SUBTRACE: JOINT AMORTIZED INFERENCE OF DISEASE PROGRESSION SUBTYPES AND TIMELINES FROM CROSS-SECTIONAL DATA

Anonymous authors
Paper under double-blind review

ABSTRACT

Event-based models infer disease progression from cross-sectional biomarkers, but subtype extensions such as SuStaIn and Bayesian EBM subtyping (BEBMS) are computationally expensive and return only ordinal event sequences. We propose SUBTRACE, a simulation-trained Transformer for amortized subtype-aware disease progression inference. Given a cohort, SUBTRACE predicts subtype-specific continuous biomarker timelines, patient subtype assignments, and patient stages in a single forward pass. It does so by learning per-patient, per-biomarker abnormality evidence and fitting subtype timelines to that evidence at the predicted disease stage. Across nine synthetic experiments, SUBTRACE improves ordering and staging over BEBMS and SuStaIn. Most importantly, the advantage is widest in the 100-biomarker setting, where the fitted baselines recover little of the ordering while SUBTRACE remains accurate. Applied to ADNI and NACC, independently trained models recover similar Alzheimer's disease subtype patterns consistent with prior research. These results support simulation-trained amortized inference as a practical alternative to per-dataset MCMC for subtype-aware disease progression modeling. Code is available at `https://gitlab.com/4peerreview/iclr2027_subtrace`.

1 INTRODUCTION

Disease progression can be understood in different ways. Here, we study it as a temporal sequence of events where biomarkers become pathological as the disease advances, as shown in Figures 2 and 3. Predicting this progression timeline is clinically important, as it enables more accurate diagnosis and helps designing therapeutical interventions (Jack et al., 2010). This task is tractable when detailed, high-quality longitudinal data are available: we can directly observe the timeline. In practice, however, such data are expensive, slow to collect, and logistically difficult to obtain. Alzheimer's disease, for instance, can span over two decades (Villemagne et al., 2013). Each clinical visit involves costly imaging, and over such a long horizon patients die or drop out, measurement standards change, and missing values accumulate. These limitations have led many researchers to focus on modeling cross-sectional datasets instead (Young et al., 2024).

Consider a concrete example. Table 1 shows a small cross-sectional dataset: each row is one participant, with a binary disease label and a set of biomarker measurements (brain region volumes, cognitive test scores, lab values). Assuming a shared disease progression, i.e., the order in which biomarkers become pathological as the disease advances, two tasks follow. *Ordering* recovers this order (e.g., amyloid $\rightarrow$ MMSE $\rightarrow$ entorhinal volume $\rightarrow$ hippocampal volume); this becomes intractable by brute-force search once the number of biomarkers grows past a few dozen, which is why the EBM family below was developed. *Staging* assigns each participant an ordinal position along this order: a stage of z means the first z biomarkers in the progression have become pathological for that participant, with $z = 0$ implying a healthy participant.

Participants need not follow one shared progression, however; subtypes are common (Murray et al., 2011; Ferreira et al., 2020; Vogel et al., 2021). With subtypes, ordering becomes finding one progression per subtype; hence, we have a third task, *subtyping*, which assigns each participant to a subtype. Subtyping is harder than adding a classification head alongside ordering and staging: the

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.

three tasks are mutually dependent. Correct subtype assignment requires knowing each subtype's ordering, and vice versa; staging depends on both. If ordering is wrong, accurate-looking subtype and stage assignments carry little clinical meaning. This coupling, not merely predicting three quantities instead of two, is what makes subtype-aware progression modeling difficult; especially since subtyping adds no new information at inference time beyond what ordering and staging already use. For example, whether we have the *subtyping* task or not, the input data is always as shown in Table 1.

Table 1: A small cross-sectional dataset. Each row is a participant with a diagnosis label and continuous biomarker measurements.

Participant	Diseased	ABETA	Entorhinal	Hippocampus	MMSE
1	True	1243	0.0026	0.0004	28
2	True	682	0.0017	0.0003	21
3	False	1812	0.0002	0.0007	30
⋮	⋮	⋮	⋮	⋮	⋮

A large body of work has developed Event-Based Models (EBMs; Fonteijn et al., 2012) to recover disease progression from cross-sectional data. The EBM family assumes that each biomarker transitions irreversibly from a benign to a pathological state at some point in the disease timeline, and uses a maximum-likelihood criterion to infer the ordering of these events. Extensions have improved robustness to non-Gaussian distributions (Firth et al., 2020), relaxed the uniform stage prior assumption (Hao et al., 2025), and modeled disease heterogeneity through multiple subtypes (Young et al., 2018; Hao & Austerweil, 2025; Tandon et al., 2024).

In particular, the Subtype and Stage Inference model (SuStaIn; Young et al., 2018) has become the standard tool for subtype-level EBM inference, with applications to Alzheimer's disease (Vogel et al., 2021; Salvadó et al., 2024; Estarellas et al., 2024), multiple sclerosis (Eshaghi et al., 2021), and Lewy body disease (Mastenbroek et al., 2024). The Bayesian Event-Based Model for Subtyping (BEBMS; Hao & Austerweil, 2025) improved on SuStaIn by placing Bayesian priors on subtype and stage distributions and updating all parameters jointly, achieving improved ordering, staging, and subtype assignment.

Yet these EBM approaches to recover subtypes in disease progression share two fundamental limitations. First, they produce only an ordinal event sequence: they tell us *which* biomarker becomes pathological before which, but not *how much earlier*. The gaps between events are unknown, which limits the clinical utility for prognosis. Second, they still struggle in ordering, subtyping, and staging, especially with high-dimensional data where the number of biomarkers grows beyond a few dozen.

TEMPO (Hao & Austerweil, 2026) addressed the first limitation for the single-subtype case, training a Transformer on synthetic data to recover continuous event timelines instead of ordinal sequences. It does not, however, extend to subtypes by simply adding a subtype-classification head, since ordering, staging, and subtyping are mutually dependent and training three independent heads discards the coupling between them (Appendix G). In this paper, we introduce SUBTRACE to solve this challenge, which ties the three tasks together structurally, described in Section 3.

Across nine synthetic experiments evaluated at both 12 and 100 biomarkers, SUBTRACE outperforms BEBMS (Hao & Austerweil, 2025) and SuStaIn (Young et al., 2018) on biomarker ordering and patient staging across nearly all settings. The advantage is especially pronounced in high-dimensional experiments: at 100 biomarkers, SUBTRACE achieves a mean Kendall τ distance of 0.119, compared to 0.378 for BEBMS (a 69% reduction over BEBMS) and 0.402 for SuStaIn (GMM). SUBTRACE also has the lowest normalized stage error in both settings. Applied to two separately collected datasets, the Alzheimer's Disease Neuroimaging Initiative (ADNI; Mueller et al., 2005) and the National Alzheimer's Coordinating Center (NACC; Morris et al., 2006), SUBTRACE recovers subtypes with meaningful progression patterns broadly consistent with established neuropathological findings (Murray et al., 2011; Braak & Braak, 1991; Jack et al., 2010; Grober et al., 2008; Jack Jr et al., 2009; Jack & Holtzman, 2013).

2 PROBLEM SETUP AND RELATED WORK

We observe a cohort of J participants with B biomarkers. For participant j, we have measurements $x_{j,1}, \ldots, x_{j,B}$ and a binary disease label $d_j \in \{0, 1\}$ (0 = healthy control, 1 = patient). Each biomarker b has a *benign* distribution ϕ_b and a *pathological* distribution θ_b. Disease progression is modeled as a sequence of *events*: biomarker b transitions irreversibly from ϕ_b to θ_b at some unknown point in the patient's disease course. The *ordering* $S = (S_1, S_2, \ldots, S_B)$ is a permutation of biomarkers specifying which becomes pathological first, second, and so on. A patient's *stage* $z_j \in \{0, 1, \ldots, B\}$ counts how many events have occurred; $z_j = 0$ means no pathology (healthy). The goal is to recover K subtype-specific orderings and jointly assign each patient to a subtype and a stage.

The Event-Based Model (EBM; Fonteijn et al., 2012) infers a single ordering S by maximum likelihood. Given S and stage z_j, biomarkers $S_1, \ldots, S_{z_j}$ are post-event (drawn from θ) and the rest are pre-event (drawn from ϕ):

$$p(\mathbf{x}_j \mid S, z_j) = \prod_{i=1}^{z_j} p(x_{j,S_i} \mid \theta_{S_i}) \prod_{i=z_j+1}^{B} p(x_{j,S_i} \mid \phi_{S_i}) \,. \tag{1}$$

Since stage is unknown, it is marginalized: $P(\mathbf{x}_j \mid S) = \sum_{z_j=0}^{B} P(z_j)\, p(\mathbf{x}_j \mid S, z_j)$, where $P(z_j)$ is a stage prior. To handle K subtypes, each with its own ordering $\mathbf{S}^{(k)}$, BEBMS (Hao & Austerweil, 2025) sums over subtypes and stages:

$$p(\mathbf{x}_j \mid \mathbf{S}) = \sum_{k=1}^{K} \pi_k \sum_{z_j=0}^{B} \pi_{z_j|k}\, p(\mathbf{x}_j \mid \mathbf{S}^{(k)},\, z_j), \tag{2}$$

where π_k is the subtype prior and $\pi_{z_j|k}$ is the stage prior for subtype k.

3 METHODS

The general idea of SUBTRACE is to predict the benign/pathological state of each (patient j, biomarker b) pair, denoted $A_{j,b}$, where $A_{j,b} = 1$ means biomarker b is pathological for patient j and $A_{j,b} = 0$ means benign, from two complementary perspectives. The first is a *data-driven perspective*: we train a proxy $p_{j,b} \approx P(A_{j,b} = 1 \mid \text{data})$ directly from ground-truth. The second is a *structural perspective*: given estimated disease stage $\hat{z}_j$ and subtype-specific event position $r_{k,b}$, we compute $\mu_{j,k,b} = \sigma(\alpha(\hat{z}_j - r_{k,b}))$, which encodes the intuitive rule that biomarker b should be abnormal for patient j if j's stage is past b's event position in their subtype. The ground truth of this rule is:

$$A_{j,b} = d_j\, \mathbf{1}\{z_j \geq r_{k_j,b}\}.$$

Training minimizes the discrepancy between these two perspectives alongside direct supervision on orderings, stages, and subtype assignments. Figure 1 shows the architecture. We describe it in the execution order: tokenization and seed routing, a provisional timeline pass, evidence and stage prediction, profile-fit refinement, a final timeline pass, and final profile fitting.

3.1 INPUTS AND TARGETS

Let $\mathbf{X} \in \mathbb{R}^{J \times B}$ be a standardized cohort with J participants and B biomarkers, and let $d_j \in \{0, 1\}$ indicate control/diseased status. In synthetic training data, we also observe the subtype label $k_j \in \{1, \ldots, K\}$, stage $z_j \in \{0, \ldots, B\}$, and subtype-specific event times $\xi_{k,b} \in [0, B]$. The stage z_j is the number of biomarker events that have occurred for patient j; we train on its normalized form $\tilde{z}_j = z_j / B$. We also normalize event times, $r^*_{k,b} = \xi_{k,b}/B$. Sorting $\xi_{k,1}, \ldots, \xi_{k,B}$ gives the ordinal ordering $\mathbf{S}^{(k)} = (S_1^{(k)}, \ldots, S_B^{(k)})$, where $S_i^{(k)}$ is the biomarker at rank i. Event times preserve spacing; the ordering preserves only relative rank.

These labels define the affected state

$$A^*_{j,b} = d_j\, \mathbf{1}\{z_j \geq \xi_{k_j,b}\} = d_j\, \mathbf{1}\{\tilde{z}_j \geq r^*_{k_j,b}\}. \tag{3}$$

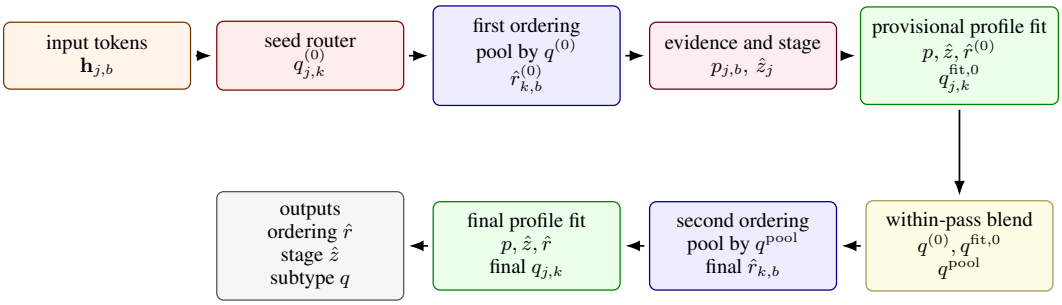


Figure 1: SUBTRACE two-pass architecture, shown as a sequential flow. Evidence and stage are computed once from the input tokens and reused by both profile-fit steps. The first ordering pass produces provisional timelines used only to form q^{pool}; the second ordering pass gives the final event timelines, and the final profile fit gives subtype probabilities. The stage output remains $\hat{z}_j$ from the stage head.

Ground-truth subtypes, stages, event times, and affected states are used only during training. At inference, SUBTRACE receives only $\mathbf{X}$ and, for the non-blind model, disease labels $\mathbf{d}$.

3.2 TOKENIZATION AND SEED ROUTING

For each patient-biomarker pair, the model builds a token from the standardized value, within-patient rank, cohort rank, and diagnosis:

$$\mathbf{u}_{j,b} = [x_{j,b}, R^{\mathrm{within}}_{j,b}, R^{\mathrm{cohort}}_{j,b}, d_j], \qquad \mathbf{h}_{j,b} = f_\psi(\mathbf{u}_{j,b}) \in \mathbb{R}^d. \tag{4}$$

where $x_{j,b}$ is the standardized measurement, $R^{\mathrm{within}}_{j,b}$ ranks biomarker b within patient j, and $R^{\mathrm{cohort}}_{j,b}$ ranks patient j within biomarker b. These rank features improve robustness to heavy-tailed measurements.

A seed router gives an initial subtype assignment. It maps each patient's biomarker values, ranks, and diagnosis label through a patient MLP and cohort Transformer to a contextual embedding $\mathbf{g}_j$, then applies

$$q^{(0)}_{j,k} = \mathrm{softmax}_k(\mathbf{w}_k^\top \mathbf{g}_j + a_k). \tag{5}$$

These seed probabilities are used for provisional pooling; final subtype probabilities come from profile fitting below.

3.3 FIRST EVENT-TIMELINE PASS

After the seed router, SUBTRACE makes a first estimate of the subtype event timelines by averaging patient-biomarker embeddings with seed-router weights:

$$\mathbf{t}^{(0)}_{k,b} = \sum_{j=1}^{J} \bar{q}^{(0)}_{j,k} \mathbf{h}_{j,b}, \qquad \bar{q}^{(0)}_{j,k} = \frac{q^{(0)}_{j,k}\omega_j}{\sum_i q^{(0)}_{i,k}\omega_i + \varepsilon}. \tag{6}$$

Here ω_j downweights controls and patients with very weak abnormality evidence; Appendix B gives its exact form.

The pooled biomarker sequence $(\mathbf{t}^{(0)}_{k,1}, \ldots, \mathbf{t}^{(0)}_{k,B})$ is passed through a biomarker Transformer and a scalar ranking head:

$$s^{(0)}_{k,b} = a_{\mathrm{ord}} \left[T_{\mathrm{ord}}(\mathbf{t}^{(0)}_{k,1}, \ldots, \mathbf{t}^{(0)}_{k,B})_b \right], \qquad \hat{r}^{(0)}_{k,b} = \sigma(s^{(0)}_{k,b}). \tag{7}$$

The provisional event position $\hat{r}^{(0)}_{k,b} \in [0,1]$ gives an initial timeline, but not the final output ordering.

3.4 Evidence and Stage Heads

The evidence branch estimates which biomarkers have already become abnormal. A separate biomarker Transformer processes each patient's biomarker tokens:

$$e_{j,b} = a_{\text{ev}}\left[T_{\text{ev}}(\mathbf{h}_{j,1}, \ldots, \mathbf{h}_{j,B})_b\right], \qquad p_{j,b} = \sigma(e_{j,b}). \tag{8}$$

The probability $p_{j,b}$ is trained to approximate $A^*_{j,b}$ in Eq. 3 using balanced binary cross-entropy.

SUBTRACE predicts one normalized stage per patient, $\hat{z}_j \in [0, 1]$, using a stage head over biomarker values, diagnosis, learned evidence, and biomarker identity embeddings:

$$\hat{z}_j = g_z(\mathbf{x}_j, \mathbf{p}_j, d_j). \tag{9}$$

$\hat{z}_j$ is the final stage output and is trained by mean squared error against $\tilde{z}_j$.

3.5 Profile Fit, Refinement, and Final Outputs

The model asks how well each provisional subtype timeline explains each patient's evidence profile at the predicted stage, using a mismatch between the structural profile implied by $\hat{z}_j$ and $\hat{r}^{(0)}_{k,b}$ and the learned evidence $p_{j,b}$; this yields a provisional subtype probability $q^{\text{fit},0}_{j,k}$, which is not an additional learned head but a quantity implied by the evidence, stage, and provisional timeline already computed. This is blended with the seed-router probabilities and used for a second, final timeline pass through the same biomarker Transformer and ranking head, giving the final event positions $\hat{r}_{k,b}$. SUBTRACE then recomputes the profile fit using this final timeline to obtain the final subtype probabilities $q_{j,k}$. The final outputs are the continuous event positions $\hat{r}_{k,b}$, subtype probabilities $q_{j,k}$, and stage $\hat{z}_j$; all refinement terms are recomputed for the current cohort inside one forward pass, with no profile-fit weights stored or reused across runs. Full equations are in Appendix B.

3.6 Training Objective and Permutation Matching

The loss combines ordering, subtype, stage, evidence, and auxiliary terms. Because subtype labels are arbitrary, ordering and subtype losses first match predicted slots to true subtype labels for each synthetic cohort. Let $C^{\text{ord}}_{k,m}$ be the ordering cost between true subtype k and predicted slot m, and let $C^{\text{sub}}_{k,m}$ be the corresponding subtype cross-entropy cost. We choose

$$\zeta^* = \arg\min_{\zeta \in \mathcal{S}_K} \sum_{k=1}^{K} \left[C^{\text{ord}}_{k,\zeta(k)} + \eta C^{\text{sub}}_{k,\zeta(k)}\right], \tag{10}$$

using the Hungarian algorithm (Kuhn, 1955). The main objective is

$$\mathcal{L} = \lambda_{\text{ord}}\mathcal{L}_{\text{ord}} + \lambda_{\text{sub}}\mathcal{L}_{\text{sub}} + \lambda_{\text{stage}}\mathcal{L}_{\text{stage}} + \lambda_{\text{evid}}\mathcal{L}_{\text{evid}} + \mathcal{L}_{\text{aux}}, \tag{11}$$

where $\mathcal{L}_{\text{ord}}$ and $\mathcal{L}_{\text{sub}}$ are computed after applying ζ^*. Auxiliary terms include seed-router supervision, a small control-stage penalty, and an evidence-side subtype contrast. Appendix B gives the exact loss definitions.

3.7 Blind Variant

SuStaIn (Young et al., 2018) does not use diagnosis labels directly. For fair comparison, the blind SUBTRACE variant uses the same architecture and training targets but removes diagnosis from model input at inference. The profile-fit equations are unchanged.

4 Synthetic Experiments

We evaluate across nine synthetic experiments varying the stages, events, and biomarker distributions (Table 5). Some experiments use an EBM-style abrupt pre/post event model (Fonteijn et al., 2012; Hao et al., 2025): for patient j, subtype k_j, and biomarker b, the affected state is determined by whether the patient's stage z_j is past the event position of b in that subtype. Other experiments use

a sigmoid measurement model (Venkatraghavan et al., 2019; Young et al., 2015; Hao et al., 2025), where biomarker measurements move gradually away from the benign state as stage advances beyond the event position.

This design isolates the effects of measurement family (EBM vs. Sigmoid), biomarker distribution (Normal vs. Non-Normal), stage prior (Dirichlet-Multinomial vs. Uniform), stage type (ordinal vs. continuous), and event-time type (discrete vs. continuous). Each experiment runs in two dimensionality settings: low-dimensional (12 ADNI-based biomarkers) and high-dimensional (100 synthetic biomarkers), with $K = 3$ subtypes, 950 training cohorts, 50 validation cohorts, and 50 held-out test cohorts. The low-dimensional setting is designed to resemble the ADNI application, while the 100-biomarker setting tests scalability. We document our design choices of synthetic data generation in Appendix D.

Baselines are BEBMS (Hao & Austerweil, 2025) and its blind version, and two variants of SuStaIn (GMM and KDE) (Young et al., 2018; Aksman et al., 2021). We also evaluate a blind SUBTRACE variant that removes diagnosis labels from the model input at inference. TEMPO (Hao & Austerweil, 2026) cannot be used directly here because it assumes a single subtype. To use it as a baseline, we first apply K-Means to cluster each dataset into 3 groups. More details of this "K-Means + TEMPO" approach can be found in Appendix H.

For each experiment, all algorithms are evaluated against the same 50 held-out test datasets. Metrics are Kendall τ distance for ordering recovery (lower is better), Adjusted Rand Index (ARI) for subtype assignment (higher is better), stage MAE, mean healthy stage, and runtime. The hard subtype label used for ARI is $\hat{k}_j = \arg\max_k q_{j,k}$. Ordering and subtype metrics use Hungarian matching (Kuhn, 1955) because subtype labels are arbitrary. Because baselines return ordinal stages, continuous SUBTRACE stages are converted to ordinal stages by rounding $\hat{z}_j \cdot B$ and clipping to the valid range. We also report event position MAE for SUBTRACE, SUBTRACE (Blind), and K-Means + TEMPO, the only methods that produce continuous event positions; this is a diagnostic of prediction granularity, not a benchmark comparison, since other baseline models cannot output continuous disease stages. SUBTRACE averages 1.74 positions of error at 12 biomarkers and 13.5 at 100, against 3.52 and 32.2 for random (Appendix N). Details on data generation and baselines are in Appendix E.

Table 2 reports results aggregated over all 450 test datasets. SUBTRACE has the best aggregate ordering, stage MAE, and healthy-stage calibration in both dimensions, and the best aggregate ARI in the 12-biomarker setting. At 100 biomarkers, SuStaIn (GMM) has higher aggregate ARI (0.169 vs. 0.108), but much worse ordering, stage MAE, and control-stage calibration.

Ordering is where SUBTRACE's advantage over the fitted baselines is clearest, in both settings. At 12 biomarkers it reaches $\tau = 0.162$ against 0.215 for BEBMS (a 25% reduction), 0.298 for SuStaIn (GMM), and 0.351 for SuStaIn (KDE). At 100 biomarkers its lead is far larger (0.119 against 0.378, 0.402, and 0.366 respectively), with the fitted baselines landing closer to the random reference (0.479) than to SUBTRACE. The two settings are separate simulator configurations whose subtype orderings differ in how similar they are to one another (Appendix E), so each compares methods fairly at its own dimensionality, but we do not attribute SUBTRACE's larger lead at 100 biomarkers to dimensionality by itself. Stage accuracy shows the same pattern after normalizing by B: SUBTRACE has the best MAE$/B$ in each setting, 0.078 at 12 bio and 0.041 at 100, against 0.103 and 0.091 for BEBMS, 0.152 and 0.198 for SuStaIn (GMM), and 0.193 and 0.150 for SuStaIn (KDE).

SUBTRACE training took roughly 14 min per 12-bio experiment and about 1.9 h per 100-bio experiment on an NVIDIA RTX 6000 Ada Generation GPU with 45 GB memory; after training, inference is a single forward pass, about 0.01 s per low-dimensional dataset (39 s for 4050) and 0.04 s per high-dimensional one (3 min for 4050). Benchmark algorithms (BEBMS and SuStaIn) were run on a CPU cluster, with Exp. 9 redone on a combination of GCP and OCI (Appendix F), and require per-dataset inference, e.g., $\sim$94 min (BEBMS) or $\sim$475 min (SuStaIn) at 100 bio.

Per-experiment results show the same structure as the aggregate results in Table 2; full per-experiment tables and discussion are in Appendix Q.

We also ran $K = 1$ and $K = 5$ robustness checks against TEMPO and the same four fitted baselines, finding the main conclusions of Table 2 unaffected (Appendix O). Table 3 ablates four architectural components. Every removal costs ordering; the seed-router loss matters most, degrading all three

Table 2: Overall results (mean ± 95% CI; $n = 450$ per setting). Healthy is the mean predicted stage of healthy control participants; true stage is 0. Runtime is minutes per dataset. The non-blind Healthy value of 0.000 holds by construction and is not comparable across methods; see Appendix M.

Dim	Algorithm	$\tau \downarrow$	ARI ↑	Stage MAE ↓	Healthy ↓	Runtime
12 bio	SUBTRACE	**0.162±0.005**	**0.288±0.017**	**0.935±0.015**	0.000	< 0.001
	SUBTRACE (Blind)	0.172±0.005	0.287±0.018	1.070±0.019	0.164	< 0.001
	BEBMS	0.215±0.006	0.270±0.019	1.233±0.030	0.143	2.7
	BEBMS (Blind)	0.235±0.007	0.258±0.019	1.308±0.034	0.354	3.2
	SuStaIn (GMM)	0.298±0.010	0.216±0.018	1.822±0.055	2.798	5.1
	SuStaIn (KDE)	0.351±0.007	0.065±0.006	2.314±0.047	3.504	4.9
	K-Means + TEMPO	0.202±0.006	0.193±0.014	0.992±0.028	0.000	< 0.001
	Random	0.427±0.004	0.000	4.445±0.028	5.982	N/A
100 bio	SUBTRACE	**0.119±0.004**	0.108±0.017	**4.114±0.071**	0.000	< 0.001
	SUBTRACE (Blind)	0.128±0.004	0.112±0.018	4.504±0.082	0.032	< 0.001
	BEBMS	0.378±0.005	0.094±0.020	9.111±0.369	0.118	94
	BEBMS (Blind)	0.381±0.004	0.089±0.019	9.429±0.394	0.881	115
	SuStaIn (GMM)	0.402±0.005	**0.169±0.025**	19.766±0.488	43.441	475
	SuStaIn (KDE)	0.366±0.005	0.037±0.009	15.042±0.377	23.452	473
	K-Means + TEMPO	0.213±0.002	0.054±0.014	5.906±0.158	0.381	< 0.001
	Random	0.479±0.001	0.000	36.341±0.214	50.122	N/A

metrics, while the profile-fit and second-pass ablations trade subtyping accuracy for lower stage error.

Table 3: Ablation study (lowdim Exp. 1; mean ± 95% CI, $n = 50$); see Appendix P.

Variant	$\tau \downarrow$	ARI ↑	Stage MAE ↓
Full model	**0.136±0.010**	**0.371±0.053**	0.926±0.026
No seed-router loss	0.189±0.015	0.297±0.052	0.981±0.033
One pass + profile fit	0.150±0.012	0.359±0.053	0.867±0.021
Seed-only (no profile-fit matching)	0.141±0.014	0.318±0.054	**0.820±0.021**
No rank features	0.141±0.012	0.353±0.054	0.980±0.028

Finally, cross-experiment transfer confirms robustness to simulator misspecification: every trained model remains competitive on ordering even when the test generator differs from the training generator; see Appendix R for the full matrices and discussion.

5 REAL WORLD EXPERIMENTS

5.1 ADNI

We apply the nine trained 12-bio SUBTRACE models to the Alzheimer's Disease Neuroimaging Initiative (ADNI) dataset (Mueller et al., 2005). The processed cohort contains 726 participants: 155 cognitively normal (CN), 182 early mild cognitive impairment (EMCI), 236 late mild cognitive impairment (LMCI), and 153 AD. We use the same 12 biomarkers, data preprocessing procedure, and biomarker distribution parameters as in BEBMS (Hao & Austerweil, 2025): amyloid (ABETA), tau (TAU, PTAU), cognitive assessments (ADAS13, RAVLT, MMSE), and structural MRI measures (Entorhinal, Hippocampus, MidTemp, Fusiform, WholeBrain, Ventricle). All 12 biomarkers are observed for all included participants, so no imputation is used.

ADNI has no ground-truth progression subtypes, orderings, or stages, so this is a qualitative analysis rather than a supervised benchmark. Each 12-bio model produces three subtype-specific orderings, subtype assignments, and stage estimates. Because subtype labels are arbitrary, we align the nine sets of subtypes using the Hungarian algorithm (Kuhn, 1955) to minimize total pairwise Kendall discordance between subtype-specific biomarker rank orderings. Displayed subtype labels are ordered by the average percentage of diseased participants assigned to the matched subtype.

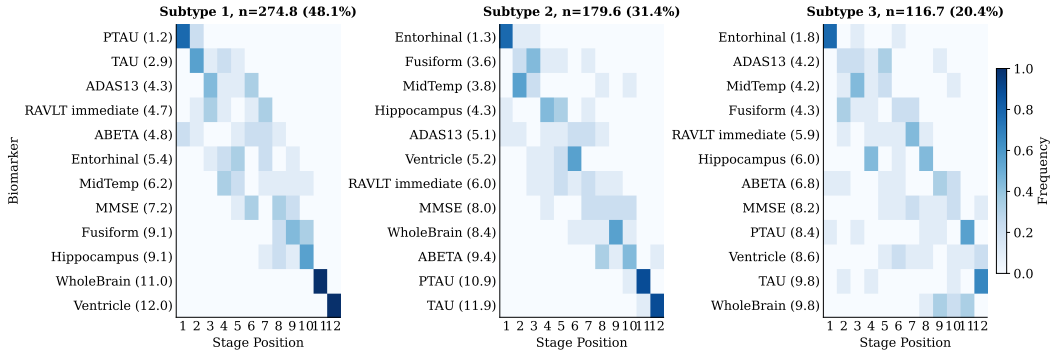


Figure 2: Consensus subtype orderings across nine SUBTRACE models on ADNI. Each cell shows how often a biomarker landed at that rank across the nine models; the number in parentheses next to each biomarker is its mean rank.

Figure 2 shows three consensus subtypes. Subtype 1 (48.1%) has CSF markers appearing early, followed by cognitive and structural changes, consistent with Typical AD (TAD; Murray et al., 2011). Subtype 2 (31.4%) has medial temporal lobe measures (Entorhinal, Fusiform, MidTemp) early, followed by global neurodegeneration, cognitive decline, and CSF markers. This is broadly compatible with Limbic-predominant AD (LPAD; Murray et al., 2011). Subtype 3 (20.4%) is more mixed, with temporal, cognitive, and CSF markers appearing before later global neurodegeneration; we interpret it cautiously as a possible variant of TAD or Hippocampal-sparing AD (HSAD; Murray et al., 2011).

The recovered orderings align with known biology. Entorhinal precedes Fusiform and MidTemp in all three subtypes, matching Braak staging (Braak & Braak, 1991) and the medial-to-lateral temporal cascade (Jack et al., 2010). MMSE is consistently late, consistent with its low sensitivity to early impairment (Tombaugh & McIntyre, 1992). Ventricle and WholeBrain also tend to be late among the three cognitive biomarkers, consistent with accelerating neurodegeneration in clinical AD (Jack Jr et al., 2009). ABETA does not always precede tau markers; this discrepancy from a simple amyloid-first cascade (Jack et al., 2010; Jack Jr et al., 2024) may reflect real cohort heterogeneity. BEBMS (Hao & Austerweil, 2025) reports a similar finding on the same data.

Table 7 shows a monotone stage pattern within every matched subtype: CN participants receive stage 0, EMCI has intermediate stages, LMCI is later, and AD is latest. Only the binary CN/non-CN disease label is used as input in these non-blind ADNI runs; the model does not receive the finer diagnostic categories EMCI, LMCI, or AD. Therefore, the monotone increase from CN to EMCI to LMCI to AD is not directly imposed by the input labels. Instead, it suggests that the continuous stages recovered by SUBTRACE track clinically meaningful disease severity. The baseline models show the same monotone pattern within their own predicted subtypes (Appendix J); SuStaIn (KDE) could not be fit on this cohort (Appendix J).

Figure 3 shows example continuous timelines from the Exp. 7 model, which are broadly consistent with the consensus subtype patterns. Events cluster earlier in the normalized timeline, suggesting faster progression early in the inferred disease course and slower progression later. This kind of spacing information is unavailable from ordinal event sequences alone and may be clinically useful when comparing early- and late-intervention windows.

5.2 NACC

We further apply SUBTRACE to the National Alzheimer's Coordinating Center (NACC) cohort (Morris et al., 2006) as an independent second real-world validation. The results are broadly consistent with those on ADNI and with established neuropathological findings; see Appendix K for details.

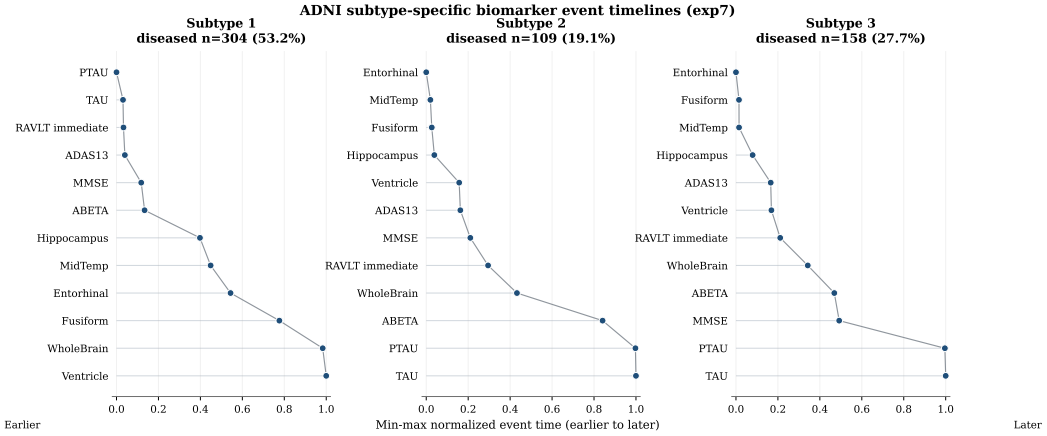


Figure 3: Example continuous ADNI timelines from the Exp. 7 12-bio SUBTRACE model. Unlike ordinal EBM outputs, SUBTRACE shows both event order and relative spacing between events.

6 DISCUSSION

We introduced SUBTRACE, a simulation-trained Transformer for amortized inference of disease progression subtypes from cross-sectional biomarker data. The subtype problem is difficult because its three outputs are tightly coupled: incorrect orderings make subtype assignments hard to interpret, and staging depends on both. SUBTRACE addresses this coupling by learning patient-biomarker abnormality evidence and fitting subtype timelines to that evidence at the predicted stage. This explicit abnormality-profile check is a major reason the advantage is largest in the 100-biomarker setting, where additional biomarkers provide more evidence instead of only enlarging the search space.

The results also clarify what remains hard. Subtype ARI is modest, especially at 100 biomarkers, and SuStaIn (GMM) sometimes obtains higher ARI. We do not view this as contradicting the main result: a high ARI can coexist with poor orderings, poor stages, and severe control-stage miscalibration. In disease progression modeling, subtype labels are useful only insofar as they correspond to plausible timelines and patient stages. The blind variant of SUBTRACE is encouraging: removing diagnosis from inference changes synthetic performance only modestly, so the method remains usable when diagnostic labels are uncertain or unavailable.

A natural concern is that SUBTRACE depends on the quality of its training simulator: if the simulator is far from the real data-generating process, the model may not transfer reliably. Our ADNI results give two reasons for cautious optimism. First, models trained under nine different synthetic assumptions recover broadly similar matched subtype orderings on ADNI, close to the detailed Exp. 7 timelines, suggesting robustness to moderate misspecification. Second, the dependence is also useful: each trained model acts as a proxy for its generative assumption, so applying models from different simulators to the same cohort compares which assumptions produce the most plausible and stable timelines.

A shared bottleneck across disease progression research is the absence of ground truth for real-world datasets, which rules out the quantitative evaluation we use in synthetic experiments; real-data analysis as in Section 5 is therefore necessarily qualitative.

Another limitation is the irreversibility assumption inherited from EBMs: once a biomarker event has occurred, the latent event state is treated as pathological. This concerns the cohort-level event state, not a claim that every patient's measured value moves monotonically or never returns toward the normal range; SUBTRACE treats partial reversions and post-event noise as variability within the pathological state. The simplification remains workable when the cohort still shows a detectable distributional difference between benign and pathological states. Finally, the number of subtypes is unknown in real data, which is also messier: often longitudinal and containing missing values. These require fundamentally different models, so we leave them to future work.

AI USE STATEMENT

In this work, we used generative AI tools to help develop theoretical models or conceptual frameworks; design or provide feedback on research methodology or experiments; implement methods; clean and reformat datasets; create or modify scientific figures or images; suggest experimental parameters; create or edit software code; draft parts of this paper; and propose a title or keywords for this paper. We did not use generative AI tools to assist with translation, generate synthetic datasets, provide critical ingredients for proving mathematical claims, formulate mathematical claims, assist in the writing of proofs, propose or refine hypotheses, support qualitative or thematic data analysis, interpret results, formulate questions for surveys or interviews, summarize or analyze existing literature, discover research topics or identify gaps, identify relevant literature, or format references. All authors read and finalized the draft; the data-generation code was manually verified, and all experimental results were manually checked. We take responsibility for the final content of this work, including text, claims, or artifacts produced with the aid of generative AI.

ETHICS STATEMENT

The authors obtained institutional IRB approval for this work, and obtained permission to use the ADNI and NACC datasets.

REPRODUCIBILITY STATEMENT

We have listed the code URL in the abstract, which contains clear instructions for reproducing our results. Note that due to randomness, some results may change from run to run.

REFERENCES

Leon M Aksman, Peter A Wijeratne, Neil P Oxtoby, Arman Eshaghi, Cameron Shand, Andre Altmann, Daniel C Alexander, and Alexandra L Young. pysustain: a python implementation of the subtype and stage inference algorithm. *SoftwareX*, 16:100811, 2021.

Heiko Braak and Eva Braak. Neuropathological stageing of alzheimer-related changes. *Acta Neuropathologica*, 82(4):239–259, 1991.

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.

Mar Estarellas, Neil P Oxtoby, Jonathan M Schott, Daniel C Alexander, and Alexandra L Young. Multimodal subtypes identified in alzheimer's disease neuroimaging initiative participants by missing-data-enabled subtype and stage inference. *Brain Communications*, 6(4):fcae219, 2024.

Daniel Ferreira, Agneta Nordberg, and Eric Westman. Biological subtypes of alzheimer disease: A systematic review and meta-analysis. *Neurology*, 94(10):436–448, 2020.

Nicholas C Firth, Silvia Primativo, Emilie Brotherhood, Alexandra L Young, Keir XX Yong, Sebastian J Crutch, Daniel C Alexander, and Neil P Oxtoby. Sequences of cognitive decline in typical alzheimer's disease and posterior cortical atrophy estimated using a novel event-based model of disease progression. *Alzheimer's & Dementia*, 16(7):965–973, 2020.

Hubert M Fonteijn, Marc Modat, Matthew J Clarkson, Josephine Barnes, Manja Lehmann, Nicola Z Hobbs, Rachael I Scahill, Sarah J Tabrizi, Sebastien Ourselin, Nick C Fox, et al. An event-based model for disease progression and its application in familial alzheimer's disease and huntington's disease. *NeuroImage*, 60(3):1880–1889, 2012.

Ellen Grober, Charles B Hall, Richard B Lipton, Alan B Zonderman, Susan M Resnick, and Claudia Kawas. Memory impairment, executive dysfunction, and intellectual decline in preclinical alzheimer's disease. *Journal of the International Neuropsychological Society*, 14(2):266–278, 2008.

Hongtao Hao and Joseph L Austerweil. Bayesian event-based model for disease subtype and stage inference. In *Proceedings of the 5th Machine Learning for Health Symposium*, volume 297. Proceedings of Machine Learning Research, 2025.

Hongtao Hao and Joseph L. Austerweil. Tempo: Transformers for temporal disease progression from cross-sectional data. In *Proceedings of the 7th Annual Conference on Health, Inference, and Learning (CHIL)*, 2026.

Hongtao Hao, Vivek Prabhakaran, Veena A Nair, Nagesh Adluru, and Joseph L. Austerweil. Stage-aware event-based modeling (SA-EBM) for disease progression. In *Proceedings of the 10th Machine Learning for Healthcare Conference*, volume 298 of *Proceedings of Machine Learning Research*. PMLR, 2025.

Clifford R Jack and David M Holtzman. Biomarker modeling of alzheimer's disease. *Neuron*, 80 (6):1347–1358, 2013.

Clifford R Jack, David S Knopman, William J Jagust, Leslie M Shaw, Paul S Aisen, Michael W Weiner, Ronald C Petersen, and John Q Trojanowski. Hypothetical model of dynamic biomarkers of the alzheimer's pathological cascade. *The Lancet Neurology*, 9(1):119–128, 2010.

Clifford R Jack Jr, Val J Lowe, Stephen D Weigand, Heather J Wiste, Matthew L Senjem, David S Knopman, Maria M Shiung, Jeffrey L Gunter, Bradley F Boeve, Bradley J Kemp, et al. Serial pib and mri in normal, mild cognitive impairment and alzheimer's disease: implications for sequence of pathological events in alzheimer's disease. *Brain*, 132(5):1355–1365, 2009.

Clifford R Jack Jr, J Scott Andrews, Thomas G Beach, Teresa Buracchio, Billy Dunn, Ana Graf, Oskar Hansson, Carole Ho, William Jagust, Eric McDade, et al. Revised criteria for diagnosis and staging of alzheimer's disease: Alzheimer's association workgroup. *Alzheimer's & Dementia*, 20(8):5143–5169, 2024.

Harold W Kuhn. The hungarian method for the assignment problem. *Naval Research Logistics Quarterly*, 2(1-2):83–97, 1955.

Colin L Mallows. Non-null ranking models. i. *Biometrika*, 44(1/2):114–130, 1957.

Sophie E Mastenbroek, Jacob W Vogel, Lyduine E Collij, Geidy E Serrano, Cécilia Tremblay, Alexandra L Young, Richard A Arce, Holly A Shill, Erika D Driver-Dunckley, Shyamal H Mehta, et al. Disease progression modelling reveals heterogeneity in trajectories of lewy-type α-synuclein pathology. *Nature Communications*, 15(1):5133, 2024.

John C Morris, Sandra Weintraub, Helena C Chui, Jeffrey Cummings, Charles DeCarli, Steven Ferris, Norman L Foster, Douglas Galasko, Neill Graff-Radford, Elaine R Peskind, et al. The uniform data set (uds): clinical and cognitive variables and descriptive data from alzheimer disease centers. *Alzheimer Disease & Associated Disorders*, 20(4):210–216, 2006.

Susanne G Mueller, Michael W Weiner, Leon J Thal, Ronald C Petersen, Clifford Jack, William Jagust, John Q Trojanowski, Arthur W Toga, and Laurel Beckett. The alzheimer's disease neuroimaging initiative. *Neuroimaging Clinics*, 15(4):869–877, 2005.

Shubhabrata Mukherjee, Seo-Eun Choi, Michael L Lee, Phoebe Scollard, Emily H Trittschuh, Jesse Mez, Andrew J Saykin, Laura E Gibbons, R Elizabeth Sanders, Andrew F Zaman, et al. Cognitive domain harmonization and cocalibration in studies of older adults. *Neuropsychology*, 37(4):409, 2023.

Melissa E Murray, Neill R Graff-Radford, Owen A Ross, Ronald C Petersen, Ranjan Duara, and Dennis W Dickson. Neuropathologically defined subtypes of alzheimer's disease with distinct clinical characteristics: a retrospective study. *The Lancet Neurology*, 10(9):785–796, 2011.

Gemma Salvadó, Kanta Horie, Nicolas R Barthélemy, Jacob W Vogel, Alexa Pichet Binette, Charles D Chen, Andrew J Aschenbrenner, Brian A Gordon, Tammie LS Benzinger, David M Holtzman, et al. Disease staging of alzheimer's disease using a csf-based biomarker model. *Nature Aging*, 4(5):694–708, 2024.

Raghav Tandon, James J Lah, and Cassie S Mitchell. s-sustain: Scaling subtype and stage inference via simultaneous clustering of subjects and biomarkers. *Proceedings of Machine Learning Research*, 248:461, 2024.

Tom N Tombaugh and Nancy J McIntyre. The mini-mental state examination: a comprehensive review. *Journal of the American Geriatrics Society*, 40(9):922–935, 1992.

Vikram Venkatraghavan, Esther E Bron, Wiro J Niessen, Stefan Klein, Alzheimer's Disease Neuroimaging Initiative, et al. Disease progression timeline estimation for alzheimer's disease using discriminative event based modeling. *NeuroImage*, 186:518–532, 2019.

Victor L Villemagne, Samantha Burnham, Pierrick Bourgeat, Belinda Brown, Kathryn A Ellis, Olivier Salvado, Cassandra Szoeke, S Lance Macaulay, Ralph Martins, Paul Maruff, et al. Amyloid β deposition, neurodegeneration, and cognitive decline in sporadic alzheimer's disease: a prospective cohort study. *The Lancet Neurology*, 12(4):357–367, 2013.

Jacob W Vogel, Alexandra L Young, Neil P Oxtoby, Ruben Smith, Rik Ossenkoppele, Olof T Strandberg, Renaud La Joie, Leon M Aksman, Michel J Grothe, Yasser Iturria-Medina, et al. Four distinct trajectories of tau deposition identified in alzheimer's disease. *Nature Medicine*, 27(5): 871–881, 2021.

Alexandra L Young, Neil P Oxtoby, Sebastien Ourselin, Jonathan M Schott, Daniel C Alexander, Alzheimer's Disease Neuroimaging Initiative, et al. A simulation system for biomarker evolution in neurodegenerative disease. *Medical Image Analysis*, 26(1):47–56, 2015.

Alexandra L Young, Razvan V Marinescu, Neil P Oxtoby, Martina Bocchetta, Keir Yong, Nicholas C Firth, David M Cash, David L Thomas, Katrina M Dick, Jorge Cardoso, et al. Uncovering the heterogeneity and temporal complexity of neurodegenerative diseases with subtype and stage inference. *Nature Communications*, 9(1):4273, 2018.

Alexandra L Young, Neil P Oxtoby, Sara Garbarino, Nick C Fox, Frederik Barkhof, Jonathan M Schott, and Daniel C Alexander. Data-driven modelling of neurodegenerative disease progression: thinking outside the black box. *Nature Reviews Neuroscience*, 25(2):111–130, 2024.

A NOTATION

Table 4 collects the symbols used across Sections 2–3 and this appendix. Superscript $*$ marks a simulator ground-truth quantity, a hat marks a model prediction, and (0) marks a provisional (first-pass) quantity later replaced by a final one.

Table 4: Notation.

Problem setup (Section 2, classical EBM)	
J, B, K	Number of participants, biomarkers, subtypes
$x_{j,b}$	Measurement, participant j, biomarker b (standardized from Section 3 on)
d_j	Diagnosis label (0 control, 1 diseased)
ϕ_b, θ_b	Biomarker b's benign / pathological distribution
z_j	True integer stage (number of events occurred)
$S = (S_1, \ldots, S_B)$	Ordering; S_i is the biomarker at rank i
π_k, $\pi_{z_j \mid k}$	Subtype prior, stage prior given subtype k
SUBTRACE *(Section 3 and Appendix B)*	
k_j	True subtype label
$\xi_{k,b}$, $r^*_{k,b} = \xi_{k,b}/B$	True event time (unnormalized, normalized)
$\tilde{z}_j = z_j/B$	Normalized true stage
$\mathbf{S}^{(k)}$	True ordinal ordering for subtype k
$A_{j,b}$, $A^*_{j,b}$	True abnormality indicator (Eq. 3)
$\mathbf{u}_{j,b}$, $\mathbf{h}_{j,b}$	Raw token features, token embedding ($\mathbf{h} = f_\psi(\mathbf{u})$)
$R^{\text{within}}_{j,b}$, $R^{\text{cohort}}_{j,b}$	Within-patient rank, cohort rank (token features)
$\mathbf{g}_j$	Seed router's contextual patient embedding
$\mathbf{w}_k$, a_k	Seed router's linear weight, bias (Eq. 5)
$q^{(0)}_{j,k}$, $q_{j,k}$	Seed (provisional), final subtype probability
ω_j	Pooling weight (downweights controls / weak evidence)
$\hat{r}^{(0)}_{k,b}$, $\hat{r}_{k,b}$	Provisional, final event position
a_{ord}, a_{ev}	Ordering, evidence scalar heads
T_{ord}, T_{ev}	Ordering, evidence biomarker Transformers
$e_{j,b}$, $p_{j,b}$	Evidence logit, evidence probability
$\hat{z}_j$, g_z	Predicted stage, stage-head function
$q^{\text{fit},0}_{j,k}$	Provisional profile-fit subtype probability
α	Profile-fit transition sharpness
w_b	Subtype-informative biomarker weight (profile fit)
ζ, ζ^*	Hungarian-matching permutation (Eq. 10)
η	Weight on subtype cost in Hungarian matching
$\lambda_{\text{ord}}, \lambda_{\text{sub}}, \lambda_{\text{stage}}, \lambda_{\text{evid}}$	Main loss weights (Eq. 11)
c, γ	Clipping constant, exponent in ω_j
κ	Mallows-Kendall dispersion (data generation only)

A few symbols still do double duty and are disambiguated by context (bold vs. plain, or the section they appear in): $\mathbf{w}_k$ (router weight vector) vs. w_b (biomarker weight, unrelated); a_k (router bias) vs. $a_{\text{ord}}, a_{\text{ev}}$ (scalar heads), all playing the same general role of an affine read-out bias; $\mathbf{S}^{(k)}$ (true ordering) vs. $s^{(0)}_{k,b}$ (an ordering score, lowercase); $\mathbf{h}_{j,b}$ (token embedding) vs. the unrelated healthy-baseline draw h_j used only in the Sigmoid simulator description (Appendix E); and d_j (diagnosis) vs. the unrelated embedding dimension d (e.g. $\mathbb{R}^d$).

B Additional Training and Evaluation Details

This appendix records the implementation-level math omitted from the main Methods section. Notation follows Section 3: k_j is the true subtype label, z_j is the true integer stage, $\tilde{z}_j = z_j/B$ is the normalized true stage, $r^*_{k,b}$ is the true normalized event position, and $\hat{z}_j$ and $\hat{r}_{k,b}$ are model predictions.

Profile fit, refinement, and final outputs (full equations for Section 3). The model asks how well each provisional subtype timeline explains each patient's evidence profile at the predicted stage. With transition sharpness α, the provisional subtype logit is the negative mismatch between the structural profile $\sigma(\alpha(\hat{z}_j - \hat{r}^{(0)}_{k,b}))$ and the learned evidence profile:

$$u^{(0)}_{j,k} = -\frac{1}{\sqrt{B}} \sum_{b=1}^{B} w_b \, \mathrm{BCELogit}\left(\alpha(\hat{z}_j - \hat{r}^{(0)}_{k,b}),\ p_{j,b}\right), \qquad q^{\mathrm{fit},0}_{j,k} = \mathrm{softmax}_k(u^{(0)}_{j,k}).$$

Here w_b upweights biomarkers whose predicted event positions differ across subtype slots; the weights are normalized to have mean near one. $q^{\mathrm{fit},0}_{j,k}$ is not an additional learned head. It is the provisional subtype probability implied by patient j's evidence profile, predicted stage, and provisional timeline.

The model blends the seed-router probabilities with these provisional profile-fit probabilities:

$$q^{\mathrm{pool}}_{j,k} = \frac{(1-\beta)q^{(0)}_{j,k} + \beta q^{\mathrm{fit},0}_{j,k}}{\sum_{\ell=1}^{K}[(1-\beta)q^{(0)}_{j,\ell} + \beta q^{\mathrm{fit},0}_{j,\ell}]}.$$

The blended weights are used for a second, final timeline pass:

$$\mathbf{t}_{k,b} = \sum_{j=1}^{J} \bar{q}^{\mathrm{pool}}_{j,k} \mathbf{h}_{j,b}, \qquad \bar{q}^{\mathrm{pool}}_{j,k} = \frac{q^{\mathrm{pool}}_{j,k}\omega_j}{\sum_i q^{\mathrm{pool}}_{i,k}\omega_i + \varepsilon}.$$

The pooled sequence is passed through the same biomarker Transformer and ranking head:

$$s_{k,b} = a_{\mathrm{ord}}\left[T_{\mathrm{ord}}(\mathbf{t}_{k,1}, \ldots, \mathbf{t}_{k,B})_b\right], \qquad \hat{r}_{k,b} = \sigma(s_{k,b}).$$

Sorting $\hat{r}_{k,1}, \ldots, \hat{r}_{k,B}$ gives the final ordinal ordering; the values preserve relative spacing between events. Finally, SUBTRACE recomputes the profile fit using the final timeline:

$$u_{j,k} = -\frac{1}{\sqrt{B}} \sum_{b=1}^{B} w_b \, \mathrm{BCELogit}\left(\alpha(\hat{z}_j - \hat{r}_{k,b}),\ p_{j,b}\right), \qquad q_{j,k} = \mathrm{softmax}_k(u_{j,k}). \tag{12}$$

The final outputs are the continuous event positions $\hat{r}_{k,b}$, subtype probabilities $q_{j,k}$, and stage $\hat{z}_j$. All refinement terms are recomputed for the current cohort inside one forward pass; no profile-fit weights are stored or reused across runs.

Pooling weights. The ordering branch downweights controls and weakly abnormal patients using

$$\omega_j = d_j \left[\frac{1}{B}\sum_{b=1}^{B} \frac{\min(|x_{j,b}|, c)}{c}\right]^{\gamma}.$$

The normalized subtype pooling weights are then

$$\bar{q}_{j,k} = \frac{q_{j,k}\omega_j}{\sum_i q_{i,k}\omega_i + \varepsilon}.$$

Balanced evidence loss. The evidence head is trained against $A^*_{j,b} = d_j \mathbf{1}\{z_j \geq \xi_{k_j,b}\} = d_j \mathbf{1}\{\tilde{z}_j \geq r^*_{k_j,b}\}$ using balanced binary cross-entropy:

$$\mathcal{L}_{\mathrm{evid}} = \frac{1}{2}\frac{\sum_{j,b} A^*_{j,b}\,\mathrm{BCELogit}(e_{j,b}, 1)}{\sum_{j,b} A^*_{j,b} + \varepsilon} + \frac{1}{2}\frac{\sum_{j,b}(1 - A^*_{j,b})\,\mathrm{BCELogit}(e_{j,b}, 0)}{\sum_{j,b}(1 - A^*_{j,b}) + \varepsilon}.$$

Subtype-informative biomarker weights. The profile-fit weights in Eq. 12 are derived from the spread of ordering scores across subtype slots:

$$\delta_b = \mathrm{sd}_k(s_{k,b}), \qquad w_b \propto \lambda_0 + (1-\lambda_0)\frac{\delta_b}{B^{-1}\sum_{b'}\delta_{b'} + \varepsilon},$$

with clipping and renormalization so that the average weight is approximately one.

Ordering cost. For true subtype k and predicted slot m, the direct event-position cost is

$$C^{\mathrm{mse}}_{k,m} = \frac{1}{B}\sum_{b=1}^{B}(\hat{r}_{m,b} - r^*_{k,b})^2.$$

The pairwise term compares relative event times. For biomarker pairs

$$\mathcal{P}_k = \{(a,b) : r^*_{k,a} \neq r^*_{k,b}\}, \qquad Y_{k,a,b} = \mathbf{1}\{r^*_{k,a} > r^*_{k,b}\},$$

we use

$$C^{\mathrm{pair}}_{k,m} = \frac{1}{|\mathcal{P}_k|}\sum_{(a,b)\in\mathcal{P}_k}\mathrm{BCELogit}(s_{m,a} - s_{m,b}, Y_{k,a,b}).$$

The ordering slot cost is

$$C^{\mathrm{ord}}_{k,m} = (1-\lambda_{\mathrm{pair}})C^{\mathrm{mse}}_{k,m} + \lambda_{\mathrm{pair}}C^{\mathrm{pair}}_{k,m}.$$

For high-dimensional settings, a fixed number of pairs is sampled rather than enumerating all $B(B-1)$ ordered pairs.

Subtype and seed-router costs. Let M_j be the subtype-training mask, which excludes controls and may exclude very early diseased patients. The subtype slot cost is

$$C^{\mathrm{sub}}_{k,m} = \frac{\sum_j M_j \mathbf{1}\{k_j = k\}[-\log q_{j,m}]}{\sum_j M_j + \varepsilon}.$$

The seed-router cost $C^{\mathrm{init}}_{k,m}$ is the same expression with $q^{(0)}_{j,m}$ in place of $q_{j,m}$.

Control-stage penalty. Controls are encouraged to remain below the first stage-rounding boundary:

$$\mathcal{L}_{\mathrm{ctrl}} = \frac{1}{|\{j : d_j = 0\}|}\sum_{j:d_j=0}\max\left(0, \hat{z}_j - \frac{1}{2B}\right),$$

weighted by $\lambda_{\mathrm{ctrl}} = 0.05$ in the total loss.

Evidence-side subtype contrast. For candidate true subtype k, define

$$A^{(k)}_{j,b} = d_j\mathbf{1}\{z_j \geq \xi_{k,b}\} = d_j\mathbf{1}\{\tilde{z}_j \geq r^*_{k,b}\}.$$

The evidence log-likelihood under candidate subtype k is

$$\ell^{\mathrm{true}}_{j,k} = \frac{1}{\sqrt{B}}\sum_b\left[A^{(k)}_{j,b}\log\sigma(e_{j,b}) + (1 - A^{(k)}_{j,b})\log\sigma(-e_{j,b})\right],$$

and the contrastive term is

$$\mathcal{L}_{\mathrm{true}} = -\frac{1}{|\{j : d_j = 1\}|}\sum_{j:d_j=1}\log\frac{\exp(\ell^{\mathrm{true}}_{j,k_j})}{\sum_k\exp(\ell^{\mathrm{true}}_{j,k})}.$$

The total training loss instantiates Eq. 11 with matched ordering, matched subtype, stage, evidence, seed-router, control-stage, and evidence-contrast terms. Optional ablation terms, such as pairwise same-subtype supervision or profile-consistency regularization, are removed by setting their coefficients to zero.

C TRAINING DETAILS

Architecture. SUBTRACE uses $d = 256$ for all hidden dimensions, 4-layer Transformer encoders with 8 attention heads, dropout $p = 0.1$, learned biomarker identity embeddings, and learned positional embeddings where used by the Transformer branches. The seed router uses a 2-layer patient MLP followed by a cohort Transformer. The patient–biomarker token MLP (f_ψ) maps the token features in Eq. 4 to $\mathbb{R}^d$. The ordering and evidence branches each use a biomarker Transformer with scalar heads, and the stage head pools biomarker-level features with attention. Evidence-weighting constants: $c = 3.0$, $\gamma = 1.0$.

Training. Models are trained with Adam, learning rate 10^{-4}, batch size 24. Low-dim models train for 30 epochs with patience 8 (early stopping on validation τ). High-dim models train for 30 epochs with patience 8. A warmup period of 4 epochs (low-dim) or 6 epochs (high-dim) suppresses $\mathcal{L}_{\text{sub}}$ for early-stage patients.

Loss weights. $\lambda_{\text{ord}} = 1.0$, $\lambda_{\text{sub}} = 0.5$, $\lambda_{\text{init}} = 0.15$, $\lambda_{\text{stage}} = 0.3$, $\lambda_{\text{evid}} = 0.35$, $\lambda_{\text{true}} = 0.25$. Optional profile-consistency and pairwise same-subtype terms are set to zero. Pairwise-to-direct ordering blend $\lambda_{\text{pair}} = 0.5$. Refinement blend $\beta = 0.5$.

Input standardization. Measurements are standardized using training-set statistics (mean and standard deviation per biomarker), then clipped to $[-6, 6]$.

D DESIGN CHOICE OF SYNTHETIC DATA GENERATION

Table 5: Nine synthetic experiments. DM = Dirichlet-Multinomial stage prior.

Exp	Name	Stage	Events	Measurement
1	Ordinal-DM, Normal	Ordinal (DM)	Discrete	EBM / Normal
2	Ordinal-DM, NonNormal	Ordinal (DM)	Discrete	EBM / Non-Normal
3	Ordinal-Unif, Normal	Ordinal (Uniform)	Discrete	EBM / Normal
4	Ordinal-Unif, NonNormal	Ordinal (Uniform)	Discrete	EBM / Non-Normal
5	ContBeta, Sigmoid	Continuous (Beta)	Discrete	Sigmoid
6	ContBeta, Normal	Continuous (Beta)	Discrete	EBM / Normal
7	ContBeta, NonNormal	Continuous (Beta)	Discrete	EBM / Non-Normal
8	ContBeta+Noise, Sigmoid	Continuous (Beta)	Continuous	Sigmoid
9	ContBeta+Noise, Normal	Continuous (Beta)	Continuous	EBM / Normal

Motivation: real cohorts (ADNI, NACC) have no ground truth, so training and evaluation require synthetic data. We keep the simulator close to real data by estimating parameters from ADNI, using the same 12-biomarker panel adopted in previous studies.

Even so, real data may differ from any single simulator, so we span a range of configurations:

- **The number of biomarkers:** we use both 12 and 100. 12 matches the ADNI panel used in previous studies (Hao et al., 2025; Hao & Austerweil, 2025; 2026). 100 balances biological reality against computation, and is the default in BEBMS.
- **Generative framework (EBM vs. Sigmoid):** The two frameworks differ in how a biomarker's measurements change after its event (Section 4): under EBM, biomarker b's value is drawn from a benign or pathological distribution depending on whether stage z_j has passed b's event position; under Sigmoid, we follow Hao & Austerweil (2026) and let the value move gradually away from the benign state as stage advances past that position, with healthy participants' values following normal distributions as in EBM.
- **Biomarker measurements:** Sigmoid is as above. For EBM we use either a normal distribution (canonical) or non-normal (more realistic), borrowing non-normal parameters from Hao & Austerweil (2026). Each biomarker randomly picks one of six non-normal configurations (exact mixture composition in Appendix E), shared across subtypes, though measurement samples are generated independently.

- **Event times:** The exact times that a biomarker transitions from being benign to pathological can be modeled as either discrete (i.e., $1, \ldots, B$ where B is the number of biomarkers) or continuous. We sample continuous event times from $\mathrm{Beta}(2, 2) \cdot B$ (symmetric bell shape) with added noise sampled from $\mathcal{N}\big(0, (0.05 \cdot B)^2\big)$ in the noisy-event-time experiments (Exp. 8–9, regardless of measurement model), and the final event times are clipped to $[0, B]$.
- **Number of events**: by definition, it is equivalent to the number of biomarkers.
- **Disease stages:** Healthy participants have $z_j = 0$. For diseased participants, we can either model their disease stages as ordinal or continuous. We can obtain the ordinal disease stages from either a uniform distribution or Dirichlet-multinomial distribution (DM). For continuous disease stages, we draw disease stages from $\mathrm{Beta}(5, 2)$ which is skewed toward later stages. This is because empirically, 68.1% of the diseased participants in ADNI belong to later stages (Late MCI and diagnosed AD). In the continuous-stage experiments, this Beta draw is a latent severity value; the ordinal ground truth z_j used for evaluation is still the number of events it has passed, obtained by counting how many sorted event times are at or below it, so every experiment has a well-defined integer z_j regardless of whether the underlying stage prior is discrete or continuous.
- **Number of stages**: It is equal to the number of participants $J = 726$, which is the number of the post-processed ADNI dataset.
- **Number of subtypes:** We chose $K = 3$ because that is the reported optimal number of subtypes for the ADNI dataset using the state-of-the-art algorithm, i.e., (Hao & Austerweil, 2025). The result was backed by prior literature on subtypes of Alzheimer's disease (Hao & Austerweil, 2025).
- **Subtype similarity:** We first sample a random consensus ordinal event sequence, s_0. Then we draw a Mallows-Kendall (MK) dispersion parameter uniformly from $[0.01, 0.5]$ to generate subtype sequences from a MK model centered on s_0. At 12 biomarkers this range gives sequences with a variety of agreements covering close to the whole range of $[0, 1]$ as measured by Kendall's W; at 100 biomarkers the same range does not, since the same dispersion produces a different effective similarity at different B (Appendix E).

E Data generation details

Subtype orderings. For each simulated cohort, K subtype orderings are drawn from the Mallows–Kendall distribution (Mallows, 1957) with concentration parameter $\kappa \sim \mathrm{Uniform}[0.01, 0.50]$ controlling inter-subtype similarity.

The same κ range does not produce comparable subtype similarity at both dimensionalities. Across the 450 test datasets, Kendall's W among the K true subtype orderings ranges from 0.085 to 0.911 (mean 0.512) at 12 biomarkers, close to the full $[0, 1]$ range, with 12.7% of datasets at $W \leq 0.3$; at 100 biomarkers it ranges from 0.344 to 0.993 (mean 0.893), and no dataset falls at or below $W = 0.3$. Equivalently, the mean pairwise Kendall distance between the K true orderings is 0.400 ± 0.099 at 12 biomarkers versus 0.146 ± 0.113 at 100 (mean $\pm$ SD over the 450 datasets): subtype orderings are substantially more similar to each other at 100 biomarkers, and the 100-biomarker test set contains none of the harder, low-similarity cases present at 12 biomarkers. This follows mechanically from drawing κ from the same range regardless of B: the per-step Mallows sampling procedure produces an absolute ordering displacement that grows linearly in B, while the normalizer $\binom{B}{2}$ used to compute both the distance and W grows quadratically, so a fixed κ range yields a narrower, higher-W (more similar) distribution as B grows. Section 4 discusses the resulting implication for the ordering-scaling comparison.

Subtype sizes. Among diseased participants, subtype sizes are drawn from a Dirichlet-Multinomial with concentration parameter selected per dataset from $\{2, 5, 20\}$ (each subtype guaranteed at least one participant before the remainder is allocated), resampling until every subtype has at least 10 participants. This minimum-size rejection rule makes concentration 2 slightly underrepresented relative to 5 and 20 in the realized datasets.

Patient stages. For ordinal experiments (Exp. 1–4), diseased-patient stages $z_j \in \{1, \dots, B\}$ are drawn from a Dirichlet-Multinomial (Exp. 1–2) or Uniform (Exp. 3–4) distribution. For continuous experiments (Exp. 5–9), stages are drawn from a Beta$(5, 2)$ distribution scaled to $[0, B]$. Healthy participants: stage 0.

Biomarker measurements. Under the EBM measurement model, biomarker b for patient j is drawn from $\mathcal{N}(\theta_{b,\mu}, \theta^2_{b,\sigma})$ if b has occurred for j, and from $\mathcal{N}(\phi_{b,\mu}, \phi^2_{b,\sigma})$ otherwise.

Under the Sigmoid model, a diseased patient's measurement is $h_j + \dfrac{\theta_{b,\mu} - \phi_{b,\mu}}{1 + \exp(-\rho_b(z_j - s_b))}$, where $h_j \sim \mathcal{N}(\phi_{b,\mu}, \phi^2_{b,\sigma})$ is the same healthy-baseline draw used for non-diseased patients, s_b is b's event time, and $\rho_b = \max\left(1, \dfrac{|\theta_{b,\mu} - \phi_{b,\mu}|}{\sqrt{\theta^2_{b,\sigma} + \phi^2_{b,\sigma}}}\right)$. Progression direction is fixed rather than randomly flipped per biomarker: it is already encoded by the sign of $\theta_{b,\mu} - \phi_{b,\mu}$, since a random flip would let the model use direction as a shortcut to infer subtype identity.

Non-Normal biomarkers are drawn from one of six mixture families, randomly assigned per biomarker per dataset (Table 6).

Table 6: Non-Normal biomarker distribution families. Each biomarker/dataset independently picks one family. Families 0–3 mix their three listed components in equal (1/3 each) proportions; Family 5's split is shown in the table. All families additionally receive $\mathcal{N}(0, (0.2\sigma)^2)$ noise and are clipped to $[\mu - 5\sigma, \mu + 5\sigma]$ (omitted below for brevity).

Family	Composition
0	Triangular + Normal (shifted) + Exponential
1	Pareto + Uniform + Logistic
2	Beta (spread) + signed Exponential + Normal (spiked)
3	Gamma + Weibull + Normal ($\pm\sigma$ spike)
4	Heavy-tailed Cauchy, clipped
5	Bimodal: Normal spike (10%) + Logistic (90%)

F CLOUD RE-RUN OF EXPERIMENT 9

Experiment 9's data generator was corrected after the original run to include the intended noisy continuous event times (the same mechanism already used in Experiment 8). Because this changed the test datasets themselves, BEBMS, BEBMS (Blind), and both SuStaIn variants had to be refit on the corrected Experiment 9 datasets so that every method, including SUBTRACE, is evaluated on the same corrected data. Experiments 1–8 were unaffected and were not rerun.

The original baseline fitting (Experiments 1–8, both dimensions) ran on a CPU cluster at the Center for High Throughput Computing (CHTC). The corrected Experiment 9 refit could not use the same cluster allocation and was instead run on commercial cloud CPU instances: Google Cloud Platform (GCP) for low-dimensional Experiment 9 and the initial share of high-dimensional Experiment 9, with the remaining high-dimensional Experiment 9 jobs completed on Oracle Cloud Infrastructure (OCI) after the GCP trial credit was exhausted. The GCP and OCI runs used the same algorithm implementations and inference settings as the original CHTC run; only the compute platform changed.

The runtimes reported in Table 2 and in the main text are the original CHTC per-dataset runtimes (Experiments 1–8), not the GCP/OCI runtimes for the corrected Experiment 9. The cloud instances were slower per dataset than the CHTC cluster, so if anything, using the cloud runtimes instead would have made the baselines look slower and SUBTRACE's relative speed advantage larger. Reporting the original, faster CHTC runtimes is therefore the conservative choice and does not favor SUBTRACE in this comparison.

G WHY TEMPO DOES NOT EXTEND TO SUBTYPES BY ADDING A HEAD

TEMPO (Hao & Austerweil, 2026) addressed the single-subtype case: rather than fitting a model to each new cohort via MCMC, it trains a Transformer on thousands of synthetic datasets generated with known ground truth (a *simulation-based supervised learning* approach). By treating the ordering and staging tasks as regressions, i.e., assigning scalars to each biomarker (as position in the timeline) and each participant (as the disease stage), it recovers continuous event timelines instead of ordinal sequences and generalizes to high-dimensional biomarker sets with only modest degradation. It does not extend to subtypes by simply adding a subtype-classification head: since ordering, staging, and subtyping are mutually dependent, training three independent heads on the same input information discards the coupling between them, and each head must implicitly recover from noisy versions of the others' errors.

K-Means + TEMPO (Appendix H) is evidence for this: subtyping is decided by K-means independently of ordering and staging, which TEMPO then fits per cluster with no feedback into the clustering. This decoupled approach clearly underperforms SUBTRACE (Table 2: ARI 0.193 vs. 0.288 at 12 biomarkers, 0.054 vs. 0.108 at 100), the same failure mode as training independent heads: subtyping decided without coupling to ordering and staging, just realized through clustering plus a per-cluster fit rather than three heads on one network.

We also do not treat a literal multi-head extension of TEMPO as a well-specified baseline in its own right: TEMPO's ordering output is per cohort, not per patient (one B-dimensional score vector shared by everyone, since there is only one subtype), so giving it K subtypes would also require changing the ordering head to emit K orderings instead of one, which "add a classification head" does not by itself specify.

H K-MEANS + TEMPO

First, we used K-means to cluster each SUBTRACE test dataset into $K = 3$ groups. We used K-means rather than a classifier, such as a neural-network classifier, because subtype labels are arbitrary across datasets. That is, for each SUBTRACE dataset, the three subtype orderings are randomly selected from all $12!$ (for low dimensional) or $100!$ (for high dimensional) possible orderings. Therefore, the subtype labeled 1 in one dataset is almost guaranteed to be completely different from subtype 1 in another dataset.

We then combined each cluster with the control participants and applied TEMPO to infer its biomarker ordering.

Third, we computed staging performance using MAE.

Finally, we used the Hungarian algorithm to find the best matching between the predicted and ground-truth subtypes. We then computed ordering performance using average Kendall's τ and subtyping performance using ARI.

I ADNI DATA PROCESSING

We use the same 12 biomarkers and preprocessing as TEMPO (Hao & Austerweil, 2026) and BEBMS (Hao & Austerweil, 2025): ABETA, TAU, PTAU (CSF), ADAS13, RAVLT immediate recall, MMSE (cognitive), and Entorhinal, Hippocampus, MidTemp, Fusiform, WholeBrain, Ventricle volumes (MRI, FreeSurfer-processed). Only participants with complete observations on all 12 biomarkers are included. Diagnosis groups are mapped to a binary disease label: CN = 0; EMCI, LMCI, AD = 1. MRI volumes are adjusted for intracranial volume and age via a linear model fit on CN participants. All biomarkers are z-score standardized using CN participants as reference.

J BASELINE MODELS ON ADNI

As a complement to Table 7, we also applied BEBMS, BEBMS (Blind), and SuStaIn (GMM) to the same ADNI cohort, using the same $K = 3$ setting as the main paper. Table 8 reports each model's own predicted stage by diagnosis, broken down by its own predicted subtypes. Because each model

Table 7: ADNI predicted direct stage by diagnosis after subtype matching. Values are averaged across the nine 12-bio models.

Subtype	CN	EMCI	LMCI	AD
1	0.00 ± 0.00	6.38 ± 2.56	8.67 ± 2.10	10.15 ± 1.01
2	0.00 ± 0.00	4.55 ± 2.40	7.01 ± 2.58	9.67 ± 1.24
3	0.00 ± 0.00	5.68 ± 2.32	7.99 ± 2.31	10.04 ± 1.36

is fit independently, subtype indices are not matched across models or to SUBTRACE's subtypes in Table 7; only within-model comparisons across diagnosis groups are meaningful here. BEBMS (Blind) subtype 3 has no CN participants assigned to it, so that cell is empty.

Table 8: ADNI predicted stage by diagnosis for the baseline models (mean $\pm$ std). Subtype indices are model-specific, not matched across rows or to Table 7.

Model	Subtype	CN	EMCI	LMCI	AD
BEBMS	1	5.11 ± 3.38	5.35 ± 2.87	9.36 ± 2.40	10.77 ± 0.68
	2	9.00 ± 4.24	7.36 ± 2.90	7.58 ± 4.45	10.07 ± 2.85
	3	0.44 ± 1.65	1.19 ± 2.46	5.60 ± 4.95	10.18 ± 2.44
BEBMS (Blind)	1	2.50 ± 2.91	3.88 ± 3.82	8.73 ± 4.56	12.17 ± 1.65
	2	3.18 ± 3.57	2.85 ± 3.76	5.79 ± 3.55	12.00 ± 0.00
	3	—	8.31 ± 2.39	10.38 ± 1.86	11.03 ± 1.17
SuStaIn (GMM)	1	1.88 ± 2.71	3.69 ± 3.61	7.88 ± 3.69	10.37 ± 1.71
	2	3.38 ± 3.19	4.98 ± 3.02	7.45 ± 3.36	9.33 ± 2.12
	3	4.53 ± 2.42	5.00 ± 3.35	7.50 ± 1.51	10.00 ± 1.41

SuStaIn (KDE) is not included: fitting its KDE mixture model on this cohort raised `LinAlgError: Singular matrix` inside `awkde`'s `GaussianKDE.fit` (a Cholesky decomposition of a singular covariance matrix during standardization), so it produced no ADNI result.

K NACC RESULTS

We obtained clinical, cerebrospinal fluid (CSF), and MRI data from the National Alzheimer's Coordinating Center (NACC) (Morris et al., 2006), supplemented with harmonized cognitive domain scores from the Alzheimer's Disease Sequencing Project Phenotype Harmonization Consortium (ADSP-PHC) (Mukherjee et al., 2023). We considered Uniform Data Set (UDS) visits conducted between September 2005 and June 2025. We use four NACC data products: the UDS clinical module with the FTLD and LBD supplements, investigator MRI, investigator CSF, and the ADSP Cognitive Composite (PHC) scores. At each clinical visit, a participant is labeled cognitively normal if `NACCUDSD`$=1$, MCI if `NACCUDSD`=3 and `NACCALZD`=1, and AD if `NACCUDSD`=4 and `NACCALZD`=1; visits meeting none of these three conditions are excluded. CSF measurements are matched to the nearest clinical visit within 180 days. A participant is included only if CSF, MRI, and PHC are all complete at a single shared visit, and we take each participant's earliest such visit. All MRI volumes are normalized by intracranial volume. The three NACC-specific cognitive scores are memory, executive function, and language (`PHC_MEM`, `PHC_EXF`, `PHC_LAN`); the remaining nine biomarkers use the same CSF and MRI measures as ADNI.

The finally processed dataset contains 141 participants after preprocessing (82 controls, 25 with MCI, and 34 with AD) and 12 biomarkers. Given the modest sample size, particularly the 59 non-control participants, the subtype-level findings reported below are exploratory and should be interpreted with caution.

We used SA-EBM (Hao et al., 2025) to obtain the theta/phi parameters of these 12 biomarkers. Then we create synthetic datasets and train SUBTRACE models the same way as in ADNI experiment 5.1. We choose $K = 3$ as that is what BEBMS (Hao & Austerweil, 2025) and SuStaIn (Young et al., 2018) both suggested based on the NACC data (Figure 4).

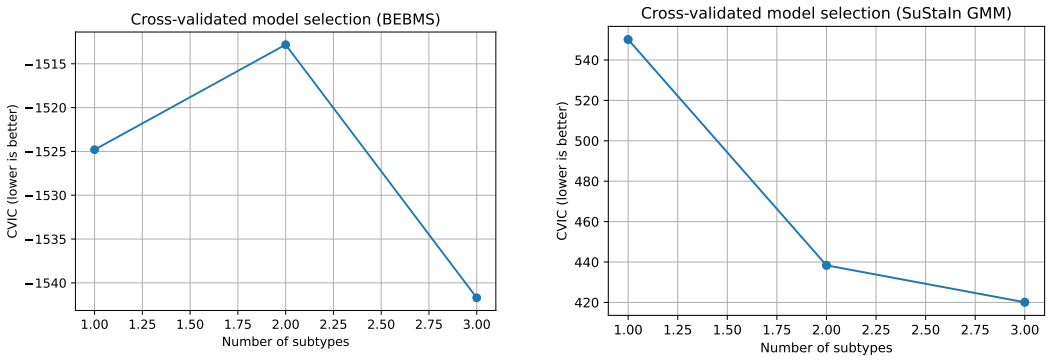


Figure 4: Cross-validated model selection on NACC. Left: BEBMS (5-fold CVIC). Right: SuStaIn mixture-GMM (5-fold CVIC). Both favor $K = 3$ within the tested range $K \in \{1, 2, 3\}$.

Table 9: NACC predicted direct stage by diagnosis after subtype matching. Values are averaged across the nine 12-bio models.

Subtype	CN	MCI	AD
1	0.00 ± 0.00	7.55 ± 2.36	8.78 ± 1.59
2	0.00 ± 0.00	7.24 ± 2.49	9.25 ± 1.42
3	0.00 ± 0.00	6.68 ± 2.52	8.76 ± 1.87

Table 9 shows a monotone stage pattern within every matched subtype: CN participants receive stage 0, MCI is intermediate, and AD is latest. The model receives only the binary CN/non-CN label, so the monotone increase across the three diagnostic categories is not imposed by its input. This replicates the ADNI pattern of Table 7 on an independent cohort. We apply model trained on datasets of Exp 7 to NACC. This is because it has the best cross-experimental ordering result (Table 10).

Table 10: Cross-experiment Kendall distance (lower is better) on synthetic 12-bio test data generated with NACC-matched simulator parameters, averaged over all nine test experiments, for each of the nine trained SUBTRACE models. Exp. 7 has the lowest row-mean distance.

Trained on	Exp1	Exp2	Exp3	Exp4	Exp5	Exp6	**Exp7**	Exp8	Exp9
Row-mean τ_{dist}	0.229	0.218	0.231	0.216	0.226	0.233	**0.215**	0.229	0.221

Across all subtypes, the PHC scores measuring memory, language and executive function are the first indicators of Alzheimer's in this NACC cohort. That memory decline precedes executive function and language decline is consistent with Grober et al. (2008). The single-subtype result in Hao et al. (2025) based on ADNI also showed that cognitive decline occurs relatively early in the progression.

The subtype 1, accounting for over half of the non-control participants, corresponds to Typical AD (Subtype 1 of ADNI) with CSF biomarkers becoming pathological before MRI structural changes. This is largely consistent with Jack et al. (2010). That whole brain atrophy and ventricular enlargement occur relatively late is largely consistent with Jack et al. (2010); Jack Jr et al. (2009). Also, Jack et al. (2010) suggested that medial temporal structural changes in Entorhinal and Hippo. precede changes in lateral temporal areas (MidTemp and Fusiform), and the subtype 1 result is largely consistent with this pattern.

Subtype 2 and 3 are relatively mixed and the results should be interpreted cautiously. Subtype 2 is similar to the subtype 1 reported in Hao & Austerweil (2025), with neurodegeneration occurring early, corresponding to the Neurodegeneration-first biomarker model of late-onset AD (Jack & Holtzman, 2013) and partially compatible with the Limbic-predominant AD (Murray et al., 2011). With Abeta marginally preceding MRI structural changes, Subtype 2 is consistent with the hypothetical cascade model of Jack et al. (2010). Subtype 3, with Hippo. occurring relatively late, is similar to the Hippocampal-sparing AD reported in Murray et al. (2011).

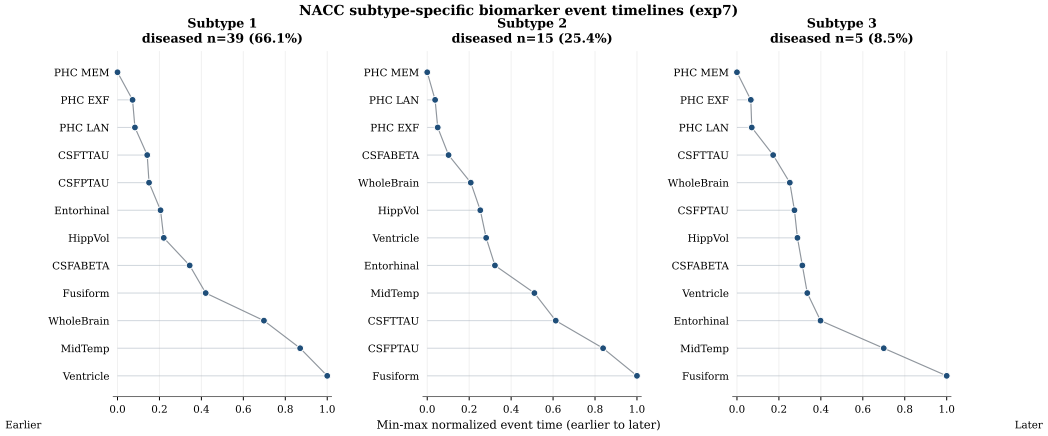


Figure 5: Example continuous NACC timelines from the Exp. 7 SUBTRACE model.

To check whether these Exp. 7-based patterns generalize, we additionally align subtype orderings across all nine trained models via the Hungarian algorithm, as in the ADNI analysis (§5.1). Figure 6 shows this nine-model consensus is largely consistent with the Exp. 7-based patterns described above.

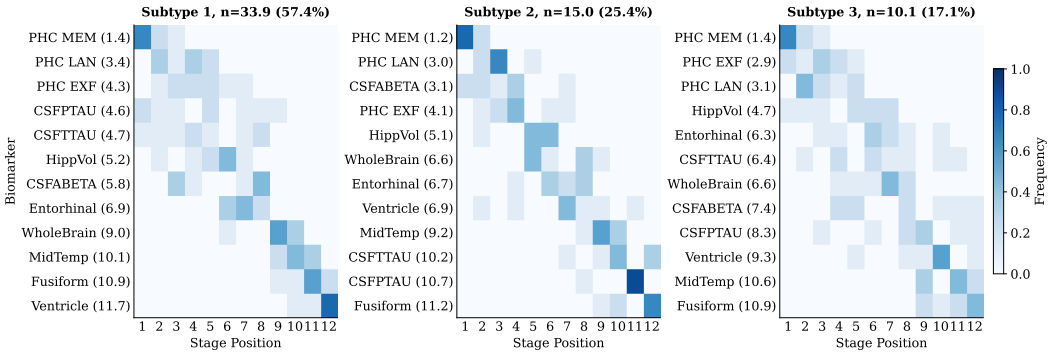


Figure 6: Consensus subtype orderings across nine SUBTRACE models on NACC, aligned via the Hungarian algorithm as in Figure 2.

L NORMALIZED STAGE MAE

Table 11: Normalized Stage MAE (MAE $/B$) across dimensionality conditions.

Algorithm	12 bio	100 bio	Change
SUBTRACE	0.078	0.041	**improves**
SUBTRACE (Blind)	0.089	0.045	improves
BEBMS	0.103	0.091	roughly stable
BEBMS (Blind)	0.109	0.094	roughly stable
SuStaIn (GMM)	0.152	0.198	worsens
SuStaIn (KDE)	0.193	0.150	improves slightly

M CONTROL-STAGE CALIBRATION

Non-blind SUBTRACE assigns healthy patients exactly stage 0.000 by construction, since it receives d_j as an input and the control-stage penalty (Appendix B) pushes controls below the first rounding boundary; this is not a result to compare against the baselines. The fair comparison is the blind variant, which loses almost nothing on ordering ($\tau = 0.172$ vs. 0.162 at 12 bio) while still calibrating controls far better than the blind baselines (0.164 vs. 0.354 for blind BEBMS at 12 bio, 0.032 vs. 0.881 at 100). SuStaIn (GMM) assigns controls an average stage of 43.4 out of a maximum of 100.

N EVENT POSITION MAE

Event position MAE is reported for SUBTRACE, SUBTRACE (Blind), and K-Means + TEMPO, the methods that predict continuous event positions $\hat{r}_{k,b} \in [0, 1]$ directly; the remaining baselines recover only an ordinal event sequence. Their predicted orderings were not retained for Experiments 1–8, so the metric is unavailable for them in Table 2. In the $K = 5$ benchmark (Table 13), all methods are scored on ordinal ranks.

Table 12 gives the values, in units of event positions (so an error of 1.736 at 12 biomarkers means the predicted position is off by about 1.7 ranks on average). SUBTRACE and its blind variant are indistinguishable from each other here and both improve on K-Means + TEMPO in both dimensionalities, with Random Guessing included only to indicate the scale of the metric.

Table 12: Event position MAE (mean $\pm$ 95% CI; $n = 450$ per setting). Lower is better. Reported only for methods that predict continuous event positions; see the caveats above.

Algorithm	12 bio ↓	100 bio ↓
SUBTRACE	1.736±0.036	13.483±0.243
SUBTRACE (Blind)	**1.712±0.032**	**13.269±0.240**
K-Means + TEMPO	2.138±0.045	17.940±0.219
Random	3.517±0.037	32.223±0.161

For SUBTRACE and SUBTRACE (Blind), the predicted event position is $\hat{r}_{k,b} \cdot B$, where B is the number of biomarkers. Ground truth is the true continuous event position when the simulator generated one (Experiments 8–9, which use continuous event times); for the remaining experiments, which only have an ordinal ground-truth ordering, the ground truth position falls back to the biomarker's ordinal rank (plus one). Event position MAE is therefore comparing a continuous prediction against an ordinal target in Experiments 1–7, which is why it is a diagnostic of prediction granularity rather than a metric with a single, experiment-independent interpretation.

K-Means + TEMPO uses a different normalization because TEMPO itself is trained against a differently-scaled target: its predicted position is $\hat{r}_{k,b} \cdot B$ against continuous ground truth (Experiments 8–9), but $\hat{r}_{k,b} \cdot (B-1)+1$ against ordinal ground truth (Experiments 1–7). This is a deliberate consequence of TEMPO's own training target, not an inconsistency introduced here.

O $K = 1$ AND $K = 5$ ROBUSTNESS CHECKS

We also ran single-subtype experiments as a sanity check against TEMPO. In the $K = 1$ low-dimensional setting, SUBTRACE improves Kendall τ distance from 0.034 for TEMPO to 0.030, but has higher stage MAE (0.610 vs. 0.549). In the 100-biomarker setting, SUBTRACE improves τ from 0.056 to 0.050, and also has lower stage MAE (2.233 vs. 2.543). We therefore treat the $K = 1$ result as evidence that the ordering branch remains strong without subtype heterogeneity, not as a claim that the current stage head uniformly dominates TEMPO at every dimensionality.

As a further robustness check, we also generated a low-dimensional, Experiment 1 benchmark with $K = 5$ subtypes instead of $K = 3$ (10 held-out test datasets) and compared SUBTRACE against the same four fitted baselines. Table 13 reports the results. SUBTRACE again has the best ordering, stage MAE, and event position MAE (its Healthy score of 0.000 holds by construction, as in Table 2, so it is not a comparable result); the one exception is subtyping accuracy, where BEBMS (Blind) has higher ARI (0.336 vs. 0.314). Event position MAE is scored on ordinal ranks for all methods; SUBTRACE's continuous predictions score 1.721 ± 0.104 on the same datasets. With only $n = 10$ test datasets the CIs on ARI are wide and overlapping, so this is best read as a mild caveat rather than a robust reversal.

Table 13: $K = 5$, low-dimensional, Experiment 1 results (mean $\pm$ 95% CI; $n = 10$). Lower τ, event position MAE, and stage MAE are better; higher ARI is better.

Algorithm	$\tau \downarrow$	ARI $\uparrow$	Stage MAE $\downarrow$	Event Pos. MAE $\downarrow$	Healthy $\downarrow$
SUBTRACE	**0.180±0.016**	0.314±0.088	**1.018±0.055**	**1.607±0.120**	0.000±0.000
BEBMS (Blind)	0.233±0.025	**0.336±0.096**	1.206±0.136	2.030±0.196	0.262±0.186
BEBMS	0.259±0.033	0.279±0.083	1.281±0.115	2.213±0.247	0.093±0.037
SuStaIn (GMM)	0.259±0.054	0.249±0.053	1.546±0.246	2.190±0.421	1.885±0.578
SuStaIn (KDE)	0.327±0.030	0.070±0.021	2.399±0.242	2.757±0.256	3.534±0.378

P ARCHITECTURE ABLATION

We also ablate four architectural components to test whether each one contributes to SUBTRACE's performance: the seed router's own auxiliary supervision, the second timeline-refinement pass, the evidence-vs-structural profile-fit matching that ties the evidence head to subtype assignment, and rank-based tokenization. Each variant is retrained from scratch on the low-dimensional Experiment 1 data and evaluated on the same 50 held-out test datasets. Table 3 reports the results.

Each intervention isolates a distinct effect rather than uniformly degrading every metric. Removing the seed router's own auxiliary loss (while keeping the router itself, since it is also the symmetry-breaking mechanism that makes the K subtype slots distinct) is the only variant that is worse on ordering, subtyping, and staging simultaneously. Skipping the second refinement pass mainly costs ordering; subtyping is largely unaffected, and staging improves too ($0.926 \rightarrow 0.867$), echoing the profile-fit removal below, though we do not have a shared explanation for why fewer refinement iterations would also sharpen the stage head. Removing the evidence-vs-structural profile-fit matching costs subtyping, consistent with this mechanism's role of tying the evidence head to subtype assignment. Stage MAE improves in this variant ($0.926 \rightarrow 0.820$), but we do not read this as evidence that the mechanism is unneeded for staging: with subtyping no longer competing for the same evidence-derived signal, the stage head has an easier, narrower target at the cost of subtype accuracy. This is precisely the kind of tradeoff the joint architecture is designed to avoid; optimizing any one task in isolation is not the goal. Removing rank-based tokenization mainly costs staging, with ordering and subtyping largely unaffected, and with no corresponding gain elsewhere. No component's removal is harmless, but the size of the effect varies widely, from broad and unambiguous (the seed-router loss) to narrow (the second pass, rank-based tokenization), and each affects a different one of the three tasks, sometimes at another task's expense.

We repeat the same four ablations at 100 biomarkers (Table 14) to check whether the pattern above holds at scale.

Table 14: Ablation of four architectural components (high-dimensional Experiment 1, mean ± 95% CI; $n = 50$).

Variant	$\tau \downarrow$	ARI ↑	Stage MAE ↓
Full model	**0.097±0.010**	0.144±0.064	3.540±0.114
No seed-router loss	0.103±0.011	**0.151±0.065**	4.030±0.201
One pass + profile fit	0.104±0.011	0.136±0.058	**3.044±0.135**
Seed-only (no profile-fit matching)	0.100±0.012	0.028±0.017	3.269±0.150
No rank features	0.098±0.011	0.104±0.060	4.478±0.152

The effect on ordering is modest and similar in size to the 12-biomarker table for every intervention, so removing a component does not become more costly for τ specifically at 100 biomarkers. Subtyping tells a different story: removing the profile-fit matching costs ARI far more severely here (0.144 → 0.028, a relative drop of about 81%) than at 12 biomarkers (0.371 → 0.318, about 14%). This supports a narrower version of the claim in Section 6: the abnormality-profile check becomes more important for subtype assignment, specifically, as B grows, rather than for ordering accuracy directly. Removing the seed-router loss no longer costs ARI at 100 biomarkers (it is marginally higher than the full model, within noise), unlike at 12 biomarkers where it is the most damaging intervention on every metric; we do not have a mechanistic explanation for this reversal and flag it as an open question rather than read too much into one experiment's ablation.

Q FULL PER-EXPERIMENT RESULTS

Tables 15–20 report per-experiment ordering, subtype, and stage results in both dimensionality conditions. Per-experiment results show the same structure as the aggregate results. SUBTRACE has the best Kendall τ distance in every experiment at both 12 and 100 biomarkers, and the best stage MAE except Exp. 9 at 12 bio. Other methods sometimes achieve higher ARI: BEBMS and SuStaIn (GMM) do so in some 12-bio experiments, and SuStaIn (GMM) is often strongest on ARI at 100 biomarkers. However, those ARI gains coincide with worse ordering and stage accuracy.

Table 15: 12-bio: Kendall τ distance per experiment (lower is better; mean $\pm$ 95% CI, $n = 50$).

Exp	SUBTRACE	BEBMS	BEBMS (B)	SuStaIn (G)	SuStaIn (K)	Random
1	**0.136±0.010**	0.179±0.019	0.216±0.019	0.217±0.021	0.314±0.016	0.432±0.015
2	**0.177±0.013**	0.239±0.020	0.286±0.022	0.374±0.019	0.387±0.017	0.417±0.013
3	**0.143±0.013**	0.174±0.016	0.198±0.018	0.211±0.020	0.334±0.017	0.427±0.012
4	**0.202±0.015**	0.224±0.017	0.250±0.018	0.374±0.019	0.402±0.013	0.412±0.010
5	**0.118±0.012**	0.214±0.016	0.218±0.019	0.307±0.027	0.323±0.022	0.436±0.014
6	**0.144±0.011**	0.209±0.016	0.218±0.019	0.213±0.025	0.305±0.019	0.422±0.013
7	**0.195±0.014**	0.248±0.019	0.275±0.019	0.398±0.025	0.402±0.017	0.427±0.013
8	**0.156±0.013**	0.217±0.016	0.209±0.012	0.332±0.018	0.348±0.017	0.431±0.014
9	**0.189±0.013**	0.232±0.016	0.247±0.016	0.259±0.020	0.345±0.016	0.437±0.013

Table 16: 12-bio: ARI per experiment (higher is better; mean $\pm$ 95% CI, $n = 50$). Bold = best per row.

Exp	SUBTRACE	BEBMS	BEBMS (B)	SuStaIn (G)	SuStaIn (K)	Random
1	0.371±0.053	**0.427±0.062**	0.384±0.061	0.356±0.059	0.097±0.025	0.000±0.001
2	0.242±0.040	**0.249±0.055**	0.228±0.047	0.114±0.031	0.049±0.015	0.000±0.001
3	0.245±0.035	0.265±0.038	0.251±0.037	**0.276±0.043**	0.055±0.013	-0.001±0.001
4	0.133±0.027	**0.183±0.031**	0.161±0.033	0.096±0.026	0.028±0.008	0.000±0.001
5	**0.451±0.064**	0.339±0.068	0.360±0.070	0.312±0.066	0.095±0.023	0.000±0.001
6	**0.412±0.053**	0.385±0.057	0.376±0.063	0.373±0.055	0.125±0.026	0.000±0.001
7	0.271±0.048	0.263±0.054	**0.272±0.056**	0.139±0.036	0.052±0.015	0.000±0.001
8	**0.277±0.041**	0.146±0.034	0.150±0.043	0.131±0.036	0.038±0.011	0.000±0.001
9	**0.189±0.034**	0.174±0.038	0.144±0.037	0.144±0.028	0.046±0.012	0.000±0.001

Table 17: 12-bio: Stage MAE per experiment (lower is better; mean $\pm$ 95% CI, $n = 50$).

Exp	SUBTRACE	BEBMS	BEBMS (B)	SuStaIn (G)	SuStaIn (K)	Random
1	**0.926±0.026**	1.136±0.075	1.268±0.093	1.406±0.085	2.368±0.089	4.181±0.040
2	**1.040±0.033**	1.587±0.073	1.769±0.093	2.567±0.151	2.840±0.119	4.142±0.036
3	**0.930±0.023**	1.068±0.040	1.123±0.047	1.416±0.089	2.386±0.099	4.524±0.025
4	**1.204±0.030**	1.465±0.056	1.567±0.062	2.656±0.117	3.064±0.123	4.530±0.036
5	**0.712±0.022**	1.355±0.098	1.391±0.097	1.672±0.100	2.142±0.061	4.302±0.026
6	**0.807±0.025**	1.053±0.051	1.094±0.070	1.351±0.059	1.948±0.071	4.280±0.030
7	**0.925±0.020**	1.462±0.075	1.540±0.069	2.233±0.086	2.397±0.074	4.273±0.024
8	**0.909±0.037**	1.015±0.049	1.029±0.053	1.580±0.055	1.846±0.067	4.890±0.056
9	0.964±0.032	**0.954±0.038**	0.994±0.047	1.518±0.042	1.834±0.072	4.883±0.061

Event position MAE is reported only for SUBTRACE, SUBTRACE (Blind), and K-Means + TEMPO, since these are the only methods that produce continuous event positions; the ordinal baselines (BEBMS, its blind variant, and both SuStaIn variants) do not, so they are omitted from Tables 21 and 22. See Appendix N for how this metric is computed and why it is a diagnostic rather than a benchmark comparison.

Table 18: 100-bio: Kendall τ distance per experiment (lower is better; mean $\pm$ 95% CI, $n = 50$).

Exp	SUBTRACE	BEBMS	BEBMS (B)	SuStaIn (G)	SuStaIn (K)	Random
1	**0.097±0.009**	0.391±0.014	0.401±0.015	0.377±0.012	0.343±0.008	0.480±0.005
2	**0.126±0.012**	0.389±0.015	0.393±0.015	0.461±0.009	0.422±0.009	0.478±0.004
3	**0.090±0.011**	0.358±0.019	0.357±0.015	0.345±0.010	0.344±0.008	0.479±0.005
4	**0.105±0.018**	0.367±0.011	0.370±0.010	0.462±0.008	0.432±0.008	0.479±0.005
5	**0.133±0.011**	0.394±0.011	0.405±0.008	0.409±0.008	0.330±0.008	0.480±0.003
6	**0.121±0.009**	0.381±0.015	0.387±0.014	0.362±0.011	0.331±0.009	0.477±0.004
7	**0.149±0.014**	0.386±0.012	0.389±0.010	0.463±0.009	0.425±0.009	0.482±0.004
8	**0.129±0.009**	0.369±0.011	0.363±0.011	0.406±0.010	0.337±0.007	0.480±0.004
9	**0.123±0.011**	0.362±0.012	0.367±0.013	0.346±0.009	0.343±0.007	0.475±0.004

Table 19: 100-bio: ARI per experiment (higher is better; mean $\pm$ 95% CI, $n = 50$). Bold = best per row.

Exp	SUBTRACE	BEBMS	BEBMS (B)	SuStaIn (G)	SuStaIn (K)	Random
1	0.144±0.064	0.127±0.065	0.163±0.073	**0.312±0.092**	0.054±0.034	0.001±0.001
2	0.102±0.055	0.114±0.073	**0.117±0.070**	0.086±0.053	0.029±0.027	0.000±0.001
3	0.037±0.024	0.103±0.068	0.096±0.053	**0.171±0.074**	0.044±0.029	0.000±0.001
4	0.024±0.017	**0.083±0.050**	0.049±0.030	0.037±0.027	0.003±0.003	0.000±0.001
5	0.186±0.067	0.063±0.052	0.060±0.050	**0.220±0.085**	0.043±0.028	0.001±0.001
6	0.146±0.057	0.133±0.077	0.123±0.074	**0.305±0.094**	0.068±0.042	0.000±0.001
7	**0.121±0.055**	0.104±0.057	0.071±0.050	0.103±0.050	0.026±0.019	0.000±0.001
8	0.126±0.050	0.056±0.037	0.077±0.044	**0.136±0.059**	0.027±0.014	0.000±0.001
9	0.089±0.038	0.064±0.043	0.044±0.038	**0.140±0.059**	0.032±0.024	0.000±0.001

Table 20: 100-bio: Stage MAE per experiment (lower is better; mean $\pm$ 95% CI, $n = 50$).

Exp	SUBTRACE	BEBMS	BEBMS (B)	SuStaIn (G)	SuStaIn (K)	Random
1	**3.541±0.114**	14.515±1.085	16.058±1.150	21.379±1.039	18.765±0.569	33.416±0.237
2	**4.473±0.170**	13.195±1.128	13.543±1.054	27.392±0.851	18.665±0.840	33.018±0.247
3	**3.325±0.097**	7.084±0.623	7.044±0.500	19.516±0.587	17.537±0.417	37.056±0.258
4	**4.208±0.093**	7.151±0.572	7.258±0.449	28.622±0.721	20.626±0.667	36.924±0.235
5	**3.747±0.165**	9.773±0.793	10.351±0.809	16.428±0.528	13.032±0.355	35.780±0.253
6	**4.489±0.219**	9.821±0.773	9.765±0.770	15.952±0.603	11.373±0.339	36.025±0.314
7	**3.933±0.101**	9.054±0.704	9.621±0.723	18.402±0.450	14.558±0.290	35.974±0.244
8	**5.044±0.258**	5.986±0.415	5.922±0.369	16.239±0.389	10.827±0.247	39.550±0.280
9	**4.263±0.144**	5.501±0.333	5.378±0.332	14.972±0.390	10.514±0.277	39.331±0.282

Table 21: 12-bio: Event position MAE per experiment (lower is better; mean $\pm$ 95% CI, $n = 50$).

Exp	SUBTRACE	SUBTRACE (Blind)	K-Means + TEMPO
1	1.641±0.085	**1.625±0.077**	1.758±0.101
2	**1.775±0.103**	1.801±0.105	2.037±0.104
3	**1.452±0.098**	1.529±0.111	2.363±0.102
4	1.936±0.114	**1.897±0.095**	2.631±0.117
5	**1.744±0.100**	1.760±0.072	1.868±0.117
6	1.841±0.099	**1.652±0.085**	2.207±0.122
7	2.007±0.120	**2.003±0.092**	2.402±0.131
8	1.648±0.077	**1.619±0.073**	1.850±0.105
9	1.579±0.083	**1.519±0.065**	2.128±0.108

Table 22: 100-bio: Event position MAE per experiment (lower is better; mean ± 95% CI, $n = 50$).

Exp	SUBTRACE	SUBTRACE (Blind)	K-Means + TEMPO
1	14.857±0.458	**14.520±0.262**	16.264±0.406
2	**13.253±0.665**	14.133±0.714	17.088±0.422
3	**11.838±0.450**	12.245±0.408	17.452±0.261
4	11.476±0.855	**9.693±1.021**	21.480±0.205
5	13.819±0.713	**13.179±0.515**	19.373±0.473
6	13.324±0.595	**12.723±0.543**	19.516±0.221
7	**13.427±1.062**	13.512±0.752	19.592±0.442
8	**14.712±0.288**	15.067±0.327	15.485±0.289
9	14.641±0.533	**14.345±0.280**	15.213±0.223

R CROSS-EXPERIMENT RESULTS

Tables 23 and 24 show the full 9×9 cross-experiment Kendall τ matrices, and Tables 25–30 report the same cross-experiment breakdown for ARI, stage MAE, and event position MAE. Entry (i, j) is the τ distance of the model trained on Exp. i tested on Exp. j.

Cross-experiment transfer tests robustness to simulator misspecification. We train each of the nine SUBTRACE models under one synthetic assumption and apply it to all nine test experiments. In 12 biomarkers, row-mean Kendall τ ranges from 0.214 to 0.246; in 100 biomarkers it ranges from 0.154 to 0.210. Thus every trained model remains competitive on ordering even when the test generator differs from the training generator. The pattern also reveals useful sensitivity information: EBM-trained models tend to transfer better than sigmoid-trained models; non-normal biomarker training often helps, especially with Dirichlet-Multinomial stages; and continuous event times help most clearly under the sigmoid measurement model.

Table 23: Cross-experiment Kendall τ distance (12 bio; mean $\pm$ 95% CI, $n = 50$). Diagonal = in-distribution. The Mean column is a point estimate (no CI available).

Train\Test	1	2	3	4	5	6	7	8	9	Mean
1	0.136±0.010	0.243±0.018	0.142±0.014	0.283±0.018	0.207±0.013	0.158±0.011	0.261±0.017	0.292±0.017	0.211±0.013	0.215
2	0.170±0.012	0.177±0.013	0.180±0.015	0.203±0.015	0.219±0.017	0.203±0.012	0.220±0.014	0.307±0.017	0.261±0.017	0.215
3	0.161±0.013	0.251±0.017	0.143±0.013	0.284±0.019	0.199±0.015	0.180±0.012	0.265±0.017	0.281±0.016	0.217±0.015	0.220
4	0.181±0.013	0.187±0.013	0.197±0.015	0.202±0.015	0.238±0.017	0.203±0.014	0.208±0.012	0.319±0.017	0.258±0.017	0.221
5	0.232±0.015	0.293±0.017	0.253±0.013	0.341±0.019	0.118±0.012	0.227±0.013	0.292±0.014	0.180±0.015	0.276±0.012	0.246
6	0.157±0.012	0.263±0.018	0.192±0.014	0.298±0.017	0.179±0.012	0.144±0.011	0.254±0.017	0.262±0.017	0.206±0.014	0.217
7	0.176±0.012	0.201±0.015	0.213±0.012	0.236±0.012	0.197±0.013	0.177±0.011	0.195±0.014	0.292±0.017	0.255±0.014	0.216
8	0.201±0.011	0.267±0.017	0.220±0.012	0.301±0.015	0.125±0.010	0.222±0.013	0.299±0.012	0.156±0.013	0.274±0.013	0.229
9	0.158±0.011	0.254±0.016	0.168±0.013	0.293±0.020	0.188±0.012	0.164±0.010	0.265±0.018	0.247±0.015	0.189±0.013	0.214

Table 24: Cross-experiment Kendall τ distance (100 bio; mean $\pm$ 95% CI, $n = 50$). Diagonal = in-distribution. The Mean column is a point estimate (no CI available).

Train\Test	1	2	3	4	5	6	7	8	9	Mean
1	0.097±0.009	0.245±0.011	0.095±0.009	0.263±0.010	0.216±0.008	0.126±0.007	0.278±0.011	0.272±0.009	0.163±0.011	0.195
2	0.135±0.010	0.126±0.012	0.164±0.008	0.140±0.012	0.199±0.011	0.173±0.007	0.171±0.012	0.234±0.010	0.202±0.009	0.171
3	0.104±0.011	0.272±0.012	0.090±0.011	0.282±0.010	0.194±0.009	0.141±0.009	0.280±0.010	0.226±0.010	0.148±0.009	0.193
4	0.128±0.011	0.123±0.015	0.112±0.011	0.105±0.018	0.200±0.011	0.163±0.008	0.165±0.016	0.226±0.011	0.177±0.010	0.156
5	0.163±0.009	0.255±0.009	0.181±0.007	0.277±0.009	0.133±0.011	0.207±0.006	0.294±0.010	0.140±0.008	0.239±0.007	0.210
6	0.107±0.010	0.243±0.011	0.112±0.009	0.260±0.010	0.192±0.007	0.121±0.009	0.272±0.010	0.240±0.010	0.130±0.009	0.186
7	0.125±0.008	0.130±0.012	0.126±0.010	0.141±0.013	0.198±0.010	0.142±0.009	0.149±0.014	0.226±0.009	0.150±0.010	0.154
8	0.169±0.008	0.260±0.010	0.172±0.007	0.283±0.009	0.128±0.011	0.201±0.007	0.297±0.009	0.129±0.009	0.240±0.009	0.209
9	0.117±0.010	0.259±0.011	0.107±0.009	0.276±0.010	0.190±0.009	0.122±0.009	0.278±0.010	0.220±0.010	0.123±0.011	0.188

Table 25: Cross-experiment ARI (12 bio; mean $\pm$ 95% CI, $n = 50$). Diagonal = in-distribution. The Mean column is a point estimate (no CI available).

Train\Test	1	2	3	4	5	6	7	8	9	Mean
1	0.371±0.053	0.178±0.041	0.239±0.032	0.101±0.023	0.318±0.053	0.404±0.050	0.200±0.040	0.163±0.033	0.198±0.036	0.241
2	0.330±0.054	0.242±0.040	0.198±0.032	0.133±0.026	0.263±0.056	0.323±0.045	0.259±0.042	0.117±0.030	0.163±0.032	0.225
3	0.351±0.046	0.159±0.039	0.245±0.035	0.089±0.021	0.255±0.047	0.353±0.042	0.181±0.035	0.141±0.028	0.187±0.037	0.218
4	0.260±0.051	0.207±0.042	0.157±0.029	0.133±0.027	0.207±0.038	0.303±0.047	0.253±0.043	0.117±0.028	0.140±0.032	0.197
5	0.226±0.048	0.131±0.037	0.133±0.027	0.064±0.018	0.451±0.064	0.229±0.044	0.113±0.032	0.279±0.043	0.094±0.023	0.191
6	0.378±0.051	0.171±0.041	0.216±0.031	0.103±0.024	0.317±0.056	0.412±0.053	0.213±0.034	0.166±0.039	0.205±0.037	0.242
7	0.304±0.047	0.202±0.038	0.157±0.030	0.110±0.027	0.280±0.053	0.315±0.047	0.271±0.048	0.140±0.031	0.161±0.031	0.215
8	0.218±0.040	0.097±0.027	0.131±0.029	0.055±0.017	0.436±0.063	0.191±0.037	0.073±0.020	0.277±0.041	0.080±0.020	0.173
9	0.319±0.049	0.151±0.033	0.179±0.030	0.091±0.022	0.255±0.044	0.345±0.054	0.172±0.033	0.153±0.026	0.189±0.034	0.206

Table 26: Cross-experiment ARI (100 bio; mean ± 95% CI, $n = 50$). Diagonal = in-distribution. The Mean column is a point estimate (no CI available).

Train\Test	1	2	3	4	5	6	7	8	9	Mean
1	0.144±0.064	0.071±0.041	0.055±0.032	0.035±0.025	0.085±0.044	0.126±0.052	0.083±0.037	0.044±0.023	0.065±0.026	0.079
2	0.101±0.052	0.102±0.055	0.046±0.035	0.052±0.039	0.072±0.038	0.084±0.045	0.141±0.064	0.026±0.013	0.030±0.021	0.072
3	0.097±0.057	0.049±0.034	0.037±0.024	0.030±0.024	0.064±0.036	0.101±0.042	0.062±0.029	0.032±0.016	0.056±0.033	0.059
4	0.070±0.048	0.063±0.038	0.013±0.016	0.024±0.017	0.050±0.041	0.057±0.043	0.076±0.039	0.015±0.009	0.032±0.025	0.045
5	0.098±0.048	0.052±0.032	0.043±0.030	0.023±0.013	0.186±0.067	0.080±0.040	0.059±0.032	0.142±0.057	0.055±0.033	0.082
6	0.126±0.057	0.065±0.036	0.036±0.025	0.027±0.020	0.114±0.062	0.146±0.057	0.100±0.048	0.045±0.023	0.065±0.032	0.080
7	0.091±0.059	0.088±0.050	0.024±0.016	0.031±0.021	0.073±0.029	0.096±0.040	0.121±0.055	0.052±0.027	0.065±0.032	0.071
8	0.073±0.044	0.037±0.026	0.033±0.022	0.024±0.014	0.165±0.067	0.058±0.033	0.050±0.027	0.126±0.050	0.042±0.029	0.068
9	0.115±0.055	0.042±0.027	0.041±0.031	0.017±0.012	0.089±0.043	0.121±0.044	0.053±0.029	0.061±0.030	0.089±0.038	0.070

Table 27: Cross-experiment stage MAE (12 bio; mean ± 95% CI, $n = 50$). Diagonal = in-distribution. The Mean column is a point estimate (no CI available).

Train\Test	1	2	3	4	5	6	7	8	9	Mean
1	0.926±0.026	1.153±0.037	1.230±0.022	1.581±0.034	0.799±0.021	1.117±0.028	1.361±0.037	1.159±0.037	1.555±0.042	1.209
2	0.937±0.027	1.040±0.033	1.231±0.019	1.425±0.030	0.953±0.025	1.219±0.040	1.305±0.030	1.364±0.037	1.804±0.062	1.253
3	0.923±0.028	1.166±0.035	0.930±0.023	1.387±0.050	0.916±0.032	0.972±0.036	1.271±0.035	0.946±0.030	1.109±0.023	1.069
4	0.962±0.028	1.080±0.029	1.002±0.019	1.204±0.030	1.012±0.020	1.177±0.044	1.319±0.033	1.157±0.026	1.606±0.052	1.169
5	0.930±0.023	1.178±0.035	1.242±0.019	1.549±0.038	0.712±0.022	0.902±0.021	1.279±0.034	1.015±0.036	1.327±0.036	1.126
6	1.102±0.039	1.396±0.046	1.432±0.024	1.749±0.041	0.998±0.037	0.807±0.025	0.995±0.033	1.021±0.045	1.039±0.027	1.171
7	1.321±0.039	1.505±0.042	1.669±0.022	1.842±0.028	0.887±0.030	0.843±0.019	0.925±0.020	1.176±0.047	1.270±0.045	1.271
8	1.081±0.028	1.242±0.035	1.443±0.020	1.720±0.034	0.817±0.033	0.873±0.018	1.073±0.018	0.909±0.037	1.220±0.028	1.153
9	1.338±0.050	1.530±0.050	1.584±0.029	1.890±0.039	1.231±0.047	0.914±0.031	1.021±0.034	1.082±0.054	0.964±0.032	1.284

Table 28: Cross-experiment stage MAE (100 bio; mean ± 95% CI, $n = 50$). Diagonal = in-distribution. The Mean column is a point estimate (no CI available).

Train\Test	1	2	3	4	5	6	7	8	9	Mean
1	3.541±0.114	5.703±0.167	3.390±0.089	7.590±0.167	5.023±0.234	3.239±0.125	6.329±0.265	5.348±0.158	4.042±0.094	4.912
2	4.094±0.127	4.473±0.170	4.110±0.128	5.351±0.151	6.176±0.254	4.082±0.132	5.472±0.195	7.484±0.291	5.956±0.201	5.244
3	3.901±0.168	5.782±0.168	3.325±0.097	7.361±0.194	4.909±0.236	3.403±0.137	6.257±0.238	5.241±0.146	4.018±0.083	4.911
4	5.067±0.174	4.402±0.122	4.307±0.134	4.208±0.093	7.858±0.284	5.057±0.198	5.896±0.256	7.904±0.284	5.973±0.183	5.630
5	7.563±0.242	11.197±0.296	7.143±0.189	13.345±0.218	3.747±0.165	9.116±0.306	17.644±0.502	3.843±0.105	8.852±0.208	9.161
6	4.719±0.204	6.986±0.269	4.390±0.125	8.960±0.213	5.004±0.181	4.489±0.219	10.314±0.468	5.140±0.164	4.366±0.154	6.041
7	6.118±0.218	6.011±0.181	6.792±0.223	7.196±0.170	6.549±0.209	4.089±0.162	3.933±0.101	5.275±0.141	3.875±0.079	5.538
8	10.524±0.223	14.214±0.298	9.455±0.209	15.392±0.235	6.194±0.301	12.980±0.368	21.659±0.399	5.044±0.258	11.887±0.300	11.928
9	4.488±0.164	7.609±0.223	4.530±0.141	10.580±0.214	4.657±0.196	4.568±0.238	11.621±0.337	4.606±0.123	4.263±0.144	6.325

Table 29: Cross-experiment event position MAE (12 bio; mean ± 95% CI, $n = 50$). Diagonal = in-distribution. The Mean column is a point estimate (no CI available).

Train\Test	1	2	3	4	5	6	7	8	9	Mean
1	1.641±0.085	2.342±0.126	1.952±0.108	2.774±0.105	2.006±0.093	1.747±0.078	2.425±0.121	2.260±0.093	2.415±0.117	2.174
2	1.908±0.097	1.775±0.103	1.943±0.114	2.046±0.116	2.267±0.136	2.100±0.089	2.118±0.112	2.764±0.093	2.663±0.120	2.176
3	1.600±0.104	2.208±0.122	1.452±0.098	2.467±0.117	2.078±0.104	1.829±0.089	2.310±0.117	2.199±0.075	2.066±0.086	2.023
4	1.952±0.103	1.924±0.100	2.085±0.119	1.936±0.114	2.612±0.118	2.268±0.099	2.057±0.083	3.092±0.108	2.913±0.109	2.315
5	2.706±0.142	3.304±0.134	3.277±0.105	3.686±0.103	1.744±0.100	2.722±0.127	3.418±0.111	2.286±0.079	3.043±0.101	2.910
6	2.023±0.103	2.805±0.153	2.632±0.113	3.192±0.106	1.974±0.101	1.841±0.099	2.730±0.145	2.447±0.075	2.338±0.112	2.442
7	2.059±0.111	2.198±0.115	2.486±0.104	2.611±0.105	1.932±0.108	1.940±0.090	2.007±0.120	2.581±0.105	2.597±0.091	2.268
8	2.059±0.091	2.565±0.131	2.336±0.114	2.849±0.099	1.534±0.086	2.309±0.111	2.923±0.100	1.648±0.077	2.571±0.072	2.310
9	1.779±0.093	2.334±0.110	1.878±0.085	2.579±0.127	1.718±0.095	1.554±0.080	2.248±0.125	1.782±0.080	1.579±0.083	1.939

Table 30: Cross-experiment event position MAE (100 bio; mean ± 95% CI, $n = 50$). Diagonal = in-distribution. The Mean column is a point estimate (no CI available).

Train\Test	1	2	3	4	5	6	7	8	9	Mean
1	14.857±0.458	21.397±0.688	13.154±0.426	21.870±0.587	18.002±0.482	15.650±0.435	23.046±0.682	20.526±0.360	13.788±0.476	18.032
2	14.575±0.545	13.253±0.665	16.122±0.371	14.526±0.640	16.935±0.775	16.764±0.486	15.252±0.733	20.769±0.491	19.996±0.416	16.466
3	15.421±0.284	23.394±0.601	11.838±0.450	21.815±0.541	17.237±0.459	17.111±0.375	23.127±0.611	19.103±0.383	17.672±0.303	18.524
4	16.439±0.339	16.375±0.495	11.777±0.520	11.476±0.855	19.693±0.564	18.337±0.408	19.003±0.655	21.216±0.395	18.934±0.295	17.028
5	16.973±0.399	21.132±0.549	16.817±0.357	22.483±0.576	13.819±0.713	18.784±0.377	24.107±0.686	17.518±0.331	23.029±0.346	19.407
6	15.761±0.372	21.281±0.710	17.349±0.279	22.899±0.566	18.529±0.408	13.324±0.595	22.560±0.674	22.854±0.311	18.681±0.433	19.249
7	13.294±0.412	13.090±0.708	14.395±0.631	13.588±0.903	16.178±0.798	14.027±0.538	13.427±1.062	19.281±0.381	17.502±0.505	14.976
8	16.596±0.379	19.944±0.544	15.814±0.363	20.247±0.514	11.658±0.624	16.648±0.404	21.835±0.573	14.712±0.288	20.153±0.322	17.512
9	14.808±0.446	21.687±0.644	15.700±0.400	22.842±0.608	16.250±0.483	11.392±0.537	21.883±0.627	19.260±0.327	14.641±0.533	17.607