 2.60$ |
| Aggregated Hawkes | $3.42 \pm 2.29$ | $2.76 \pm 2.09$ |
| EMBER | $2.48 \pm 2.09$ | $1.97 \pm 1.57$ |

### D.5 Scoring the comparators

Let $V_\Delta^{(1)}, \ldots, V_\Delta^{(S)}$ denote predictive volume samples. The point forecast is their posterior mean. The continuous ranked probability score (Gneiting & Raftery, 2007) is estimated by

$$\mathrm{CRPS} = \frac{1}{S}\sum_s |V_\Delta^{(s)} - v_\Delta| - \frac{1}{2S^2}\sum_{s,t} |V_\Delta^{(s)} - V_\Delta^{(t)}|.$$

It evaluates both the predictive location and its spread in mL. MAE and CRPS are averaged over the 20 participants, with between-participant standard deviations.

The imaging regression baseline predicts change volume from baseline features and elapsed time. The neural CDE forecast is initialized by the same baseline features and propagated to the requested time. Midpoint latent Hawkes is fitted using midpoint times for internal training events; aggregated Hawkes integrates observed event intervals but omits hidden activity. All event models share the change-volume emission and baseline feature encoder. Thus the public comparison tests transfer of the temporal representation while holding image measurement and clinical-variable availability fixed.

The open study complements the internal multi-visit experiments and assesses distributional forecasting in a new MRI setting. Posterior-refinement learning is evaluated through the internal multi-visit experiments: the risk ablations, matched scan removal, and the exact probability identities.

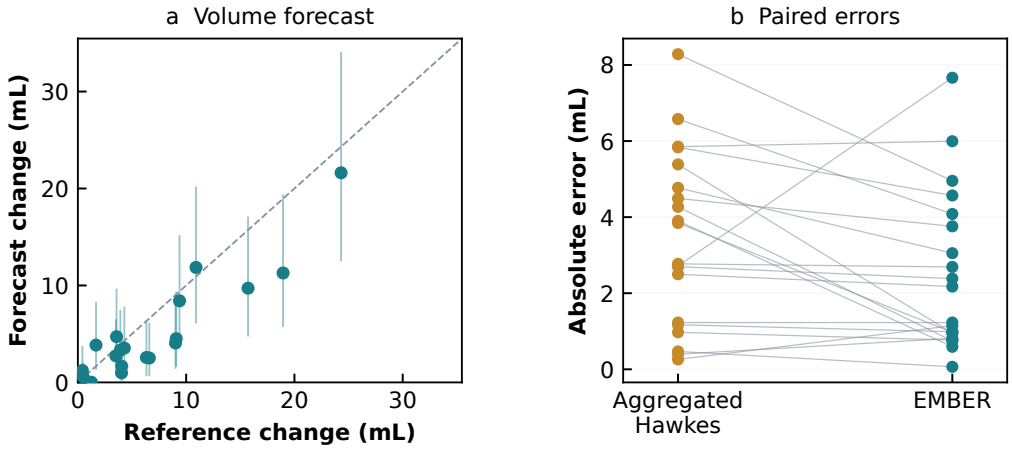



Figure 7: **Open longitudinal imaging forecast.** (a) Participant-level change volume and posterior-mean forecast; vertical bars show central 90% predictive intervals. (b) Paired absolute prediction errors, one line per participant. Follow-up change masks supply the scoring reference for the forecast.

# E Clinical cohort and endpoint adjudication

## E.1 Cohort partition and acquisition

The two-center cohort contains 240 eligible participants. Center A contributes 160 participants, partitioned by patient into 128 training and 32 validation records. Center B contributes 80 independent test participants. Eligibility is assessed before partitioning, and every examination from a participant remains in the same partition. Table 9 reports the development and test populations.

| Characteristic | Development (A) | Test (B) |
|---|---|---|
| Participants | 160 | 80 |
| Women, $n$ (%) | 106 (66.2) | 52 (65.0) |
| Age, years, median (IQR) | 40.8 (35.4–46.4) | 41.9 (35.6–48.0) |
| Disease duration, years, median (IQR) | 6.6 (4.5–10.0) | 6.9 (4.6–9.7) |
| Landmark EDSS, median (IQR) | 2.5 (2.0–3.5) | 2.5 (2.0–4.0) |
| Usable pre-landmark scans, median (IQR) | 3 (3–4) | 3 (3–4) |
| Scan interval, median (IQR), months | 10.3 (7.5–11.2) | 10.5 (7.4–11.6) |
| No new lesion detected in final year, $n$ (%) | 65 (40.6) | 32 (40.0) |
| 24-month PIRA onset, $n$ (%) | 31 (19.4) | 15 (18.8) |

Table 9: Cohort characteristics at the prediction landmark. Scan-interval summaries pool adjacent intervals within each center. PIRA onset is subsequently confirmed for 12 months.

Eligibility comprises adults aged 18–65 years with relapsing-remitting MS, landmark EDSS no greater than 5.5, at least three usable three-dimensional T1/FLAIR examination pairs within the

24-month look-back, with a first scan 21–24 months before the landmark, and sufficient clinical follow-up to classify the primary outcome. The same eligibility and imaging-quality rules apply at both centers. Scans obtained specifically for an acute relapse are excluded from the primary surveillance sequence. Pre-landmark relapses remain available as clinical covariates.

T1 images undergo lesion filling before volume estimation. Registration and segmentation outputs are inspected for motion, incomplete brain coverage, and tissue-label failures before quantitative features are accepted. Within-subject registration aligns each visit to an unbiased longitudinal template. New FLAIR lesions are matched with spatial correspondence and manual adjudication of uncertain cases. Paired change maps are obtained from registered FLAIR images with expert review of candidate changes; their nonzero-voxel volume defines the imaging-forecast target, including lesion appearance, enlargement, shrinkage, and disappearance. Large confluent lesions and unresolved split/merge events are flagged; they contribute continuous lesion burden but not uncertain additional births. The primary event list includes lesions with reliable first-appearance intervals that remain visible and matchable on subsequent scans through the landmark. This persistence filter is applied using pre-outcome imaging only. Lesions seen exclusively at an intermediate visit enter its burden measurement, not the persistent-event list. After that visit is removed, its burden is also removed.

Compartment assignment uses an ordered, non-overlapping rule within each hemisphere: infratentorial tissue, periventricular white matter, juxtacortical white matter, and the remaining deep frontal, deep parietal, and deep temporo-occipital white matter. Boundary precedence prevents a lesion from contributing multiple regional events. Lesion burden, growth, and tissue volumes are normalized by intracranial volume where appropriate, then standardized using development statistics. Scanner and protocol changes are recorded. Longitudinal tissue-volume features require compatible T1 acquisitions and acceptable registration; scanner information also enters the emission model as a nuisance covariate.

### E.2 Time origin and PIRA

Each patient's last scan in the look-back defines $T$. All predictive inputs are truncated at this time. A participant may have an onset at $T + 23$ months and a valid confirmation at $T + 35$ months; this is a positive primary outcome. A new onset at $T + 27$ months is outside the primary horizon even if confirmation is available. This distinction prevents a 24-month prediction target from being silently replaced by a later disability label.

The primary rule follows the 12-month standardized PIRA framework (Müller et al., 2025). A significant EDSS worsening is an increase of at least 1.5 from a reference EDSS of 0, at least 1.0 for reference scores from 1.0 through 5.5, or at least 0.5 above 5.5. The reference is re-established after qualifying relapses, prior PIRA, and confirmed improvement. Worsening must occur without a relapse since the preceding eligible EDSS assessment. Confirmation measurements must remain above the qualifying threshold for at least 12 months; assessments within 30 days after relapse are ineligible for confirmation. The clinical adjudication algorithm is identical across study partitions.

The fixed-horizon primary analysis uses complete adjudication. A missing confirming assessment does not become a negative label. The operational workflow first checks endpoint evaluability, then assigns the outcome. Clinical follow-up after $T$ is used exclusively for endpoint ascertainment. Treatment history is encoded only through $T$; the estimated probability reflects progression during subsequent usual care, not a counterfactual treatment regime.

### E.3 The clinically quiet subgroup

The MRI-quiet subgroup has no newly detected focal lesion at surveillance examinations in $(T - 12, T]$, with an earlier reference examination available for comparison. Membership follows scan-based detection dates and is assigned before outcomes are examined. It includes 32 test participants and 5 outcomes. The model's emission features provide a complementary route from longitudinal atrophy and chronic lesion evolution to progression risk. The subgroup analysis uses the full-cohort fitted model and does not tune a separate threshold.

# F COMPARATORS AND UNCERTAINTY

## F.1 COMPARATOR INPUT PARITY

The clinical-plus-MRI comparator uses regularized logistic regression with age, sex, disease duration, landmark EDSS, prior relapses, treatment class, baseline and landmark lesion burden, lesion-count increments, and tissue-volume slopes. GRU-D and neural CDE receive the corresponding visit-level sequence, masks, and actual acquisition times. A latent-ODE model is part of development architecture selection; neural CDE is the continuous-time comparator retained for the clinical performance table.

Transformer Hawkes receives one token per incident lesion, with midpoint time, anatomical type, and a scan-derived lesion embedding. Its pooled history is concatenated with the same clinical and tissue summary used by other models. Aggregated Hawkes integrates birth times within intervals using a multivariate event model without hidden subprocesses. The shared-kernel latent comparator uses one common decay shape, midpoint times, and the same maximum latent dimension. All event comparators have access to the landmark tissue head. These choices isolate the observation model and hidden dynamics rather than withholding useful imaging from a baseline.

The rank-informed comparator computes lagged count covariance features after placing events at midpoint times, and selects latent complexity by a singular-value threshold tuned only on validation patients. Its final risk head has the same width as EMBER, giving an operational shared-kernel baseline for this task.

## F.2 CALIBRATION AND PATIENT UNCERTAINTY

Validation selects the clinical refinement weight and stopping point; the latent dimension is selected by held-out imaging likelihood. Risk recalibration fits an intercept on the validation participants with a standard-normal prior, keeping the prediction-logit coefficient fixed at one. Test-center calibration plots use four groups containing equal numbers of patients sorted by predicted probability. The Brier score is the mean squared difference between predicted probability and the binary outcome. Calibration slope is estimated by logistic regression of the outcome on the prediction log-odds, with a free intercept. AUROC and AUPRC use the same fixed participant set.

Patient bootstrap intervals resample center B participants with replacement within outcome strata, retaining both the test-set size and event count. Subgroup intervals use the same outcome-stratified scheme within each subgroup. The all-participant row in Figure 8 uses the same bootstrap draws as Table 1. For a method difference, the same resampled indices are used for both predictions. The 90% history-risk interval is instead computed from the weighted 5th and 95th percentiles of $R(H)$, and describes sensitivity to unresolved timing and activity histories under the fitted model.

## F.3 CALIBRATION AT THE INDEPENDENT CENTER

Center B contains 80 participants and 15 PIRA outcomes (18.8%), and all models are frozen after development at center A. Table 10 reports the calibration intercept and slope that characterize the risk gradient; these coefficients are diagnostic fits and do not alter the evaluated predictions. Table 1 gives the discrimination and Brier results, including the paired comparison with the shared-kernel latent comparator.

| Model | Intercept (95% CI) | Slope (95% CI) | Mean risk (%) |
|---|---|---|---|
| Clinical + MRI | 0.40 (-0.82–2.11) | 1.30 (0.44–2.68) | 19.6 |
| Shared-kernel latent | -0.06 (-0.70–0.93) | 0.92 (0.46–1.71) | 18.3 |
| EMBER | -0.05 (-0.57–0.87) | 1.07 (0.68–1.88) | 20.2 |

Table 10: **Calibration at the independent test center.** Parentheses show 95% patient-bootstrap intervals. Intercept and slope are jointly fitted, with reference values 0 and 1. Mean risk uses the frozen model probabilities.

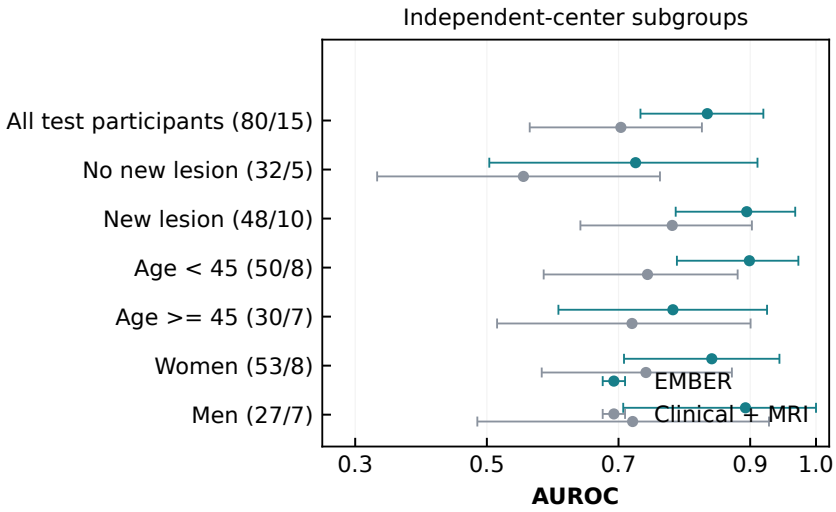



Figure 8: **Internal held-out subgroup discrimination.** Points use the same frozen predictor in center B. Horizontal bars show 95% patient-bootstrap intervals from 1,200 outcome-stratified resamples within each displayed group. Parentheses report participants and PIRA outcomes. Groups overlap and are evaluated without subgroup-specific fitting.

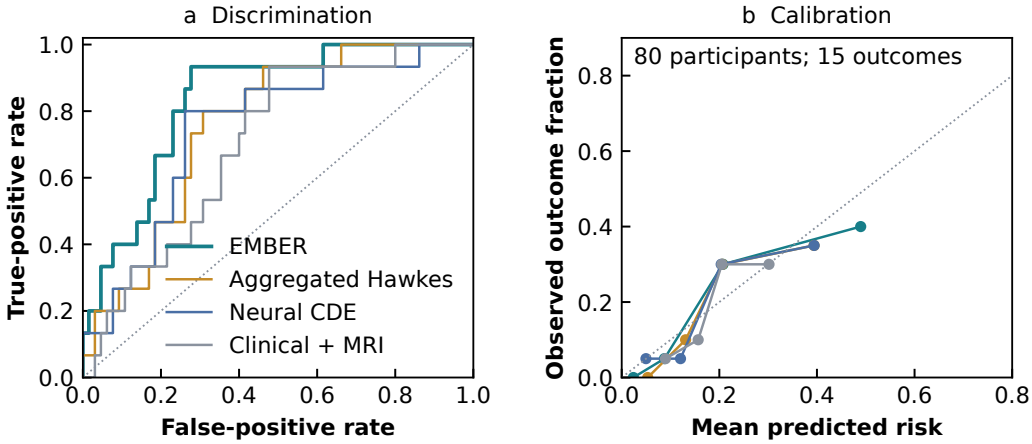



Figure 9: **PIRA performance at the independent test center.** (a) ROC curves use the same held-out participants. (b) Calibration is shown in four equal-frequency prediction groups; the dashed line denotes perfect calibration.

### F.4 Scan-coarsening protocol

For every intermediate scan, an independent removal indicator is sampled at the specified probability. First and landmark examinations remain. Registration, lesion correspondence, and tissue features are recomputed from the retained images. Birth intervals use the retained negative/positive examinations; intermediate tissue observations are removed. The clinical covariates and endpoint remain unchanged. Five training initializations and fixed acquisition masks give the seed mean and standard deviation shown in Figure 4, and the corresponding discrimination is reported alongside the center-B estimate in Table 1.

The refinement gap in Table 3 uses posterior-generated compatible fine observations, not the observed fine prediction alone. The ablation summary uses eight compatible refinements under the 0.50 random-removal protocol. The fixed one-visit reconstruction analysis uses 128 refinements, so its residual is reported separately. In the invariant ablation, training explicitly matches each coarse prediction to its single observed fine partner. The Brier and discrimination measures distinguish this behavior from the distributional averaging that the posterior-refinement objective targets.

### F.5 Controlled network motifs

The mechanism benchmark contains 12 observed compartments and up to two hidden activity processes. It includes independent regional innovations, one hidden common driver, a hidden two-stage path with unequal lengths, and a mixed network with regional excitation. The full integrated branching matrix is rescaled to spectral radius below 0.85. Kernels combine 3-, 9-, and 24-month components with edge-specific weights. Baseline rates span clinically sparse lesion activity; zero-count intervals are retained.

Training and test sets contain 512 and 256 independent histories per replicate, with ten parameter replicates. Observations use median gaps of 1, 3, 6, or 12 months with 15% scheduling jitter. Ground-truth parent responsibilities are available for this controlled benchmark. Branch-flow error is the Frobenius difference between the inferred and true expected regional responsibility matrices, divided by the norm of the true matrix with a fixed positive floor. In the pure common-driver motif, all cross-regional direct edges are absent. The false-edge statistic is the fraction of these absent ordered pairs selected using a development-fixed posterior threshold.

Figure 3c,d reports the principal mechanism comparisons. The factorial analysis in Appendix I varies the conditions jointly and reports replicate uncertainty.

## G Additional spatial results

### G.1 Spatial memory across patients and centers

To evaluate the regional dynamics learned from uncertain histories, we examine temporal response shapes, sensitivity to patient composition, and agreement at the independent center. These analyses connect the first contribution to spatial patterns beyond patient-level risk. Four prominent relationships selected using training patients have distinct short-term responses and long tails within the monotone exponential-mixture family (Figure 12). Removing latent activity redistributes responsibility toward observed regions; the matched network and threshold comparisons in Figure 13 locate this change mainly among weak relationships.

Figure 10 connects population summaries to individual variation. At a common responsibility threshold $\tau = 0.025$, the 128 training patients and 80 test patients each retain the same 12 directed relationships. Parameters and display settings are fixed after development. All 12 training relationships exceed this threshold in all 1,200 patient-bootstrap means. The threshold grid shows how their support changes as the required attribution increases. This resampling measures variation in patient composition conditional on the fitted model.

Cross-center agreement holds at both scales. The rank correlation is 0.95 for the 12 relationships selected by the training threshold and 0.94 across all 132 off-diagonal pairs, so the population ordering is preserved. Among the selected relationships, the fraction of test patients whose individual responsibility exceeds $\tau$ lies between 73.8% and 100%. The side-by-side mean intervals and patient fractions show how each selected relationship persists at the population level and how it varies between individuals.

### G.2 Where regional prediction improves

We next evaluate spatial memory against withheld MRI observations. The public masked-region task has 60 MSSEG-2 test patients; the fixed hidden-visit task in Section 6.3 has 80 independent-center patients. Each supplies 12 dependent regional queries. Figure 11 maps the mean paired NLL improvement, defined as comparator NLL minus EMBER NLL. Positive values favor EMBER. The

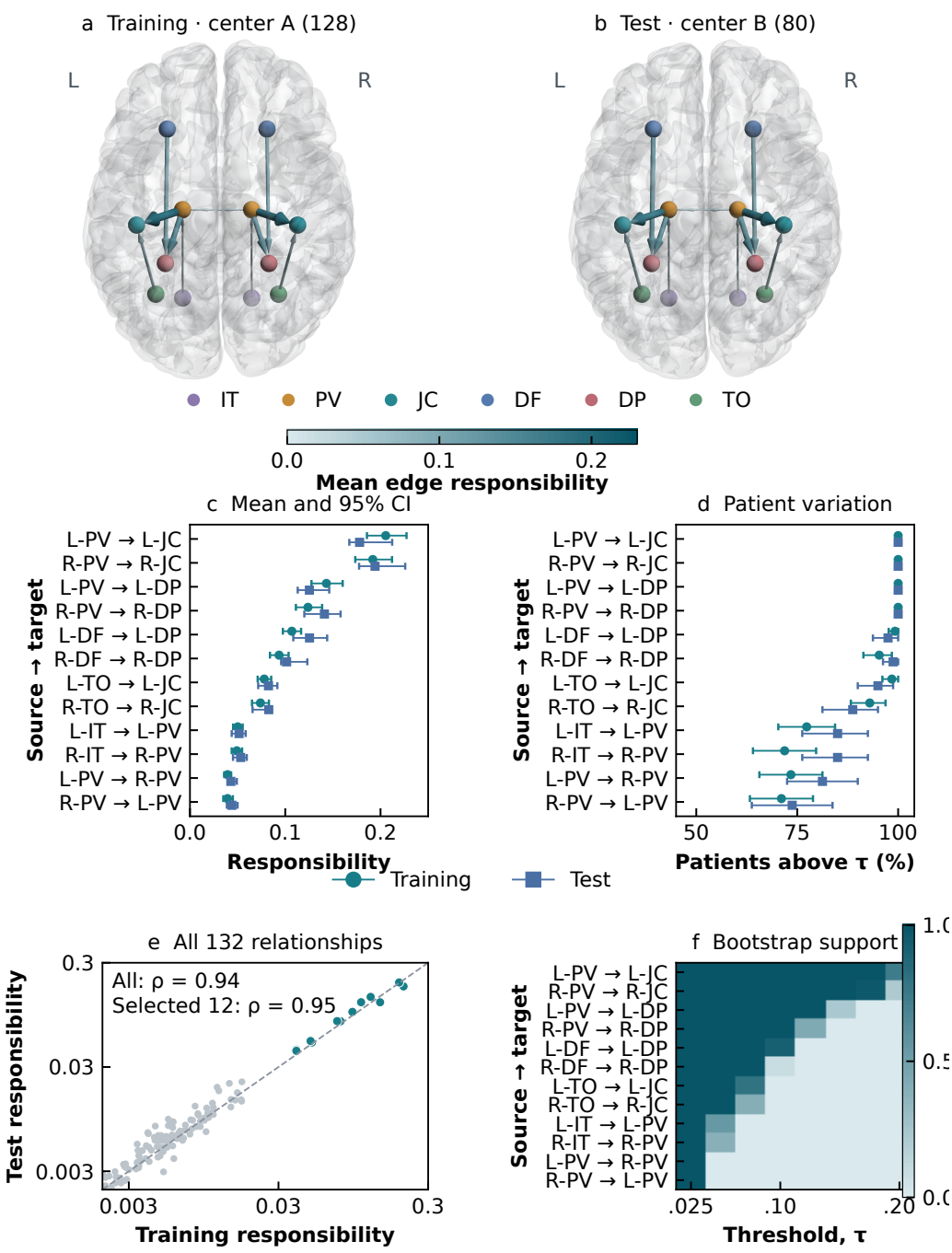



Figure 10: **Population network agreement and individual variation.** (a,b) Training and independent-test mean attributions, rendered with BrainNet Viewer on the same Ch2 template. Node colors identify compartments; arrow direction, width, color, and threshold are shared. (c) Means and 95% patient-bootstrap intervals for the 12 training-selected relationships. (d) Fractions of individual patients with responsibility at least $\tau = 0.025$, with patient-bootstrap intervals. (e) All 132 off-diagonal pairs; teal marks the training-selected 12, with logarithmic axes. Both rank correlations are reported. (f) Fraction of 1,200 training-patient bootstrap means exceeding each threshold. Row order in (c,d,f) follows training mean responsibility. These frequencies summarize population-level attribution under the fixed model.

comparator is regional Poisson for public masking and midpoint timing for hidden visits, matching the distinct observation tasks.

Mean NLL improves in 11 of 12 public compartments and 10 of 12 internal compartments. Periventricular regions show the largest gains, alongside larger numbers of positive targets or nonempty retained brackets. Showing the signed effects, patient-bootstrap intervals, and contributing counts together separates predictive benefit from regional activity prevalence. The two tasks retain their own denominators and comparators.

For hidden visits, 631 of the 960 queries have no incident lesion in the retained bracketing interval, so the retained visits determine their hidden count. We additionally score the 329 nonempty brackets, contributed by 80 of the 80 patients, using only retained observations to define this subset. NLL is 0.788 for EMBER and 0.934 for midpoint timing; the paired improvement is 0.146 nats (95% patient-bootstrap CI [0.059, 0.234]). Thus the advantage remains on queries for which lesion timing is unresolved. This analysis changes the scoring subset while keeping every prediction and model parameter fixed.

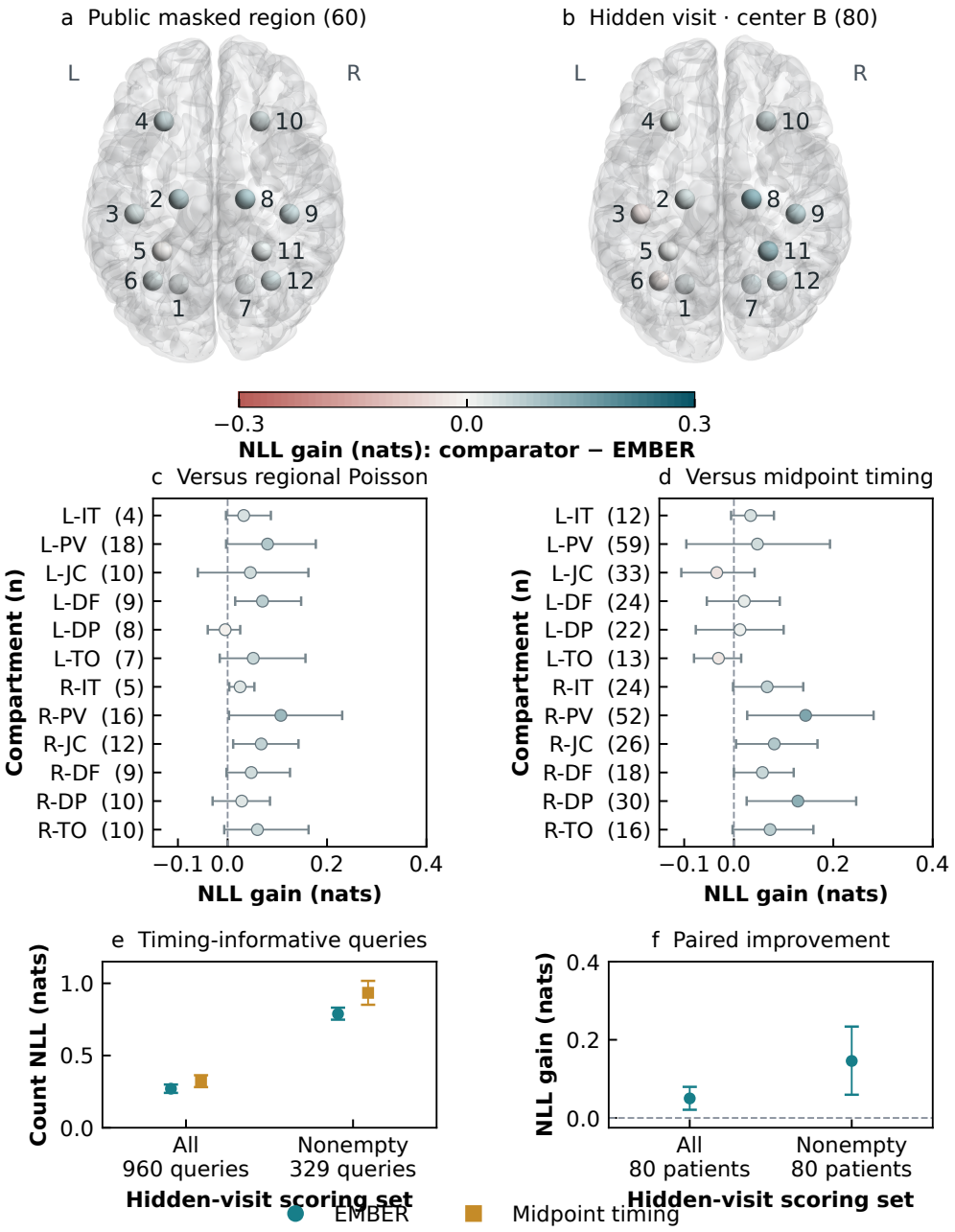



Figure 11: **Spatial predictive utility and informative lesion intervals.** (a,b) Signed regional NLL gains on a common diverging scale, with fixed-size BrainNet Viewer nodes and no display threshold. Numbers 1–12 follow the compartment row order in (c,d). These are compartment summaries, not voxelwise statistical maps. (c,d) Paired mean gains with pointwise 95% intervals from 1,200 whole-patient bootstrap samples. Parentheses give patients with a positive hidden-region target in (c), and a nonempty retained bracket in (d); all 60 or 80 patients remain in the regional score. (e) Hidden-count NLL on all 960 queries and the 329 nonempty queries. (f) Corresponding paired improvements. Subset estimates pool eligible queries and resample the 80 complete patient clusters, recomputing the denominator; 80 contribute nonempty brackets.

# H PROTOCOLS FOR SPATIAL AND HIDDEN-OBSERVATION EVALUATION

## H.1 REGIONAL ATTRIBUTION AND KERNEL DISPLAY

For each receiving compartment, the displayed attribution averages the posterior fraction of incident appearances assigned to each observed parent compartment. Observed-parent, latent-parent, and immigrant responsibilities sum to one. The figure shows only the observed block. Contributions with no observed events are integrated under the conditional predictive distribution, avoiding an undefined empirical fraction for quiet regions. The risk head instead receives the burden-normalized memory features defined in Appendix A.

The four displayed off-diagonal relationships have the largest mean training-patient responsibilities. Selection uses the 128 training patients, while the bootstrap independently resamples the patient compositions 1,200 times. Parameters remain fixed throughout this bootstrap; the intervals describe uncertainty in population composition conditional on the model. A selected arrow is a regional conditional attribution. The normalized kernel $h_{ij}(t)/\int h_{ij}(u)du$ describes the associated temporal shape, and has units of inverse months. Its area is one even when the integrated edge strength differs. The nonnegative exponential components imply a maximum at lag zero followed by monotone decay.

MSSEG-2 regional NLL uses the same 60 test patients and 12 masking queries defined in Appendix D.3. The region-level records include nonzero-case counts and patient-bootstrap intervals. All patients contribute to every region, including zero-count targets. Regional difficulty depends on the distribution of activity, not on a change of denominator. The schematic compartment positions facilitate comparison with the matrix indices, and region numbering is shared across all panels.

## H.2 BRAINNET VIEWER DISPLAY AND THRESHOLD SENSITIVITY

BrainNet Viewer (Xia et al., 2013) renders the three-dimensional scenes in Figure 13. We use its Ch2 surface with cerebellum and six bilateral compartment classes. Each class occupies a broad anatomical territory, so its representative display anchor is fixed in MNI coordinates rather than estimated as a patient-specific or atlas-parcel centroid (Table 11). The infratentorial anchor lies in the cerebellar territory; it represents the full modeled compartment. The spatial layout provides an anatomical reference for the regional dependency model, and arrows encode conditional parent attribution between compartments. Latent activity has no assigned anatomical node. This display uses existing model outputs and requires no additional MRI acquisition.

Let $\bar{r}_{ij}$ denote mean observed-parent responsibility with receiving region $i$ and source region $j$. We retain $i \neq j$ and $\bar{r}_{ij} \geq \tau$, using the same $\tau = 0.025$ for every network view. Node radius is fixed at 4.5 template millimeters; edge radius is $10\bar{r}_{ij}$ in the same display units. The common color scale spans $[0, 0.23]$. Camera, node coordinates, surface opacity, and these mappings are shared by all dorsal views. The threshold curve uses the unrounded matrix values without smoothing or a fixed edge budget. At each threshold, its denominator is the same 132 possible directed inter-compartment relationships.

The attribution balance is computed before thresholding: observed-parent mass is $12^{-1}\sum_i \sum_j \bar{r}_{ij}$, including the diagonal. Latent-parent and immigrant fractions use the corresponding receiving-region means, and the three fractions sum to one. Each receiving compartment has equal weight. These summaries describe model attribution at fixed parameters; predictive validity and uncertainty in population composition are evaluated by the separate scoring and bootstrap analyses in Figure 12.

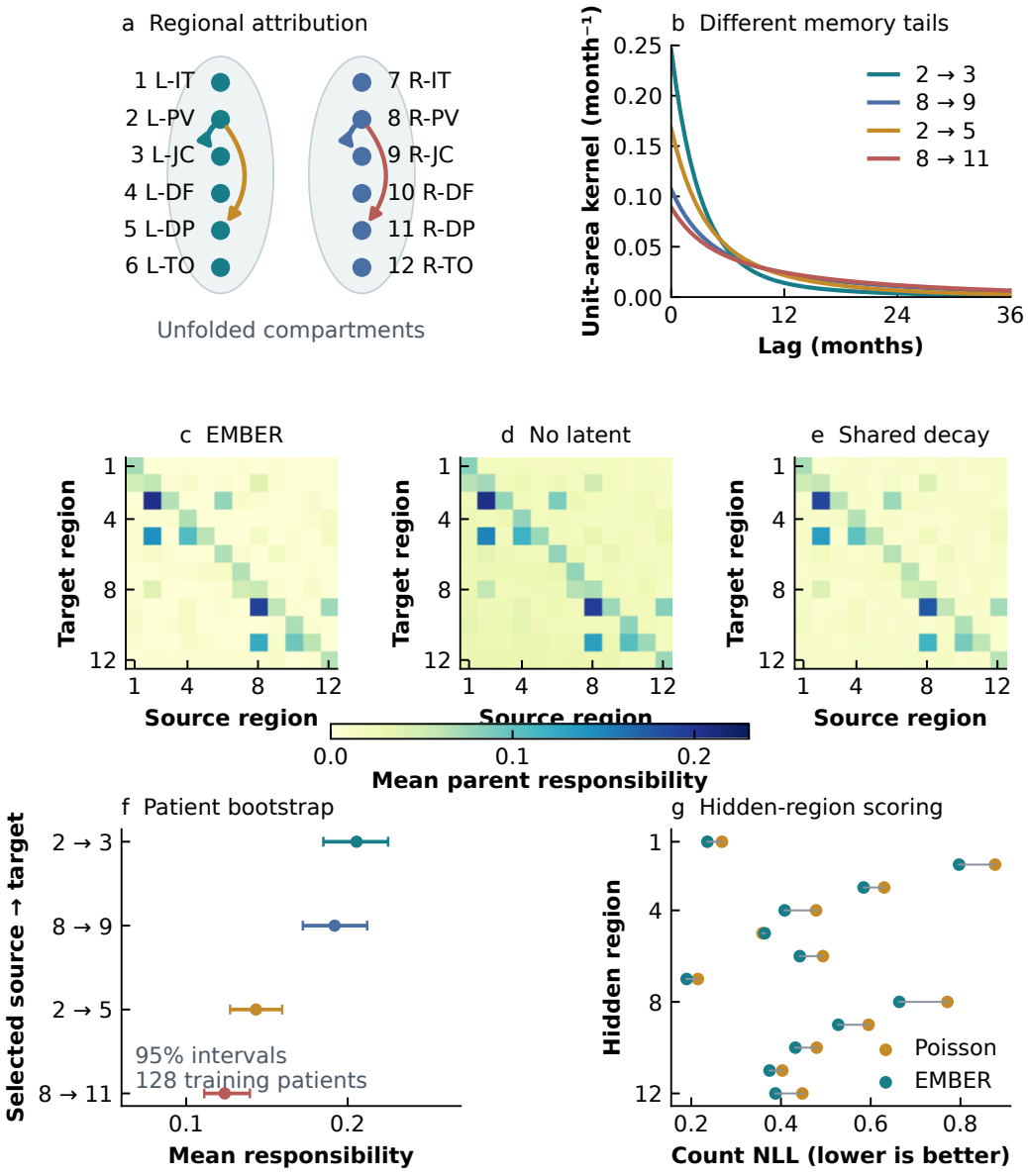



Figure 12: **Spatial dependencies, temporal memory, and predictive verification.** (a) Unfolded schematic of the 12 compartments; arrows show four relationships selected from training patients. Region numbering is shared throughout. (b) Unit-area kernel shapes for those relationships. (c–e) Mean posterior attribution to each observed parent region conditional on the receiving region, shown on one color scale. (f) 95% intervals from 1,200 resamples of the 128 training patients at fixed parameters. (g) Count NLL for each hidden region in all 60 MSSEG-2 test patients. Abbreviations: IT, infratentorial; PV, periventricular; JC, juxtacortical; DF/DP/TO, remaining deep frontal, parietal, and temporo-occipital white matter. L/R denote hemisphere.

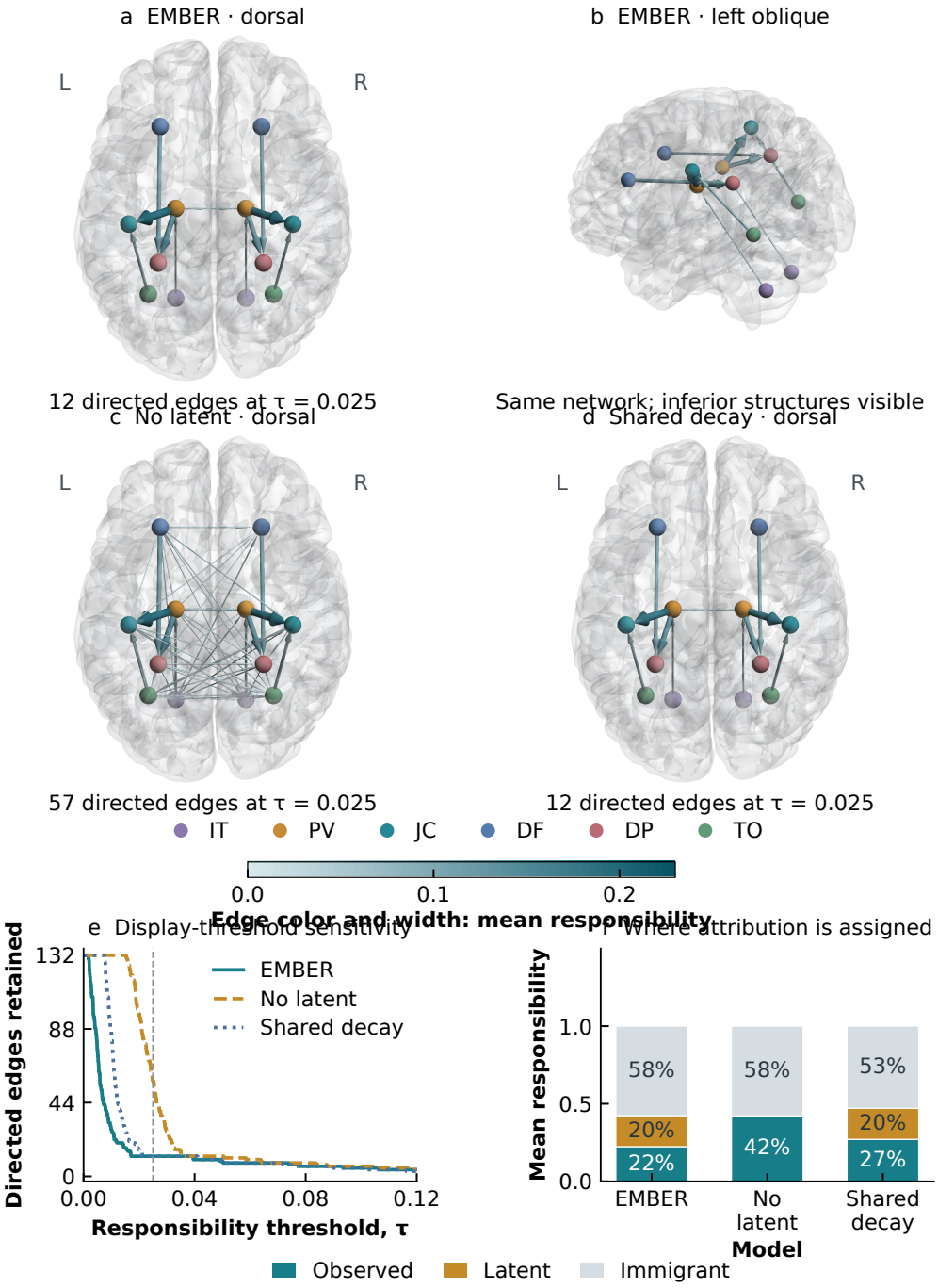



Figure 13: **Brain-network views of conditional regional attribution.** (a,b) EMBER in dorsal and left-oblique views; (c,d) ablations at the same scale. Fixed-size nodes mark representative MNI display anchors on the Ch2 template, including the cerebellum; colors identify compartments. Arrows point from source to receiving region, with common color and radius mappings for responsibility. The displayed threshold is 0.025. (e) Counts over all 132 possible off-diagonal directed relationships as the threshold changes. (f) Attribution averaged across receiving compartments, including self-parent responsibility. All summaries use the same 128 training patients. Coordinates and display conventions are given in Appendix H.2.

| Compartment (left/right index) | Left $x$ | Right $x$ | $y$ | $z$ |
|---|---|---|---|---|
| IT (1/7) | $-18$ | 18 | $-62$ | $-34$ |
| PV (2/8) | $-18$ | 18 | $-16$ | 26 |
| JC (3/9) | $-42$ | 42 | $-24$ | 45 |
| DF (4/10) | $-26$ | 26 | 26 | 26 |
| DP (5/11) | $-27$ | 27 | $-44$ | 35 |
| TO (6/12) | $-32$ | 32 | $-60$ | 6 |

Table 11: **Representative display anchors in MNI millimeters.** Negative $x$ denotes the left hemisphere; positive $y$ and $z$ denote anterior and superior directions. Anchors locate broad compartments for visualization and do not define their lesion-assignment boundaries.

### H.3 Patient resampling and cross-center network comparison

Let $r_{pij}$ be the conditional observed-parent responsibility from source $j$ to receiving compartment $i$ for patient $p$, integrating compatible event histories at fixed learned parameters. For a partition $\mathcal{P}$ of size $n$, define

$$\bar{r}_{ij}^{\mathcal{P}} = \frac{1}{n}\sum_{p\in\mathcal{P}} r_{pij}, \qquad A_{ij}^{\mathcal{P}}(\tau) = \mathbf{1}\{i \neq j,\ \bar{r}_{ij}^{\mathcal{P}} \geq \tau\}.$$

The training partition contains 128 center-A patients and the independent test partition contains 80 center-B patients. The 32 validation patients are used neither in the displayed training mean nor in the test mean. The model is evaluated with development-fitted weights at both centers. We fix $\tau = 0.025$ and select the 12 displayed relationships using training means only, sorting them by decreasing training mean.

For each partition, 1,200 resamples draw $n$ complete patient matrices with replacement. The same sampled patients contribute to every edge. Pointwise percentile intervals describe mean attributions conditional on model parameters; parameters are not refitted. Bootstrap exceedance at threshold $t$ is

$$s_{ij}(t) = \frac{1}{1200}\sum_{b=1}^{1200} \mathbf{1}\{\bar{r}_{ij}^{*(b)} \geq t\}.$$

This measures the sensitivity of a thresholded population summary to patient sampling. Figure 10f evaluates the selected relationships at $t = 0.025, 0.05, 0.075, 0.10, 0.125, 0.15, 0.175, 0.20$, tracing how their support changes as the required attribution increases. The across-center comparison reports Spearman correlation both over all 132 directed off-diagonal pairs and over the training-selected 12. At $\tau = 0.025$, both mean networks contain those same 12 pairs, giving Jaccard agreement one, with weights that agree closely across centers.

The patient fraction is a different quantity,

$$u_{ij}^{\mathcal{P}}(\tau) = \frac{1}{n}\sum_{p\in\mathcal{P}} \mathbf{1}\{r_{pij} \geq \tau\}.$$

Its intervals use the same whole-patient resamples, but threshold each patient's attribution before averaging. In contrast, $s_{ij}$ thresholds each resampled population mean. Separating these quantities explains how a relationship can be present in every bootstrap mean while exceeding the display threshold in fewer than all individual patients. The selected test fractions range from 73.8% to 100%.

### H.4 Regional predictive gains

For each task, let $\ell_{pi}^{M}$ be count NLL for model $M$ on patient $p$ and region $i$. We compute the paired gain

$$\widehat{\Delta}_i = \frac{1}{n}\sum_{p}(\ell_{pi}^{\text{comparator}} - \ell_{pi}^{\text{EMBER}}).$$

Regional percentile intervals resample complete patient score vectors, pairing the two methods within every draw. The public task uses all 60 MSSEG-2 test patients and regional Poisson as comparator. The internal hidden-visit task uses all 80 center-B patients and midpoint timing as comparator. A zero target remains a valid scored outcome. Parenthetical counts in Figure 11c,d provide activity

context and do not replace these denominators. Intervals are pointwise summaries, and the maps use a common diverging color scale across both tasks.

Let $B_{pi}$ be the incident count between the two retained visits bracketing the hidden examination. Both its count and the eligibility mask $M_{pi} = \mathbf{1}(B_{pi} > 0)$ are available from retained observations. Of the 960 queries, 631 have $B_{pi} = 0$; the nonempty subset has 329 queries in 80 patients. Its pooled score is

$$\widehat{L}_M = \frac{\sum_p \sum_i M_{pi} \ell_{pi}}{\sum_p \sum_i M_{pi}}.$$

Each bootstrap draw still samples 80 whole patient vectors, with every patient contributing at least one eligible query. The numerator and denominator are recomputed jointly, preserving within-patient dependence and variability in the number of eligible queries. For model differences, this same ratio is applied to paired score differences. Figure 11e,f reports these estimates alongside the all-query result, whose denominator is always 960. This separates the all-query result from the timing-informative subset without altering predictions or excluding patients based on hidden targets.

The brain displays use the same representative MNI anchors as Table 11. Continuous node color encodes signed mean NLL gain in nats, with a common range of $[-0.30, 0.30]$, fixed radius 5.5 template millimeters, and no display threshold. Left/right and anterior/posterior conventions match the network views. BrainNet Viewer performs the actual surface and node rendering, and the forest plots report the paired effect magnitudes with their bootstrap intervals.

### H.5 Hidden-visit selection and permissible information

Let $v_j$ be the interior examination nearest $(v_0 + T)/2$, with an earlier-visit tie rule. Exactly one scan is removed per test patient. Coarse-model inputs are computed exclusively from the retained scans: registration, lesion matching, regional measurements, and event intervals are all recomputed after removal, so the coarse model receives exactly the information available in the coarsened record.

The count target is the number of persistent incident lesions already visible at $v_j$ among those whose first appearance lies in $(v_{j-1}, v_{j+1}]$. The two retained bracketing scans identify the total number in that bracket, so the hidden count is bounded by that total. Hidden images supply the scoring reference only. Total lesion volume at the hidden visit is a separate continuous target: existing-lesion growth can change volume without another incident event. Reconstruction is retrospective conditional on pre-landmark boundaries, whereas the Ljubljana forecast conditions on baseline alone.

For count distribution $p(K)$ and hidden reference $k$, the score is $-\log p(k)$. Central discrete intervals use the smallest integer quantiles whose cumulative probabilities reach 0.05 and 0.95. Coverage, width, and NLL are reported together to give a complete picture of the predictive count law, which is necessarily discrete on sparse brackets. The nominal 90% count intervals cover 99.9% of the regional targets, with mean width 0.46 lesions per region.

Volume CRPS uses 512 stratified predictive quantiles of the nonnegative volume emission. Calibration curves evaluate central intervals at nominal levels 0.50, 0.60, 0.70, 0.80, 0.90, and 0.95. Coverage, width, and proper scores are first averaged within patient when there are multiple targets. Uncertainty resamples the 80 complete patient score vectors, with 1,200 paired draws and no region-level pseudoreplication.

Table 2b reports the corresponding proper scores, volume coverage, and interval width.

### H.6 Posterior refinements and clinical risk bands

Figure 15 connects three descriptive risk bands ($< 0.10$, $[0.10, 0.30)$, and $\geq 0.30$) at the same prediction landmark with the observed endpoint. Removing an intermediate scan moves patients between bands in both directions, and the four cases plotted in Figure 14c show how different these movements can be. The lowest-ranked of the four carries 11.3 mL of lesion load on the remaining scans; its fine-history risk is 0.020, and after deletion the coarse risk rises to 0.059 within the same band. The case at rank 30 has the smallest lesion load of the four (2.4 mL) and moves the other way, from 0.090 to 0.009. At rank 50 the risk falls from 0.225 to 0.046, crossing below $0.10$, although the retained lesion load is 2.6 mL. The highest-ranked case carries 7.7 mL, is one of the 15 PIRA

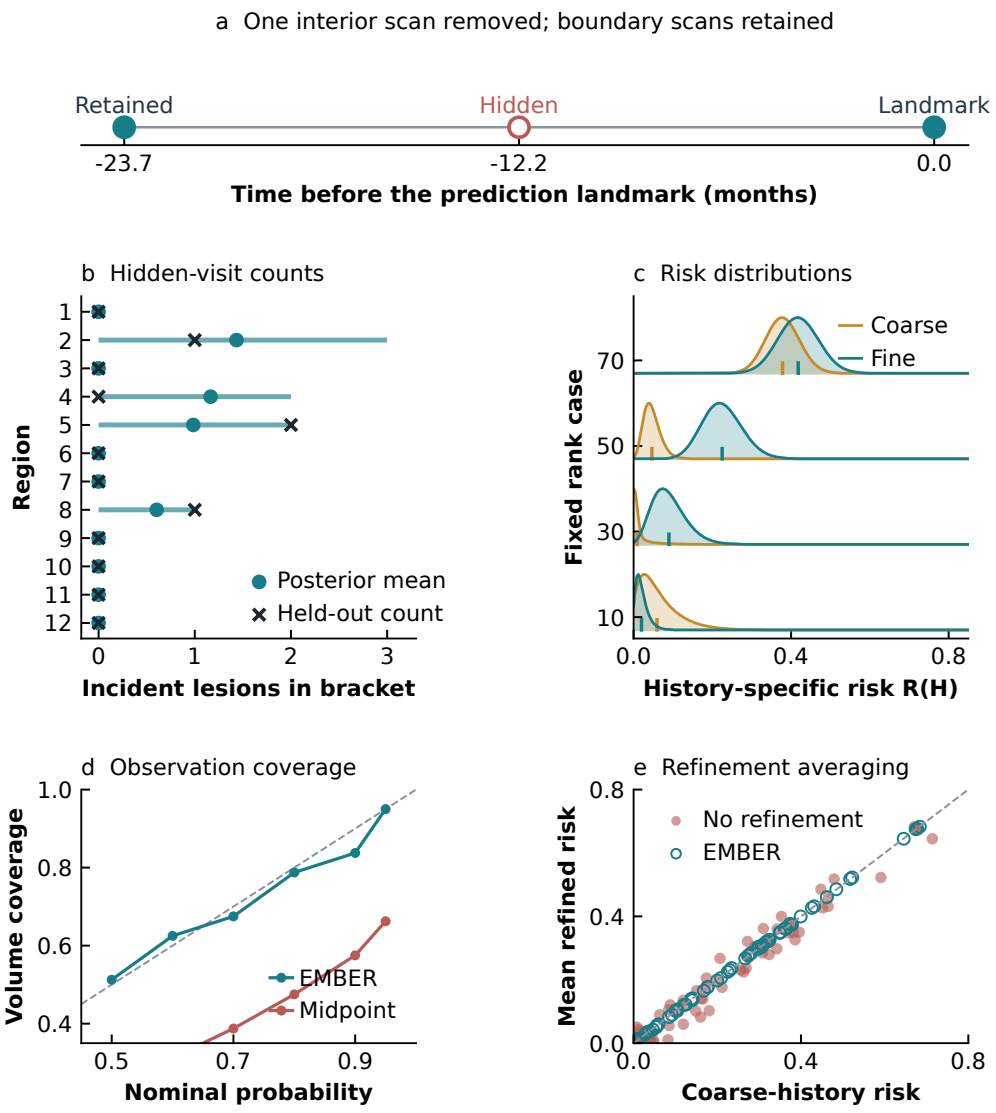



Figure 14: **What is lost when an interior scan is hidden?** (a) Retained and hidden scan times for a case chosen using retained-bracket event counts. (b) Predicted incident appearances by the hidden visit, within its retained bracketing interval; bars are central 90% predictive count intervals. (c) Four cases at fixed ranks of fine-history risk, with fine and coarse distributions of $R(H)$. Densities are individually peak-normalized for display; ticks indicate means. (d) Empirical predictive coverage of the hidden total lesion volume across 80 patients. (e) Coarse risk versus the mean over 128 compatible refinements, evaluated with and without refinement learning. Panels (b,d) summarize observation targets, whereas panel (c) summarizes the risk distribution.

outcomes, and stays in the top band throughout, from 0.418 to 0.378. Across the cohort the two directions are close to balanced: deletion raises risk for 43 patients and lowers it for 37.

Band boundaries are fixed for display and are applied identically to the fine- and coarse-history risks, and the middle nodes represent the same person and horizon under less imaging information. Every patient's mass is conserved across both links, and the 15 PIRA outcomes are unchanged by scan removal. The paired Brier difference in Section 6.4 evaluates the probability changes independently of these cutoffs.

Fine-history clinical probabilities exactly match the frozen center-B predictions in Table 1, and each coarse prediction is recomputed after its fixed scan deletion. Compatible finer observations are integrated using 128 stratified refinements for this evaluation, exceeding the eight used per training update. The reported residual compares the coarse prediction with their weighted mean rather than with the single observed fine-history risk.

Uncertainty summaries use 256 weighted history-risk values at each observation resolution; point predictions retain the frozen 128-history outputs. Each curve in Figure 14c is divided by its own peak for visibility, while its horizontal shape, mean, and quantiles retain their interpretation. Cases are the patients at sorted fine-risk ranks 10, 30, 50, and 70, with patient identifier breaking ties, and the count example in Figure 14b is chosen at the median number of active compartments in the retained bracket. Both rules depend only on retained scan counts and the fine-risk ordering.

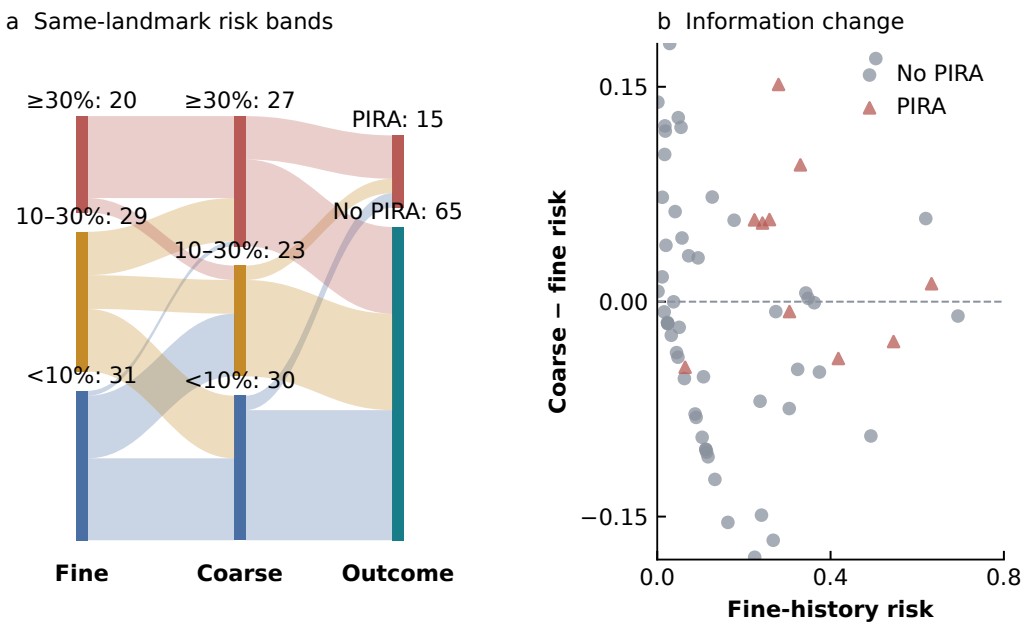



Figure 15: **Risk reclassification under information loss.** (a) Flow widths count the same 80 patients across fine-history risk, coarse-history risk, and 24-month PIRA outcome. Counts are printed at each node. (b) Individual probability changes, with outcome shown only for evaluation. Both increases and decreases are compatible with posterior refinement.

# I FACTORIAL CONTROLS AND UNCERTAINTY

## I.1 JOINT EFFECTS OF SPARSE SCANS AND HIDDEN ACTIVITY

Two factorial grids separate the effects of acquisition and model misspecification (Figure 16). Both use median scan gaps of 1, 3, 6, and 12 months. The first varies the fraction of observed events descending from latent immigrants from 0 to 0.8; the second varies edge-specific kernel dispersion from a shared shape to heterogeneous mixtures. Each cell uses ten independent parameter replicates, with 512 training and 256 test histories per replicate. Both grids are controlled mechanism evaluations, scored against reference branch flows and independently of the clinical cohort.

We hold the expected observed burden at six events per 24-month history. Adjusting immigrant allocation changes latent-origin fraction without changing the integrated excitation matrix within a replicate; changing mixture shape also leaves this matrix fixed. Thus the two factorial grids probe distinct aspects of latent organization at matched overall burden. At a 12-month gap and latent-origin fraction 0.8, branch-flow error is 0.183 for EMBER, 0.392 with midpoint timing, and 0.390 without latent activity. The no-latent model is competitive when latent-origin activity is absent; the shared-decay model is competitive at zero kernel dispersion. Their gaps widen in the conditions that violate their respective assumptions. Table 12 reports replicate-level uncertainty and false-edge summaries for this high-pressure condition.

## I.2 FACTORIAL DESIGN AND SCORING

For each of ten parameter replicates, let $C$ be the nonnegative integrated matrix of 12 observed and two hidden components. Its maximum column sum is below 0.85. With stationary initialization and constant immigration vector $\mu$, the rate vector is $\Lambda = (I - C)^{-1}\mu$. Partition immigration into observed-origin and latent-origin terms, $\mu_O$ and $\mu_L$. We set

$$f_L = \frac{\sum_{i=1}^{12}[(I-C)^{-1}\mu_L]_i}{\sum_{i=1}^{12}\Lambda_i}, \qquad \sum_{i=1}^{12}\Lambda_i = 0.25 \text{ month}^{-1}.$$

The first constraint controls the fraction descending from latent immigrants; the second gives six expected observed events in 24 months. This lineage fraction includes descendants reached through intervening observed events, and differs from the immediate hidden-parent fraction. Realized counts fluctuate around this expectation. All-zero intervals are retained. Observations use 15% scheduling jitter around each specified median gap.

For the second grid, edge weights are $w_{ij}(\alpha) = (1-\alpha)(1/3, 1/3, 1/3) + \alpha w^*_{ij}$. Each vector remains on the simplex. Holding the integrated edge coefficients fixed preserves stability and the stationary mean rate while changing the mixture of 3-, 9-, and 24-month memory. The no-latent model integrates event times but omits hidden components; the midpoint model retains hidden components and heterogeneous kernels but collapses observed timing; shared decay retains interval integration and hidden components but forces one kernel shape. These single-component controls differ from the combined shared-kernel, midpoint clinical comparator.

Each replicate contributes one mean error per condition after averaging its 256 test histories. All four methods use the same test histories and observation masks within a replicate. The recorded intervals are mean $\pm t_{9,0.975}s/\sqrt{10}$ across parameter replicates; cell-level comparisons pair the replicate identifiers. The grid evaluates the mechanism-level contrast at every cell. Branch-flow error is normalized by the known reference flow norm as in Appendix F.5. The specificity score is the fraction of absent off-diagonal observed edges selected by a fixed threshold on posterior-mean responsibility. The numerical records contain both metrics for every cell, together with the stable matrices and stationary-rate controls.

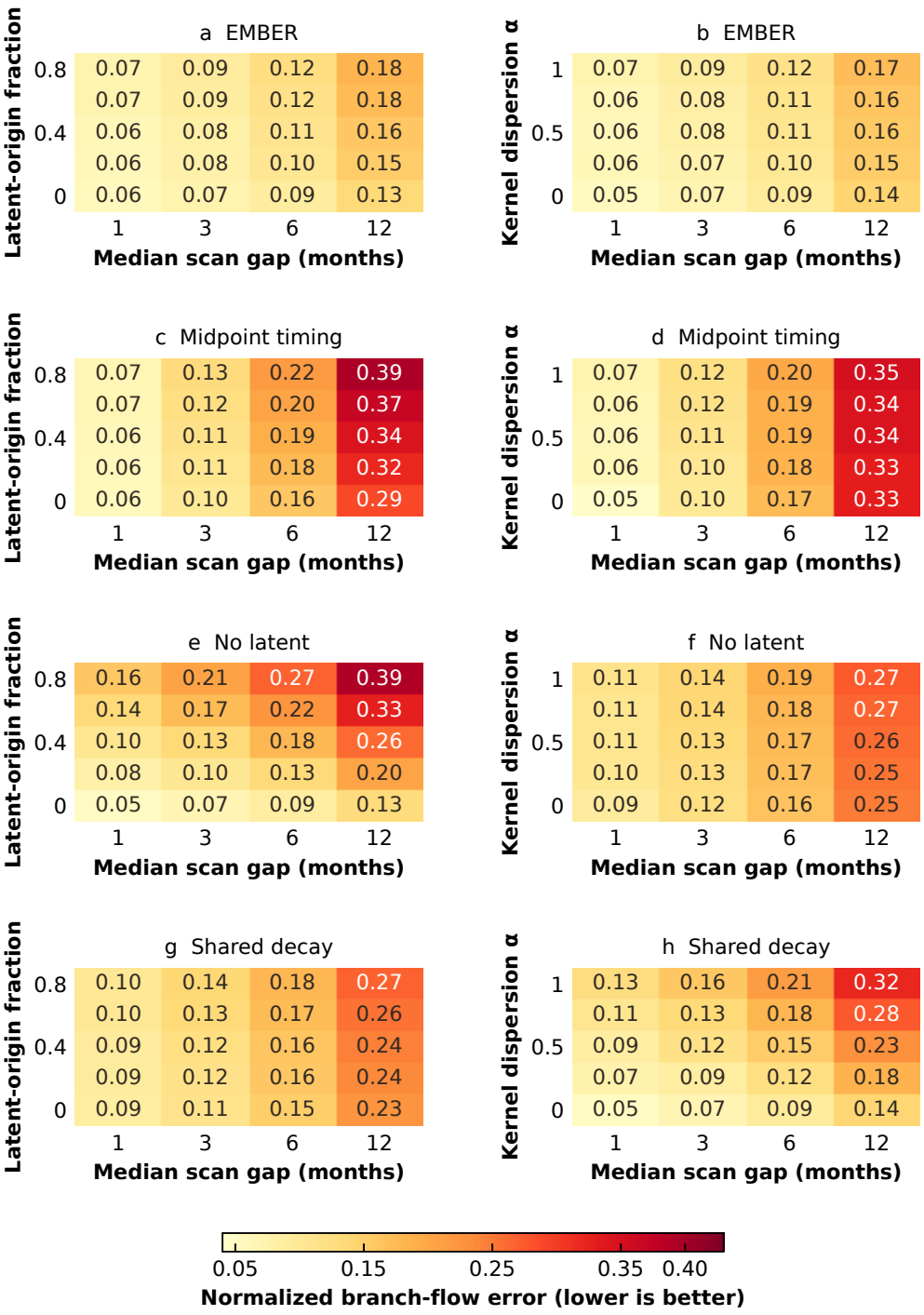



Figure 16: **Observation sparsity interacts with distinct model assumptions.** Every cell shows mean normalized branch-flow error over ten parameter replicates, with one color scale across all eight panels. Left: latent-immigrant origin fraction varies at kernel dispersion $\alpha = 0.6$. Right: dispersion varies at latent-origin fraction 0.4. At $\alpha = 0$, every edge shares mixture weights $(1/3, 1/3, 1/3)$; at $\alpha = 1$, edges use their individual weights on the 3-, 9-, and 24-month basis kernels. Expected observed event burden, sequence length, sample sizes, stability, and development-selected thresholds are controlled across the grid. Cells are shown without spatial smoothing.

| Model | Branch-flow error (95% CI) | False-edge fraction (95% CI) |
|---|---|---|
| EMBER | 0.183 [0.180, 0.186] | 0.072 [0.070, 0.075] |
| Midpoint timing | 0.392 [0.389, 0.396] | 0.222 [0.221, 0.224] |
| No latent | 0.390 [0.386, 0.393] | 0.298 [0.296, 0.300] |
| Shared decay | 0.268 [0.264, 0.273] | 0.109 [0.107, 0.112] |

Table 12: **A high-pressure condition.** Median gap 12 months, latent-origin fraction 0.8, and kernel dispersion 0.6. Intervals use ten independent parameter replicates. False-edge proportions use only absent off-diagonal observed edges; lower values are better.