Bayesian Event-Based Model for Disease Subtype and Stage Inference

Hongtao Hao HHAO9@WISC.EDU
University of Wisconsin–Madison, USA

Joseph L. Austerweil JOSEPH.AUSTERWEIL@GMAIL.COM
Chiba Institute of Technology, Japan & University of Wisconsin-Madison, USA

for the Alzheimer's Disease Neuroimaging Initiative[*]

Abstract

Chronic diseases often progress differently across patients. Rather than randomly varying, there are typically a small number of subtypes for how a disease progresses across patients. To capture this structured heterogeneity, the Subtype and Stage Inference Event-Based Model (SuStaIn) estimates the number of subtypes, the order of disease progression for each subtype, and assigns each patient to a subtype from primarily cross-sectional data. It has been widely applied to uncover the subtypes of many diseases and inform our understanding of them. But how robust is its performance? In this paper, we develop a principled Bayesian subtype variant of the event-based model (BEBMS) and compare its performance to SuStaIn in a variety of synthetic data experiments with varied levels of model misspecification. BebmS substantially outperforms SuStaIn across ordering, staging, and subtype assignment tasks. Further, we apply BEBMS and SuStaIn to a real-world Alzheimer's data set. We find BEBMS has results that are more consistent with the scientific consensus of Alzheimer's disease progression than SuStaIn.

Keywords: Event-based model, Disease progression, Bayesian methods, Subtypes, Alzheimer's disease

Data and Code Availability BEBMS can be installed by `pip install bebms` (https://github.com/jpcca/bebms_pkg). After installing the BEBMS package, the codes necessary for replicating the experiments using High-Throughput Computing (HTC) on top of a cluster environment are available at https://github.com/hongtaoh/bebms. ADNI data can be requested at https://adni.loni.usc.edu/.

Institutional Review Board (IRB) The IRB at University of Wisconsin-Madison has reviewed and approved the research (#2025-1254).

1. Introduction

Understanding how chronic diseases progress is crucial for early diagnosis, prognosis, and therapeutic development (Jack et al., 2010). These diseases rarely unfold along a single pathway. In the case of Alzheimer's disease, patients often follow distinct trajectories characterized by different progression sequences (Estarellas et al., 2024; Vogel et al., 2021; ten Kate et al., 2018; Poulakis et al., 2022; Jellinger, 2021). Identifying subtypes is critical for uncovering disease mechanisms, and personalizing medicine via improved diagnoses and tailored interventions.

Progression modeling frameworks, such as the Event-Based Model (EBM; Fonteijn et al., 2012), are powerful tools for reconstructing disease progression from cross-sectional data. However, most EBM variants assume a canonical trajectory, limiting their ability to capture heterogeneity. One exception is the Subtype and Stage Inference Event-Based Model (SuStaIn; Young et al., 2018; Aksman et al., 2021). It addressed this limitation by extending EBM to multiple subtypes, and it has since become the de facto standard, applied to a wide range of neurodegenerative diseases in high-profile publications (Estarellas et al., 2024; Young et al., 2018; Vogel et al.,

[*] Data used in preparation of this article were obtained from the Alzheimer's Disease Neuroimaging Initiative (ADNI) database (adni.loni.usc.edu). As such, the investigators within the ADNI contributed to the design and implementation of ADNI and/or provided data but did not participate in analysis or writing of this report. A complete listing of ADNI investigators can be found in Appendix A.

2021). Despite this impact, SuStaIn has not been rigorously evaluated for robustness, particularly under model misspecification (e.g., non-Gaussian biomarker distributions, continuous disease stages, and uneven subtype and stage distributions). We investigated its performance on realistic synthetic datasets with known ground truth and its performance was brittle.

In this paper, we introduce the Bayesian Event-Based Model for Subtyping (BEBMS). Our approach retains the interpretability of EBMs while embedding them in a Bayesian framework, enabling more accurate inference of biomarker orderings, disease staging, and subtype assignment. Across realistic synthetic datasets, and clinical data from ADNI, we show that BEBMS improves biomarker ordering (27%), disease staging (89%), and subtype assignment (56%) over SuStaIn, while reducing runtime.

2. Past Work

The EBM (Fonteijn et al., 2012) formulates disease progression as a sequence of biomarker events, where each biomarker switches from a healthy to a pathological distribution at an unknown position in the sequence. Once the latent disease stage exceeds this position, the biomarker is considered affected. This framework enabled estimation of progression patterns in several neurodegenerative diseases from primarily cross-sectional data (Young et al., 2018; Fonteijn et al., 2012; Oxtoby et al., 2021; Wijeratne et al., 2023; Firth et al., 2020).

Over time, a number of extensions to EBM have been proposed. The Discriminative EBM (DEBM; Venkatraghavan et al., 2019) relaxed the assumption of one ordering by allowing subject-specific variability as random Mallows-distributed noise around the canonical ordering. The Temporal EBM (TEBM; Wijeratne et al., 2023) reformulated progression in continuous rather than discrete time. The Parsimonious EBM (P-EBM; Cs et al., 2025) captures cases where multiple biomarkers become pathological simultaneously. The KDE EBM (Firth et al., 2020) introduced nonparametric likelihoods, enabling the model to estimate data likelihood under non-Gaussian biomarker distributions. The Stage Aware EBM (SA-EBM; Hao et al., 2025) introduced stage distributions and improved inference, resulting in substantial performance improvements over the original EBM, KDE EBM, and DEBM. Despite these advances, all of these models assume a single central ordering, and therefore cannot capture disease heterogeneity across patients.

SuStaIn (Young et al., 2018) addressed this limitation by extending EBM to incorporate multiple progression patterns, enabling subtype inference. SuStaIn has been widely adopted across 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), and is supported by an open-source implementation (Aksman et al., 2021). As SuStaIn is computationally demanding, scalable variants for high-dimensional data (s-SuStaIn; Tandon et al., 2024) and incomplete data (Estarellas et al., 2024) have been developed.

While SuStaIn and its extensions represent a major step forward in capturing disease heterogeneity, they also have important limitations. First, they typically assume a uniform prior over disease stages, even though later stages are underrepresented in real-world cohorts such as ADNI (Donohue et al., 2014). Second, they rely on static estimates of biomarker distribution parameters, which remain fixed during inference of subtype orderings and may bias results when the true ordering and stage distribution are unknown. Despite its widespread adoption, the robustness of SuStaIn has rarely been systematically evaluated on synthetic datasets with known ground truth. As a result, the model's robustness under misspecification remains unclear.

Building on prior work, we introduce the Bayesian Event-Based Model for Subtyping (BEBMS). By embedding EBMs in a Bayesian framework, BEBMS provides more accurate estimation of disease subtypes, biomarker orderings, and patient stages, while also reducing runtime, as demonstrated in both synthetic and ADNI experiments.

3. Method

3.1. Model Specification

BEBMS models disease progression as a sequence of biomarker events. Each of the N biomarkers can exist in either a "pre-event" (healthy) or "post-event" (pathological) state. A participant j's disease state is defined by their stage k_j. We define $k_j = -1$ for a healthy participant (no events), and $k_j \in [0, 1, ..., N-1]$ for a progressing participant, where k_j corresponds to the highest rank of the event that has occurred.

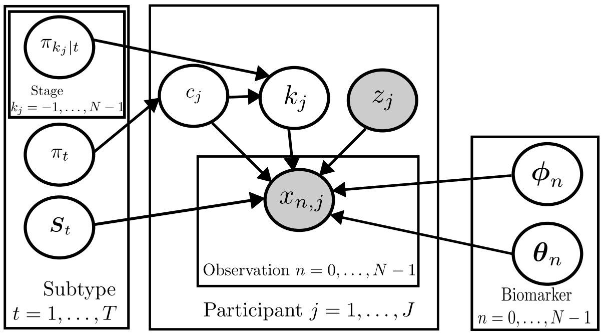

Figure 1: BEBMS as a graphical model.

We model T disease subtypes, each defined by a unique sequence of biomarker events. Let $\boldsymbol{S}$ be a $T \times N$ matrix where $S_{t,n}$ is the 0-based **rank** of biomarker n in the progression sequence for subtype t. We also write $S(t,i)$ for the **biomarker** at position i in subtype t, so $S(t,i) = n$ if and only if $S_{t,n} = i$. A biomarker n is considered post-event for a participant in subtype t and stage k_j if $k_j \geq S_{t,n}$. Let c_j denote participant j's subtype. π_t and $\pi_{k_j|t}$ reflect the probabilities of subtype t ($c_j \sim \pi_t$) and of a participant being in disease stage k_j given they belong to subtype t ($k_j|c_j \sim \pi_{t|c_j}$, respectively. Figure 1 presents the approach as a graphical model.

Biomarker measurements in both states are modeled with Gaussian distributions. **We assume a shared parameterization for biomarkers across subtypes**:

$$x_{j,n} \sim \begin{cases} \mathcal{N}(\phi_{n,\mu}, \phi_{n,\sigma}^2), & \text{if pre-event,} \\ \mathcal{N}(\theta_{n,\mu}, \theta_{n,\sigma}^2), & \text{if post-event.} \end{cases} \tag{1}$$

with ϕ and θ denoting healthy and afflicted distribution parameters, respectively. $\boldsymbol{X}$ is the whole dataset, and $\boldsymbol{X}_j$ is the biomarker measurements of a specific participant. We use $x_{j,n}$ to denote the value of biomarker n in participant j.

The model is a mixture over subtypes and stages. For a healthy participant ($z_j = 0$, i.e., $k_j = -1$), the data likelihood is

$$p(\boldsymbol{X}_j \mid \boldsymbol{S}, z_j = 0) = \prod_{n=0}^{N-1} p(x_{j,n} \mid \phi_n) \tag{2}$$

For a diseased participant ($z_j = 1$), the likelihood is marginalized over all subtypes t and all possible disease stages k_j for participant j, weighted by their respective priors π_t and $\pi_{k_j|t}$:

$$p(\boldsymbol{X}_j \mid \boldsymbol{S}, z_j = 1) = \sum_{t=1}^{T} \pi_t \sum_{k_j=0}^{N-1} \pi_{k_j|t}\, p(\boldsymbol{X}_j \mid \boldsymbol{S}_t, z_j, k_j) \tag{3}$$

where $p(\boldsymbol{X}_j \mid \boldsymbol{S}_t, z_j = 1, k_j)$ denotes the likelihood of participant j given they are in subtype t at stage k_j:

$$\begin{aligned} p(\boldsymbol{X}_j \mid \boldsymbol{S}_t, z_j{=}1, k_j) = &\underbrace{\prod_{i=0}^{k_j} p(x_{j,S(t,i)} \mid \theta_{S(t,i)})}_{\text{post-event}} \\ &\times \underbrace{\prod_{i=k_j+1}^{N-1} p(x_{j,S(t,i)} \mid \phi_{S(t,i)})}_{\text{pre-event}}. \end{aligned} \tag{4}$$

The total data likelihood across all J participants is the product of their individual likelihoods:

$$P(\boldsymbol{X} \mid \boldsymbol{S}) = \prod_{j=1}^{J} P(\boldsymbol{X}_j \mid \boldsymbol{S}, z_j) \tag{5}$$

We place Dirichlet priors on the subtype prior $\boldsymbol{\pi} \in \mathbb{R}^T \sim \mathrm{Dir}(\boldsymbol{\alpha})$ and the stage prior $\boldsymbol{\pi}_{\cdot|t} \in \mathbb{R}^{T \times N} \sim \mathrm{Dir}(\boldsymbol{\alpha}_{\cdot|t})$, with weakly informative priors (Gelman et al., 2017) of 1.0 for $\boldsymbol{\alpha}$ and $\boldsymbol{\alpha}_{\cdot|t}$.

3.2. Inference Procedure

The biomarker parameters $(\boldsymbol{\theta}, \boldsymbol{\phi})$ are initialized by K-Means using controls for pre-event and progressing participants for post-event. After clustering, the cluster containing the majority of controls is labeled pre-event ($\boldsymbol{\phi}$) and the other post-event ($\boldsymbol{\theta}$). K-Means estimates are immediately refined with weighted conjugate updates. We assume a Normal-Inverse Gamma (NIG) conjugate prior on the unknown mean and variance, $(\mu, \sigma^2) \sim \mathrm{NIG}(m_0, n_0, s_0^2, \nu_0)$. The posterior is also a NIG distribution with updated parameters.

We set $n_0 = 1, \nu_0 = 1$, in the spirit of weakly informative priors (Gelman et al., 2017). m_0 and s_0^2 are the raw mean and variance of each cluster, respectively. The updates use a soft assignment of each measurement to the pre- and post-event clusters. Each cluster has a weight vector ($\boldsymbol{w} \in \mathbb{R}^J$), indicating the posterior probability of each subject belonging to the cluster. For initialization, we assign

1 to all entries of $\boldsymbol{w}$. See Appendix B for how we update distribution parameters using conjugate priors.

A key innovation of BEBMS is the iterative estimation of distribution parameters $(\boldsymbol{\theta}, \boldsymbol{\phi})$, and stage $(\boldsymbol{\pi}_{\cdot|t})$ and subtype $(\boldsymbol{\pi})$ priors through a Metropolis-Hastings Markov chain Monte Carlo (MCMC) sampler (See Algorithm 1 in Appendix C).

Specifically, we compute the intermediate stage $(\tilde{P}_{\text{stage}} \in \mathbb{R}^{J\times T\times N})$ and subtype $(\tilde{P}_{\text{subtype}} \in \mathbb{R}^{J\times T})$ posteriors as follows:

$$P_{\text{stage}}(k \mid j, t) = \frac{\pi_{k|t} \cdot p(\boldsymbol{X}_j \mid \boldsymbol{S}_t, k, \boldsymbol{\theta}, \boldsymbol{\phi})}{\sum_{k'=0}^{N-1} \pi_{k'|t} \cdot p(\boldsymbol{X}_j \mid \boldsymbol{S}_t, k', \boldsymbol{\theta}, \boldsymbol{\phi})}$$

$$P_{\text{subtype}}(t \mid j) = \frac{\pi_t \sum_{k=0}^{N-1} P_{\text{stage}}(k \mid j, t)}{\sum_{t'=1}^{T} \pi_{t'} \sum_{k=0}^{N-1} \pi_{k|t'} P_{\text{stage}}(k \mid j, t')}$$

In updating distribution parameters, the weights $(\boldsymbol{w})$ are the marginalized probabilities of being in pre/post-event states across all subtypes and stages:

$$w_{j,n,\theta} = \sum_{t=1}^{T} \sum_{k=0}^{N-1} \mathbf{1}_{\{k \geq S_{t,n}\}} P_{\text{stage}}(k \mid j, t) P_{\text{subtype}}(t \mid j),$$

$$w_{j,n,\phi} = 1 - w_{j,n,\theta}$$

For healthy participants, assignments are fixed: $w_{j,n,\phi} = 1, w_{j,n,\theta} = 0$ for all biomarkers. If a proposal is accepted, we update subtype and stage priors with posterior counts:

$$\boldsymbol{\pi} \sim \text{Dir}\left(\boldsymbol{\alpha} + \sum_{j=1}^{J} P_{\text{subtype}}(t \mid j)\right)$$

$$\boldsymbol{\pi}_{\cdot|t} \sim \text{Dir}\left(\boldsymbol{\alpha}_{\cdot|t} + \sum_{j=1}^{J} P_{\text{stage}}(k \mid j, t) P_{\text{subtype}}(t \mid j)\right)$$

3.3. Model Selection

The above inference procedure assumes we know the number of subtypes T. In real-world scenarios, however, this information is missing. Following SuStaIn (Aksman et al., 2021; Young et al., 2018), we applied K-fold cross-validation to find the optimal T. For each candidate T, we trained the model on the training folds and obtained the out-of-sample log-likelihood on the held-out fold. We then aggregated log-likelihoods across folds to compute a cross-validation information criterion (CVIC), and selected the T with the lowest CVIC score. When multiple T have similar CVIC scores (difference less than 6), we chose the smallest T within the group. Specifically, we performed stratified K-fold cross-validation to maintain the proportion of progressing/healthy in each fold. CVIC score is computed as:

$$\text{CVIC} = 2 \cdot \sum_{i=1}^{T} i\text{th fold log likelihood} \tag{6}$$

We chose the threshold of 6 for consistency with SuStaIn's procedure (Aksman et al., 2021).

4. Model Evaluation

We use SuStaIn as the baseline for evaluation because it is the only publicly available implementation to reconstruct subtypes of disease progression using cross-sectional data. Missing data SuStaIn (Estarellas et al., 2024) referred to SuStaIn as the code source and s-SuStaIn (Tandon et al., 2024) does not have publicly available code. For the single-subtype case, BEBMS is essentially the same as the Conjugate Prior variant of SA-EBM (Hao et al., 2025), which has improved performance compared to DEBM (Venkatraghavan et al., 2019), KDE-EBM (Firth et al., 2020) and UCL GMM (Firth et al., 2020). Thus, SuStaIn is the only benchmark algorithm in this study.

Our evaluation relies on both synthetic and real-world datasets. The real-world data comes from the Alzheimer's Disease Neuroimaging Initiative (ADNI, Mueller et al., 2005). The ADNI was launched in 2003 as a public-private partnership, led by Principal Investigator Michael W. Weiner, MD. The primary goal of ADNI has been to test whether serial magnetic resonance imaging (MRI), positron emission tomography (PET), other biological markers, and clinical and neuropsychological assessment can be combined to measure the progression of mild cognitive impairment (MCI) and early Alzheimer's disease (AD).

Our analysis was based on the `adnimerge` table (updated on September 7, 2023) from the Alzheimer's Disease Cooperative Study data system. We restrict the study cohort to baseline visits from participants with a diagnosis of CN, Early and Late MCI, or AD. The biomarker selection consists of 12 measures commonly used in the field (Cs et al., 2025; Young et al., 2014; Archetti et al., 2019). These biomarkers cover a wide range of categories: cognitive assessment, Cerebrospinal fluid (CSF) markers and MRI-derived measurements of the brain regions. See Table 1 and 2 (in Appendix D and E, respectively) for details about these biomarkers.

For the brain regions MRI measurements, following best practices, we applied intracranial volume (ICV) normalization because people's brain sizes vary. We excluded participants with missing values from any of these 12 biomarkers, and de-duplicated the final dataset. The final version of ADNI fitting these criteria had 726 participants, distributed from three ADNI protocols: ADNI (275, 37.9%), ADNIGO (76, 10.5%) and ADNI2 (375, 51.7%). There were 413 (56.9%) men and 313 (43.1%) women, diagnosed as AD (153, 21.1%), late MCI (236, 32.5%), Control (155, 21.3%) and early MCI (182, 25.1%).

We obtained the biomarker distribution parameters for synthetic datasets based on ADNI. To provide SuStaIn with the best opportunity to perform well, we applied their method (Gaussian Mixture Model) on the processed ADNI dataset to obtain $\boldsymbol{\theta}$ and $\boldsymbol{\phi}$.

We employed two models to generate synthetic datasets: the EBM and the Sigmoid model (defined later). First, we uniformly pick a number from 1 to 5 (inclusive) for the number of subtypes. We chose [1,5] because [2,5] is the most likely range for AD according to the literature (Young et al., 2018; Estarellas et al., 2024; ten Kate et al., 2018; Jellinger, 2021), and 1 to account for no subtypes. We then randomly pick a dispersion parameter between 0.01 and 0.5, and use the Top-K Mallow's Model implemented by Boujaada et al. (2022) to get the event sequence for each subtype based on a random permutation of all the 12 biomarkers. We chose [0.01, 0.5] because it allows the generated subtypes' sequences to have varied agreement covering the whole range of [0,1] as measured by Kendall's W (See Appendix H). We made sure no two subtypes share the same event sequence, and there were at least ten progressing participants in each subtype.

Healthy participants do not belong to any subtype. We used a Dirichlet-Multinomial (DM) distribution to assign subtypes to progressing participants. The prior for the DM distribution is selected uniformly at random from 0.1, 2, 5, and 20. This allows both sparse and uniform distributions for participants' subtype assignments. DM distribution is also employed to generate disease stages for progressing participants, but the prior setting depends on experimental configurations (See below and Appendix F).

Given the event sequence ($\boldsymbol{S}_t$) of subtype t, a biomarker n, and a participant j with diagnosis $z_j \in \{0,1\}$ and disease stage k_j, the generative model of EBM defines the biomarker measurement of $x_{j,n} \mid \boldsymbol{S}_t, k_j, \boldsymbol{\theta}_n, \boldsymbol{\phi}_n, z_j$ as:

$$x_{j,n} \sim \begin{cases} p(x_{n,j} \mid \boldsymbol{\theta}_n), & z_j = 1,\ S_{t,n} \leq k_j, \\ p(x_{n,j} \mid \boldsymbol{\phi}_n), & z_j = 1,\ S_{t,n} > k_j, \\ p(x_{n,j} \mid \boldsymbol{\phi}_n), & z_j = 0. \end{cases} \tag{7}$$

The Sigmoid model, adapted from Young et al. (2015) and Venkatraghavan et al. (2019), is motivated by the biomarker cascade hypothesis of Jack et al. (2010), which postulates that AD biomarkers follow sigmoid-shaped trajectories: slow to change in early stages, accelerating during symptomatic onset, and plateauing later. As in EBM, measurements of healthy individuals are drawn from normal distributions. Formally:

$$x_{j,n} \sim \mathcal{N}(\mu_{n,\phi}, \sigma^2_{n,\phi})$$

The measurements of progressing participants monotonically deviate from the healthy state:

$$x_{n,j} \sim \mathcal{N}(\mu_{n,\phi}, \sigma^2_{n,\phi}) + \frac{(-1)^{I_n} R_n}{1 + e^{-\rho_n (k_j - \xi_n)}}$$

where $I_n \sim$ Bernoulli (0.5) randomly flips the direction of the deviation, $R_n = \mu_{n,\theta} - \mu_{n,\phi}$ controls the range of the measurements, and $\rho_n = \max\left(1, \frac{|R_n|}{\sqrt{\sigma^2_{n,\theta} + \sigma^2_{n,\phi}}}\right)$ sets the slope.

SuStaIn (Aksman et al., 2021) has two variants: GMM (gaussian mixture model) and KDE. GMM assumes that biomarker measurements follow Gaussian distributions whereas KDE does not. Both BEBMS and SuStaIn rely on three key assumptions: (1) biomarker measurements follow Gaussian distributions (except for SuStaIn KDE); (2) disease stages are ordinal; and (3) biomarker events occur in an ordinal sequence with approximately even spacing. To evaluate robustness, we systematically relaxed each assumption. Non-Gaussianity was tested both by generating EBM datasets with non-normal biomarker distributions (Table 2 in Appendix E has more details) and by using the Sigmoid model. Continuous disease stages were introduced to test violations of the second assumption. We examined violations of the third assumption by introducing uneven event spacing across biomarkers. Experimental results show that the last test revealed a shared limitation of the underlying modeling paradigm. See Appendix I for more details. Details about experimental specifications are available in Appendix F. In Appendix G, we have plotted the theoretical and empirical distributions of all twelve biomarkers (Fig. 4).

4.1. Experiment Setup

For each subtype t, when the BEBMS infers the event sequences $\boldsymbol{S}_t$, i.e., doing the **ordering task**, it assumes knowledge of the diagnosis labels, i.e., whether healthy or progressing. In the **subtyping task** and the **staging task**, when inferring the most likely subtype and disease stage for each participant, BEBMS is blind to the diagnosis labels. It also discards the stage and subtype priors inferred by the model and relies only on the estimated biomarker distribution parameters $\boldsymbol{\theta}, \boldsymbol{\phi}$.

SuStaIn's methodology differs, as it leverages diagnosis labels only to estimate the biomarker distribution parameters, excluding them from subsequent inference tasks. While this design aims for robustness against diagnostic uncertainty, it introduces a potential issue: if the labels are unreliable, their use in any estimation step risks biasing the model's core parameters. To enable a direct comparison, we developed BEBMS (BLIND). This variant strictly mirrors SuStaIn's philosophy, utilizing diagnosis labels only for the initialization of biomarker distribution parameters and not in any of the three tasks mentioned above.

We used four different total participant sizes: $J = 300, 500, 1000, 1500$ and three different healthy ratios: $R = 0.25, 0.5, 0.75$. For each $J-R$ pair, we generated 10 datasets. With eleven experiments, we have 1,320 datasets in total. BEBMS and BEBMS (BLIND) use 10,000 MCMC iterations with 500 burn-in and no thinning. Per SuStaIn (Aksman et al., 2021) recommendation, we applied 25 parallel start points for their E-M algorithm and 100,000 MCMC iterations. All other settings are the default of SuStaIn. We applied both the GMM and the KDE version of SuStaIn. For the estimation of T, in cross-validation, we tested $T = 1$ to $T = 5$ (inclusive). We used the maximum-likelihood ordering (the E-M solution) from SuStaIn as the estimated subtype progression pattern, and the `ml_subtype` and `ml_stage` outputs from `run_sustain_algorithm` as the subject-level subtype and stage assignments, respectively. Posterior MCMC samples (`samples_sequence`) were used only to quantify uncertainty in the ordering results.

4.2. Evaluation Metrics

For the ordering task, we computed a cost matrix of normalized Kendall's τ distances and applied the Hungarian algorithm (Kuhn, 1955; Munkres, 1957) to optimally match estimated and true sequences, reporting the mean distance across matched pairs. Subtype assignment accuracy was measured with the Adjusted Rand Index (Hubert and Arabie, 1985) between the true and inferred subtype labels. Note that the performance on the subtyping task is only relevant when the ground truth has more than 1 subtype. As staging error is confounded with subtype accuracy, we evaluated the staging accuracy by only reporting the mean estimated stage for control participants with the ground truth of 0. We reported the Mean Absolute Error (MAE) as the accuracy for the estimation of the number of subtypes. To compare the computational efficiency, we also report the runtime for each dataset. Due to the computational cost of cross-validation, we only report the full runtime of model selection for Experiment 1.

5. Results

5.1. Synthetic Datasets

We conducted all experiments on the CHTC cluster at the University of Wisconsin-Madison (Center for High Throughput Computing, 2006). Across the 1,320 synthetic datasets, 2 timed out. SuStaIn (KDE) failed to process an additional 152 due to an "IndexError", highlighting its fragility in practice. As shown in Figs. 2, 6, 7 and 8, BEBMS consistently outperforms SuStaIn across ordering, subtyping, and staging tasks, and is faster. Unless otherwise mentioned, the results refer to Exp. 1-9.

Ordering: BEBMS and its blind variant achieved Kendall's τ distances of 0.24 ± 0.01 and 0.29 ± 0.01, respectively—substantially lower (better) than SuStaIn KDE (0.33 ± 0.01) and SuStaIn GMM (0.39 ± 0.01). For reference, random guessing yielded 0.45 ± 0.00. Variability is 95% CI. SuStaIn GMM performed particularly poorly when biomarker measurements deviated from Gaussian assumptions and also showed a marked decline with increasing proportions of healthy participants, even under normally distributed data. By contrast, BEBMS—although also

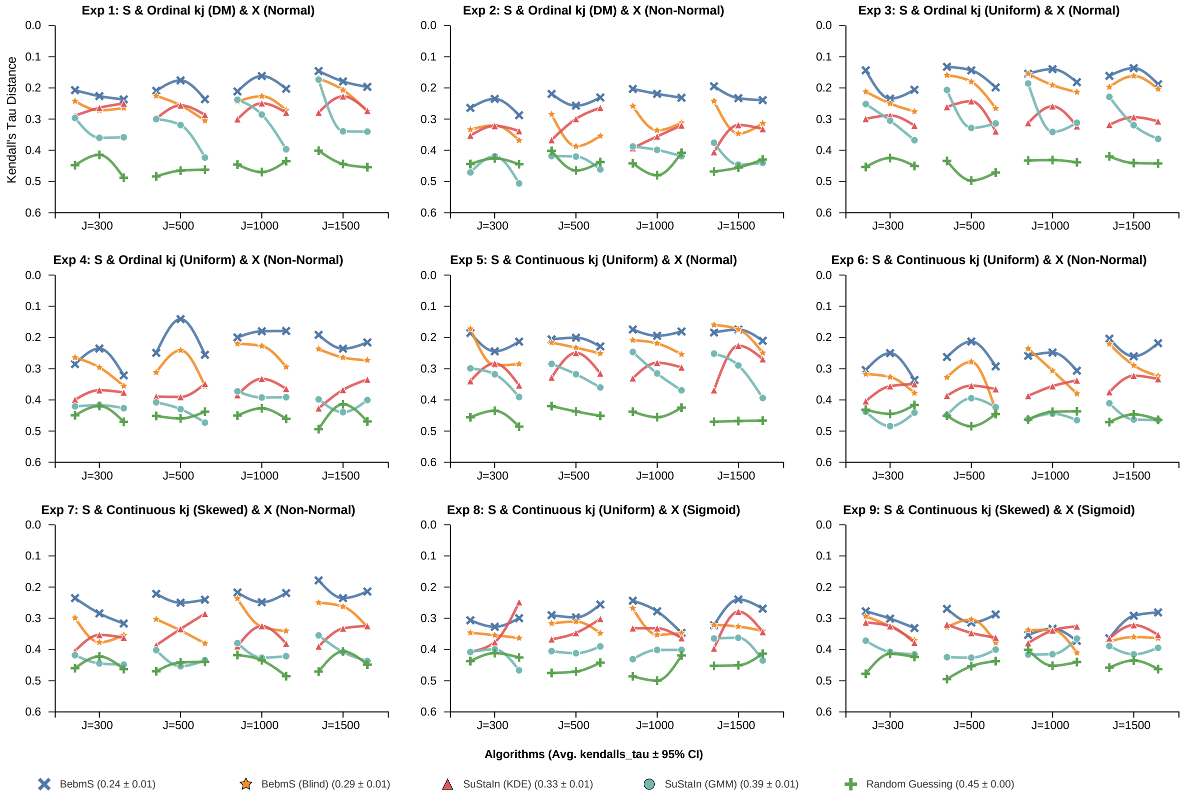

Figure 2: Normalized Kendall's τ across all nine synthetic experiments. Each panel corresponds to an experiment; within each panel, participant sizes (J) are shown across columns, and within each column three healthy ratios ($R = 0.25, 0.5, 0.75$) are displayed from left to right. BEBMS reduced ordering error by 27.3% relative to SuStaIn, with BEBMS (Blind) performing nearly identically. SuStaIn results were consistently lower, with margins narrowing under model misspecification (Experiments 8–9). Performance was largely insensitive to participant size and healthy ratio.

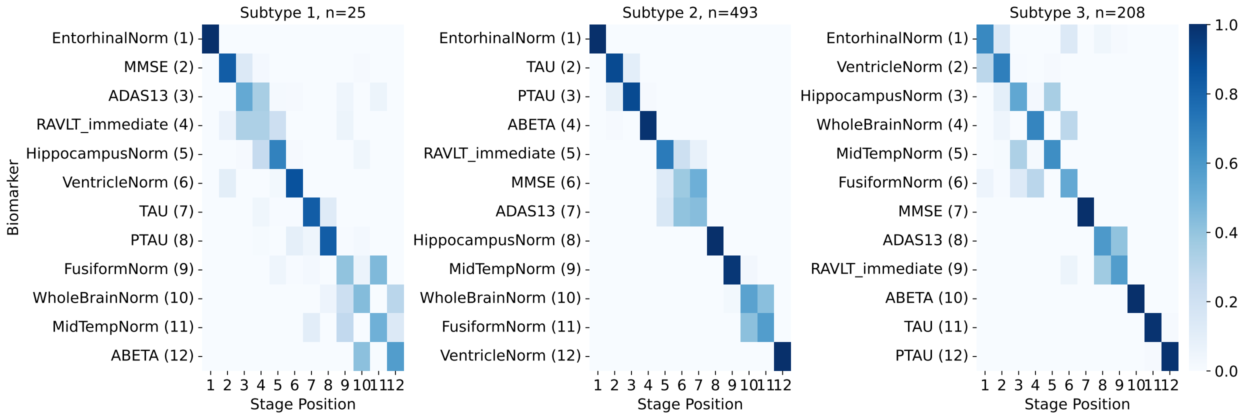

Figure 3: BEBMS ADNI ordering. All subtypes begin in the entorhinal region, with Subtype 1 showing early cognitive decline, Subtype 2 early CSF changes, and Subtype 3 early neurodegeneration.

assuming Gaussian distributions—was robust to non-Gaussian misspecification, and outperformed SuStaIn KDE in most cases. Both BEBMS and SuStaIn struggled in Experiment 9, where the ordinal assumption of disease stages was violated and data were generated from the sigmoid framework. Regarding participant size, performance saturated around 300 participants. BEBMS was also robust to varying healthy ratios: holding the total number of participants fixed, increasing the proportion of controls had only a modest effect on ordering accuracy.

Subtyping. As shown in Fig. 6 (Appendix I), subtyping proved to be a challenging task overall. BEBMS achieved an Adjusted Rand Index (ARI) 0.25 ± 0.02, followed by BEBMS (BLIND) (0.24 ± 0.02). SuStaIn GMM and KDE obtained 0.16 ± 0.02 and 0.13 ± 0.02, respectively. Random assignment yielded 0.00. Subtype accuracy appeared insensitive to sample size, though increasing the healthy ratio impaired performance.

Staging. In staging tasks (Appendix I, Fig. 7), BEBMS assigned average stages of 0.16 ± 0.03 to control participants, while BEBMS (BLIND) achieved 0.62 ± 0.09. By comparison, SuStaIn KDE and GMM performed substantially worse, assigning average stages of 1.45 ± 0.30 and 3.03 ± 0.24, respectively. Importantly, SuStaIn's staging accuracy degraded most severely when datasets contained fewer controls (i.e., smaller healthy ratios).

Runtime. BEBMS achieved better performance and was faster (Appendix I, Fig. 8), with an average runtime across all datasets of 2.37 ± 0.26 minutes, followed by BEBMS (BLIND) (4.04 ± 0.43 minutes). SuStaIn KDE (6.57 ± 0.58 minutes) and SuStaIn GMM (6.88 ± 0.61 minutes) took about twice as long.

Subtype number estimation. In estimating the optimal number of subtypes (Appendix I, Fig. 9), BEBMS (1.08 ± 0.14) performed marginally better than SuStaIn GMM (1.16 ± 0.20), followed by BEBMS (BLIND) (1.27 ± 0.16) and SuStaIn KDE (1.92 ± 0.23) which had the same performance as random guessing (1.92 ± 0.33). SuStaIn tends to overfit by identifying more subtypes than the ground truth, whereas both variants of BEBMS exhibit symmetric relative-error distributions centered at zero (See Fig. 11 in Appendix I). Although BEBMS and SuStaIn GMM achieved comparable accuracy, BEBMS was faster (Appendix I, Fig. 10), requiring 40.11 ± 15.28 minutes on average compared to 63.68 ± 15.11 minutes for SuStaIn GMM.

Stress-test experiments (Exp. 10-11). BEBMS outperformed SuStaIn across all tasks in the two stress-test experiments. Interestingly, the best-performing variants of each method achieved higher subtyping (Fig. 13) and staging (Fig. 14) accuracy than in the standard experiments (Exp. 1–9). However, for ordering (Fig. 12), the performance of both methods degraded to the level of random guessing.

5.2. ADNI

Cross-validation and setup. We performed 5-fold cross-validation on the ADNI dataset to select the number of subtypes, testing values from 1 to 6 (the upper bound allowing us to assess potential overfitting). All other settings matched those used in the synthetic experiments. SuStaIn-GMM selected six subtypes, whereas BEBMS selected three (see explanations, Table 3 and Figures 15–16 in Appendix J). SuStaIn-KDE failed due to a singular matrix error. For both algorithms, we then ran ten replications with different random seeds and retained the solution with the highest data likelihood. For BEBMS, each run consisted of 20,000 MCMC iterations (200 burn-in, no thinning). Fig. 3 shows the inferred progression patterns for the three BEBMS subtypes based on the final 18,000 iterations. The trace plot indicates stable convergence (Fig. 17, Appendix J).

In the results below, we present the inferred ordering, subtype assignments, and disease staging on ADNI. Biomarkers are grouped as CSF amyloid (A: $A\beta_{1-42}$), CSF tau (T: TAU, PTAU), cognition (C: ABETA, MMSE, ADAS13), and neurodegeneration (N: all remaining biomarkers; see Fig. 3).

Ordering. BEBMS identified three subtypes:

1. N → C → N → T → N → A (25, 3.4%)
2. N → T → A → C → N (493, 67.9%)
3. N → C → A → T (208, 28.7%)

SuStaIn GMM identified six subtypes (See Fig. 20 in Appendix J):

1. A → T → C → N → C → N (342, 47.1%)
2. A → N → C → N → C → T (124, 17.1%)
3. C → N → A → N → T (148, 20.4%)
4. N → C → A → T (54, 7.4%)
5. A → C → T → N (46, 6.3%)

6. T → N → C → N → C → N → A → N (12, 1.7%)

Table 4 and 5 in Appendix J have more detailed data of the distributions of subtypes over participants of different diagnoses.

Staging. BEBMS scored 1.10 average stage, compared to 2.48 for SuStaIn GMM for healthy patients. Fig. 18 and 19 in Appendix J have more details.

6. Discussion

The ability of SuStaIn (Young et al., 2018) to estimate subtypes of diseases from cross-sectional data has been a breakthrough for data-driven medical research. However, the robustness of its performance is unclear. In this paper, we presented a Bayesian variant of the EBM with subtypes (the BEBMS). Across a range of synthetic data experiments, we found that the BEBMS performed better in all tasks: ordering, subtyping, staging, and estimating the number of subtypes, while being computationally more efficient. BEBMS can handle data with missing entries and shows great potential in scaling to high-dimensional data (See Appendix K). The blind variant of BEBMS has a slightly lower performance while also being computationally more demanding; domain expertise is needed to decide which version of our model to use, depending on the uncertainties in the diagnosis labels.

BEBMS's improved performance has important implications for clinical and research practice. Our model's advantages can directly enhance clinical trial enrichment and patient stratification in precision medicine. Also, BEBMS remains robust across variation in the proportion of healthy participants, and the performance of all methods saturates at around 300. In many real-world settings—especially for new or rare diseases—recruiting large patient cohorts is difficult. Our results suggest that (1) cohorts larger than 300 may not be necessary, and (2) when using BEBMS, increasing the proportion of healthy controls—which is typically far easier to recruit—has only a modest effect on performance.

The results on the real-world ADNI dataset show that BEBMS yields subtype patterns that align more closely with the current scientific consensus on Alzheimer's disease progression than those identified by SuStaIn. Neuropathological evidence indicates three major AD subtypes (Murray et al., 2011). The most common subtype, Typical AD (TAD, ∼75%), generally follows the ATNC sequence—amyloid (A) abnormality preceding tau (T), then neurodegeneration (N), and cognitive decline (C) (Jack Jr et al., 2024), although only about one-third of patients strictly follow this ordering (Mendes et al., 2025). In contrast, Limbic-predominant AD (LPAD, ∼14%) is characterized by early and severe involvement of the hippocampus and medial temporal cortex, with relatively limited neocortical involvement. Finally, Hippocampal-sparing AD (HSAD, ∼11%) exhibits pronounced neocortical pathology—particularly in parietal and frontal association areas—while the hippocampus remains comparatively preserved.

Regarding the subtype patterns, BEBMS Subtype 2 (CSF-first, 67.9%) corresponds to Typical AD (TAD), as CSF biomarkers become abnormal earliest; this aligns with SuStaIn Subtype 1 (47.1%) and possibly Subtype 6 (1.7%). BEBMS Subtype 3 (entorhinal/hippocampal-first, 28.7%) corresponds to Limbic-predominant AD (LPAD), characterized by early medial temporal involvement; SuStaIn Subtypes 4 (7.4%) and potentially Subtype 2 (17.1%) reflect a similar pattern. Finally, BEBMS Subtype 1 (neocortical/cognitive-first with delayed hippocampal involvement, 3.4%) corresponds to Hippocampal-sparing AD (HSAD), which matches SuStaIn Subtypes 3 (20.4%) and 5 (6.3%).

Overall, BEBMS yields subtype proportions and progression patterns that more closely match the canonical findings reported in (Murray et al., 2011), and also produces cleaner clustering of jointly progressing biomarkers. However, BEBMS is not flawless: all subtypes show early entorhinal involvement, which may reflect mild overfitting or a strong entorhinal signal in ADNI. We caution against overinterpretation and believe our results should be treated as converging evidence to be assessed by domain experts, and not as definitive.

Our study has several limitations. First, both BEBMS and SuStaIn struggle when event times of when a disease progresses were assumed to be continuous rather than ordinal. If the event times of a disease are not relatively evenly spaced, then neither model is a good choice. Future work should explore continuous-time disease progression, e.g., TEBM (Wijeratne et al., 2023), with subtypes using longitudinal data. Second, the advantages of BEBMS are testing on only one real-world dataset. More work should be done to validate BEBMS on other datasets.

Acknowledgments

We thank the CHTC at the University of Wisconsin-Madison for computing support. JLA was funded by the Japan Probabilistic Computing Consortium Association (JPCCA).

Data collection and sharing for this project was funded by the Alzheimer's Disease Neuroimaging Initiative (ADNI) (National Institutes of Health Grant U01 AG024904) and DOD ADNI (Department of Defense award number W81XWH-12-2-0012). ADNI is funded by the National Institute on Aging, the National Institute of Biomedical Imaging and Bioengineering, and through generous contributions from the following: AbbVie, Alzheimer's Association; Alzheimer's Drug Discovery Foundation; Araclon Biotech; BioClinica, Inc.; Biogen; Bristol-Myers Squibb Company; CereSpir, Inc.; Cogstate; Eisai Inc.; Elan Pharmaceuticals, Inc.; Eli Lilly and Company; EuroImmun; F. Hoffmann-La Roche Ltd and its affiliated company Genentech, Inc.; Fujirebio; GE Healthcare; IXICO Ltd.; Janssen Alzheimer Immunotherapy Research & Development, LLC.; Johnson & Johnson Pharmaceutical Research & Development LLC.; Lumosity; Lundbeck; Merck & Co., Inc.; Meso Scale Diagnostics, LLC.; NeuroRx Research; Neurotrack Technologies; Novartis Pharmaceuticals Corporation; Pfizer Inc.; Piramal Imaging; Servier; Takeda Pharmaceutical Company; and Transition Therapeutics. The Canadian Institutes of Health Research is providing funds to support ADNI clinical sites in Canada. Private sector contributions are facilitated by the Foundation for the National Institutes of Health (www.fnih.org). The grantee organization is the Northern California Institute for Research and Education, and the study is coordinated by the Alzheimer's Therapeutic Research Institute at the University of Southern California. ADNI data are disseminated by the Laboratory for Neuro Imaging at the University of Southern California.

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.

D. Archetti, S. Ingala, V. Venkatraghavan, V. Wottschel, A. L. Young, F. Barkhof, D. C. Alexander, N. P. Oxtoby, G. B. Frisoni, A. Redolfi, for the Alzheimer's Disease Neuroimaging Initiative, and for EuroPOND Consortium. Multi-study validation of data-driven disease progression models to characterize evolution of biomarkers in alzheimer's disease. *NeuroImage: Clinical*, 24: 101954, 2019.

Ahmed Boujaada, Fabien Collas, and Ekhine Irurozki. top-k-mallows. https://github.com/ekhiru/top-k-mallows, 2022. Accessed: 2025-09-06.

Center for High Throughput Computing. Center for high throughput computing, 2006. URL https://chtc.cs.wisc.edu/.

Parker Cs, NP Oxtoby, DC Alexander, H Zhang, AL Young the Alzheimer's Disease Neuroimaging Initiative, et al. Parsimonious EBM: generalising the event-based model of disease progression for simultaneous events. *NeuroImage*, page 121162, 2025.

Michael C Donohue, Hélène Jacqmin-Gadda, Mélanie Le Goff, Ronald G Thomas, Rema Raman, Anthony C Gamst, Laurel A Beckett, Clifford R Jack Jr, Michael W Weiner, Jean-François Dartigues, et al. Estimating long-term multivariate progression from short-term data. *Alzheimer's & Dementia*, 10:S400–S410, 2014.

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.

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.

Andrew Gelman, Daniel Simpson, and Michael Betancourt. The prior can often only be understood in the context of the likelihood. *Entropy*, 19(10): 555, 2017.

Hongtao Hao, Vivek Prabhakaran, Veena A Nair, Nagesh Adluru, and Joseph L. Austerweil. Stage-aware event-based modeling (SA-EBM) for disease progression. In Monica Agrawal, Kaivalya Deshpande, Matthew Engelhard, Shalmali Joshi, Shengpu Tang, and Iñigo Urteaga, editors, *Proceedings of the 10th Machine Learning for Healthcare Conference*, volume 298 of *Proceedings of Machine Learning Research*. PMLR, 15–16 Aug 2025. URL https://proceedings.mlr.press/v298/hao25a.html.

Lawrence Hubert and Phipps Arabie. Comparing partitions. *Journal of classification*, 2(1):193–218, 1985.

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, 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.

Kurt A Jellinger. Pathobiological subtypes of alzheimer disease. *Dementia and Geriatric Cognitive Disorders*, 49(4):321–333, 2021.

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

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.

Augusto J Mendes, Federica Ribaldi, Michela Pievani, Cecilia Boccalini, Valentina Garibotto, Giovanni B Frisoni, and Alzheimer's Disease Neuroimaging Initiative. Validating the amyloid cascade through the revised criteria of alzheimer's association workgroup 2024 for alzheimer disease. *Neurology*, 104(11):e213675, 2025.

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.

James Munkres. Algorithms for the assignment and transportation problems. *Journal of the society for industrial and applied mathematics*, 5(1):32–38, 1957.

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.

Neil P Oxtoby, Louise-Ann Leyland, Leon M Aksman, George EC Thomas, Emma L Bunting, Peter A Wijeratne, Alexandra L Young, Angelika Zarkali, Manuela MX Tan, Fion D Bremner, et al. Sequence of clinical and neurodegeneration events in parkinson's disease progression. *Brain*, 144(3): 975–988, 2021.

Konstantinos Poulakis, Joana B Pereira, J-Sebastian Muehlboeck, Lars-Olof Wahlund, Örjan Smedby, Giovanni Volpe, Colin L Masters, David Ames, Yoshiki Niimi, Takeshi Iwatsubo, et al. Multi-cohort and longitudinal bayesian clustering study of stage and subtype in alzheimer's disease. *Nature communications*, 13(1):4566, 2022.

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.

Mara ten Kate, Ellen Dicks, Pieter Jelle Visser, Wiesje M van der Flier, Charlotte E Teunissen, Frederik Barkhof, Philip Scheltens, Betty M Tijms, and Alzheimer's Disease Neuroimaging Initiative. Atrophy subtypes in prodromal alzheimer's disease are associated with cognitive decline. *Brain*, 141 (12):3443–3456, 2018.

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.

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.

Peter A Wijeratne, Arman Eshaghi, William J Scotton, Maitrei Kohli, Leon Aksman, Neil P Oxtoby, Dorian Pustina, John H Warner, Jane S Paulsen, Rachael I Scahill, et al. The temporal event-based model: Learning event timelines in progressive diseases. *Imaging Neuroscience*, 1:1–19, 2023.

Alexandra L Young, Neil P Oxtoby, Pankaj Daga, David M Cash, Nick C Fox, Sebastien Ourselin, Jonathan M Schott, and Daniel C Alexander. A data-driven model of biomarker changes in sporadic Alzheimer's disease. *Brain*, 137(9):2564–2577, 2014.

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.

Appendix A. ADNI Information

A complete listing of ADNI investigators can be found at: http://adni.loni.usc.edu/wp-content/uploads/how_to_apply/ADNI_Acknowledgement_List.pdf

Appendix B. Conjugate Prior Update for Distribution Parameters

The updating rule:

$$W = \sum_{j=1}^{J} w_j, \bar{x} = \frac{1}{W}\sum_{j=1}^{J} w_j x_j$$

$$S = \sum_{j=1}^{J} w_j (x_j - \bar{x})^2$$

$$m' = \frac{n_0 m_0 + W\bar{x}}{n_0 + W}, n' = n_0 + W$$

$$\nu' = \nu_0 + W, s' = \frac{1}{\nu'}\left[S + \nu_0 s_0^2 + \frac{n_0 W}{n'}(\bar{x} - m_0)^2\right]$$

From these, the posterior mean of the mean is taken as:

$$\hat{\mu} = m',$$

and the posterior expectation of the variance is:

$$\hat{\sigma}^2 = \frac{\nu' \cdot {s'}^2/2}{\nu'/2} = (s')^2 \tag{8}$$

with posterior standard deviation $\hat{\sigma} = \sqrt{\hat{\sigma}^2}$. This initialization ensures that $\boldsymbol{\theta}, \boldsymbol{\phi}$ are informed by both clustering structure and prior uncertainty.

Note that we knew the statistically correct calculation of $\hat{\sigma}^2$ should be:

$$\hat{\sigma}^2 = \begin{cases} \frac{\nu'(s')^2}{\nu' - 2}, & \nu' > 2, \\ (s')^2, & \nu' \leq 2. \end{cases} \tag{9}$$

but we decided to use $\hat{\sigma}^2 = (s')^2$ because it led to better empirical results.

Appendix C. BEBMS Inference Algorithm

Algorithm 1: BEBMS Metropolis–Hastings Sampler

```
Input: Data $\boldsymbol{X} \in \mathbb{R}^{J \times N}$, Number of subtypes $T$
Output: Samples $\{\boldsymbol{S}^{(i)}, \boldsymbol{\theta}^{(i)}, \boldsymbol{\phi}^{(i)}, \boldsymbol{\pi}_t^{(i)}, \boldsymbol{\pi}_{k|t}^{(i)}, P_{\text{stage}}^{(i)}, P_{\text{subtype}}^{(i)}, \ell^{(i)}\}_{i=1}^{M}$
Init: $\boldsymbol{S}^{(0)}, \boldsymbol{\theta}^{(0)}, \boldsymbol{\phi}^{(0)}, \boldsymbol{\alpha}_t, \boldsymbol{\alpha}_{k|t}, \boldsymbol{\pi}_t^{(0)}, \boldsymbol{\pi}_{k|t}^{(0)}$;
Compute initial posteriors $P_{\text{stage}}^{(0)}, P_{\text{subtype}}^{(0)}$;
Compute total log-likelihood $\ell^{(0)}$;
for $i = 1$ to $M$ do
    // Step 1: Propose new orderings
    Propose $\boldsymbol{S'}$ from $q(\boldsymbol{S'} \mid \boldsymbol{S}^{(i-1)})$;
    Compute intermediate posteriors $\tilde{P}_{\text{stage}}, \tilde{P}_{\text{subtype}}$ using $\boldsymbol{S'}$ and $(\boldsymbol{\theta}^{(i-1)}, \boldsymbol{\phi}^{(i-1)})$;
    Update $(\boldsymbol{\theta}', \boldsymbol{\phi}')$ using $\boldsymbol{S'}, \tilde{P}_{\text{stage}}, \tilde{P}_{\text{subtype}}$;
    // Step 2: Compute likelihood
    Recompute $P'_{\text{stage}}, P'_{\text{subtype}}$ and $\ell'$;
    // Step 3: Acceptance probability
    $\alpha \leftarrow \min\Big(1, e^{\ell' - \ell^{(i-1)}}\Big)$;
    $U \sim \text{Uniform}(0, 1)$;
    // Step 4: Accept/Reject
    if $U < \alpha$ then
        $S^{(i)} \leftarrow \boldsymbol{S'}$;
        $\boldsymbol{\theta}^{(i)} \leftarrow \boldsymbol{\theta}'$, $\boldsymbol{\phi}^{(i)} \leftarrow \boldsymbol{\phi}'$;
        $P_{\text{stage}}^{(i)} \leftarrow P'_{\text{stage}}$, $P_{\text{subtype}}^{(i)} \leftarrow P'_{\text{subtype}}$;
        $\ell^{(i)} \leftarrow \ell'$;
        $\boldsymbol{\pi}_t^{(i)} \sim \text{Dir}(\boldsymbol{\alpha}_t + \text{counts})$;
        $\boldsymbol{\pi}_{k|t}^{(i)} \sim \text{Dir}(\boldsymbol{\alpha}_{k|t} + \text{counts})$;
    end
end
```

At each MCMC iteration, we propose a new event sequence $\boldsymbol{S'}$ by randomly selecting two subtypes. For each of these two subtypes, we then randomly select two biomarkers and swap their positions within the subtype's ordering. If there is only one subtype, we randomly select two biomarkers from it and swap their positions. This defines the proposal distribution $q(\boldsymbol{S'} \mid \boldsymbol{S}^{(i-1)})$.

Appendix D. ADNI Select Biomarker Glossary

Table 1: Glossary of ADNI Biomarkers with Source, Units, and Interpretation

Abbrev.	**Full Name**	**Source Modality**	**Unit / Scale**	**Higher Values Indicate**
MMSE	Mini-Mental State Examination	Cognitive test	Score (0–30)	Less pathology (better global cognition)
ADAS13	Alzheimer's Disease Assessment Scale – Cognitive Subscale (13-item)	Cognitive test	Score (0–85)	More pathology (worse cognition)
RAVLT-immediate	Rey Auditory Verbal Learning Test – Immediate Recall	Cognitive test	Score (0–75)	Less pathology (better memory encoding)
ABETA	Amyloid Beta (Aβ_{1-42})	CSF	pg/mL	Less pathology (less amyloid deposition)
TAU	Total Tau	CSF	pg/mL	More pathology (axonal degeneration / neuronal injury)
PTAU	Phosphorylated Tau (p-Tau$_{181}$)	CSF	pg/mL	More pathology (neurofibrillary tangle burden)
Ventricles	Ventricular Volume (ICV-normalized)	Structural MRI	Fraction of ICV	More pathology (greater brain atrophy)
WholeBrain	Whole Brain Volume (ICV-normalized)	Structural MRI	Fraction of ICV	Less pathology (greater structural integrity)
Hippocampus	Hippocampal Volume (ICV-normalized)	Structural MRI	Fraction of ICV	Less pathology (greater structural integrity)
Entorhinal	Entorhinal Cortex Volume (ICV-normalized)	Structural MRI	Fraction of ICV	Less pathology (greater structural integrity)
Fusiform	Fusiform Gyrus Volume (ICV-normalized)	Structural MRI	Fraction of ICV	Less pathology (greater structural integrity)
MidTemp	Middle Temporal Gyrus Volume (ICV-normalized)	Structural MRI	Fraction of ICV	Less pathology (greater structural integrity)

Appendix E. Biomarker Parameters and Non-Normal Distribution Parameter Details

Table 2: Biomarker Parameterization and Irregular Sampling Distributions

Biomarker	θ_{mean}	θ_{std}	ϕ_{mean}	ϕ_{std}	**Irregular Distribution (Per Implementation)**
MMSE	25.31	2.38	29.17	0.81	Triangular($\mu-2\sigma, \mu-1.5\sigma, \mu$); $\mathcal{N}(\mu+\sigma, (0.3\sigma)^2)$; Exp($0.7\sigma$)+ $(\mu-0.5\sigma)$ (equal mixture).
ADAS13	21.79	9.51	9.32	3.91	Same mixture structure as MMSE (triangular + Gaussian + exponential).
RAVLT_immediate	27.50	7.93	45.39	9.36	Same mixture structure as MMSE (triangular + Gaussian + exponential).
ABETA	661.23	195.29	1331.37	214.57	Pareto(1.5)$\cdot\sigma + (\mu-2\sigma)$; $\mathcal{U}(\mu-1.5\sigma, \mu+1.5\sigma)$; Logistic($\mu, \sigma$) (equal mixture).
TAU	385.84	138.95	208.11	58.84	Same mixture structure as ABETA.
PTAU	37.21	15.09	17.88	5.13	Same mixture structure as ABETA.
VentricleNorm	0.0359	0.0128	0.0198	0.0069	Beta(0.5,0.5)$\cdot 4\sigma + (\mu-2\sigma)$; Exp($0.4\sigma$) with ± sign; $\mathcal{N}(\mu, (0.5\sigma)^2) + 0$, 2σ spike.
HippocampusNorm	0.00390	0.00065	0.00511	0.00059	Same mixture structure as VentricleNorm.
WholeBrainNorm	0.6311	0.0346	0.6949	0.0389	Gamma($2, 0.5\sigma$) + $(\mu-\sigma)$; Weibull(1.0) $\cdot\, \sigma + (\mu-\sigma)$; $\mathcal{N}(\mu, (0.5\sigma)^2) \pm \sigma$.
EntorhinalNorm	0.00217	0.00050	0.00253	0.00038	Same mixture structure as WholeBrainNorm.
FusiformNorm	0.01116	0.00167	0.01186	0.00140	Standard Cauchy(μ, σ) + $\mathcal{N}(0, (0.2\sigma)^2)$, clipped to $[\mu-4\sigma, \mu+4\sigma]$.
MidTempNorm	0.01241	0.00179	0.01344	0.00140	10% $\mathcal{N}(\mu, 0.2\sigma)$ spike + 90% Logistic($\mu+\sigma, 2\sigma$).

Implementation Notes:

- μ & σ use θ parameters for affected (pathological) and ϕ for nonaffected (intact).
- For non-normal components, **After sampling, all values are perturbed by additional noise** $\mathcal{N}(0,\ (0.2\sigma)^2)$ **and clipped to** $[\mu-5\sigma,\ \mu+5\sigma]$.

Appendix F. Experiment Specifications

1. Ordinal $\boldsymbol{S}$ & Ordinal k_j with bell-shape Dirichlet priors & $x_{j,n}$ sampled from normal distributions according to the EBM model.
2. Same as Experiment 1, but non-normal distributions for $x_{j,n}$.
3. Same as Experiment 1, but with uniform distributions for ordinal k_j.
4. Same as Experiment 2, but with uniform distributions for ordinal k_j.
5. Ordinal $\boldsymbol{S}$ & Continuous k_j with uniform distributions & $x_{j,n}$ sampled from normal distributions according to the EBM model.
6. Same as Experiment 5, but with non-normal distributions for $x_{j,n}$.
7. Same as Experiment 6, but with a scaled Beta distribution ($\lambda = N, \alpha = 5, \beta = 2$) for k_j.
8. Ordinal $\boldsymbol{S}$ & Continuous k_j with uniform distributions & $x_{j,n}$ generated using the Sigmoid model.
9. Same as Experiment 8, but with a scaled Beta distribution ($\lambda = N, \alpha = 5, \beta = 2$) for k_j.
10. Continuous event times ($\text{Beta}(2,2) \times N$) & scaled Beta distribution ($\lambda = N, \alpha = 5, \beta = 2$) for k_j & $x_{j,n}$ sampled from normal distributions according to the EBM model.
11. Same as Experiment 10, but $x_{j,n}$ was generated using the Sigmoid model.

Appendix G. Theoretical and Empirical Biomarker Distributions

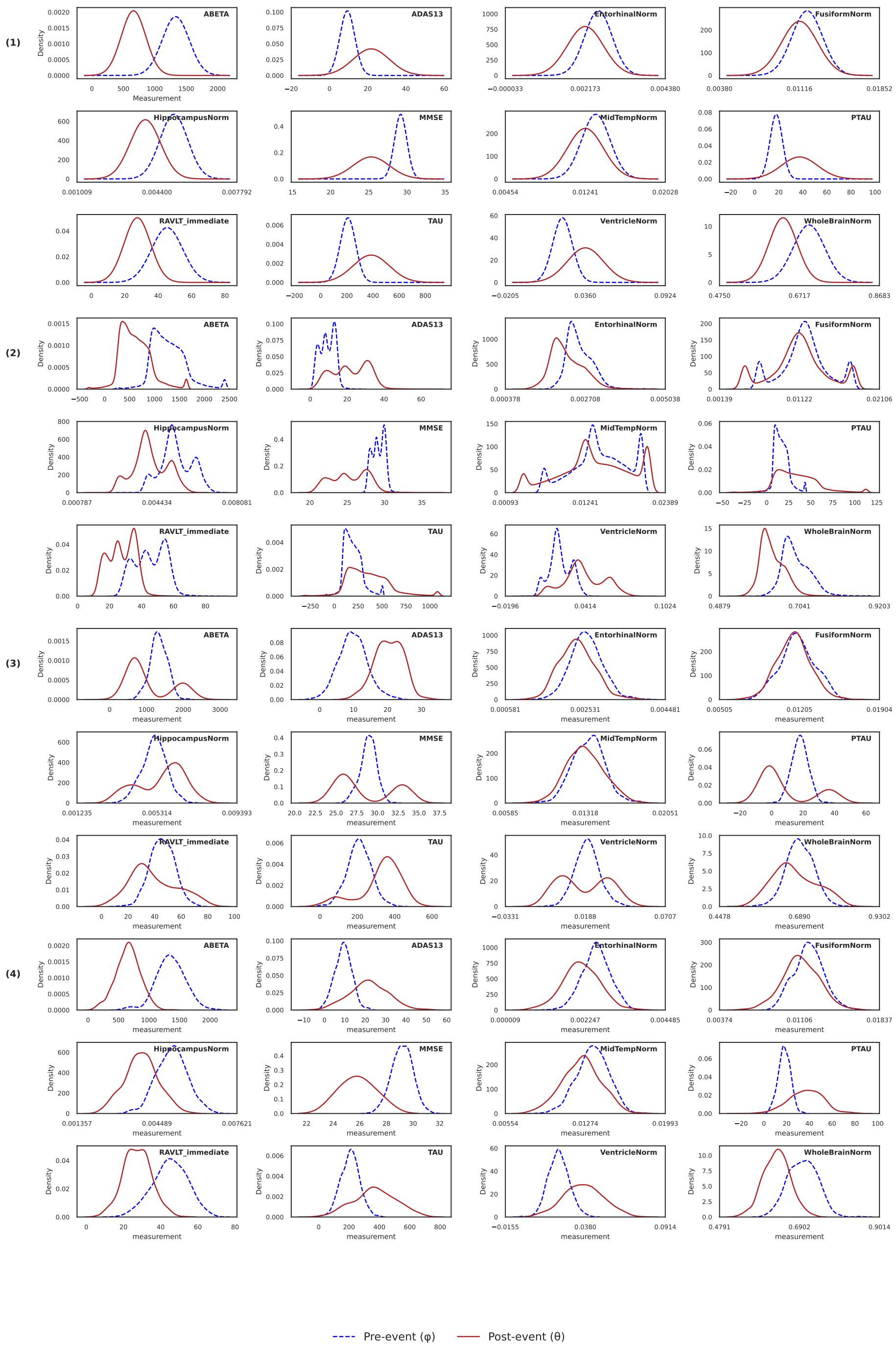

Figure 4: (1) Theoretical normal distributions; (2) Theoretical non-normal distributions; (3) Empirical distributions in one synthetic dataset of Exp. 9; (4) Empirical distributions in one synthetic dataset of Exp. 1.

Appendix H. Kendall's W

We used Kendall's W to measure the similarity among disease progression of different subtypes. W ranges from 0 to 1, with 0 indicating no similarity at all and 1 indicating exactly the same. We aim for the full range of $[0, 1]$ in synthetic experiments because we want to make sure the performances of all algorithms are not dependent on the similarity among subtypes.

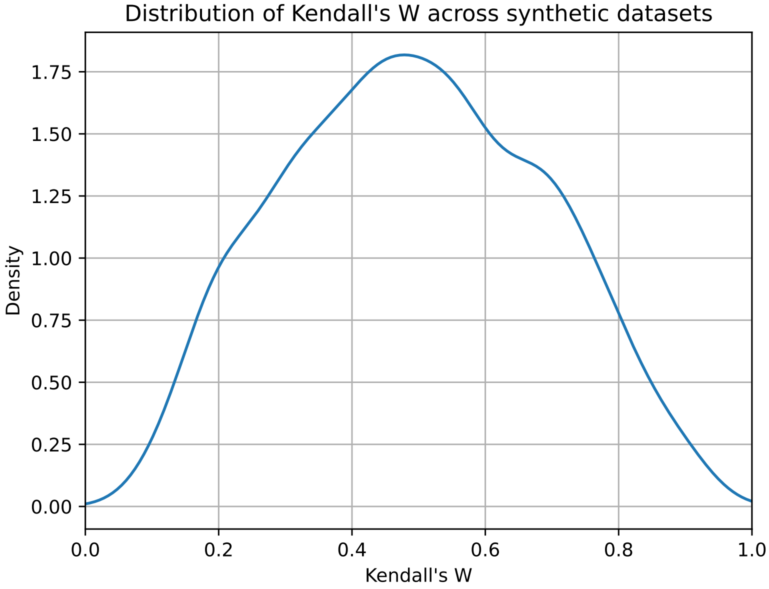

Figure 5: Distribution of Kendall's W across all 1,318 synthetic datasets (Exp. 1-11).

Appendix I. Synthetic Experiment Results

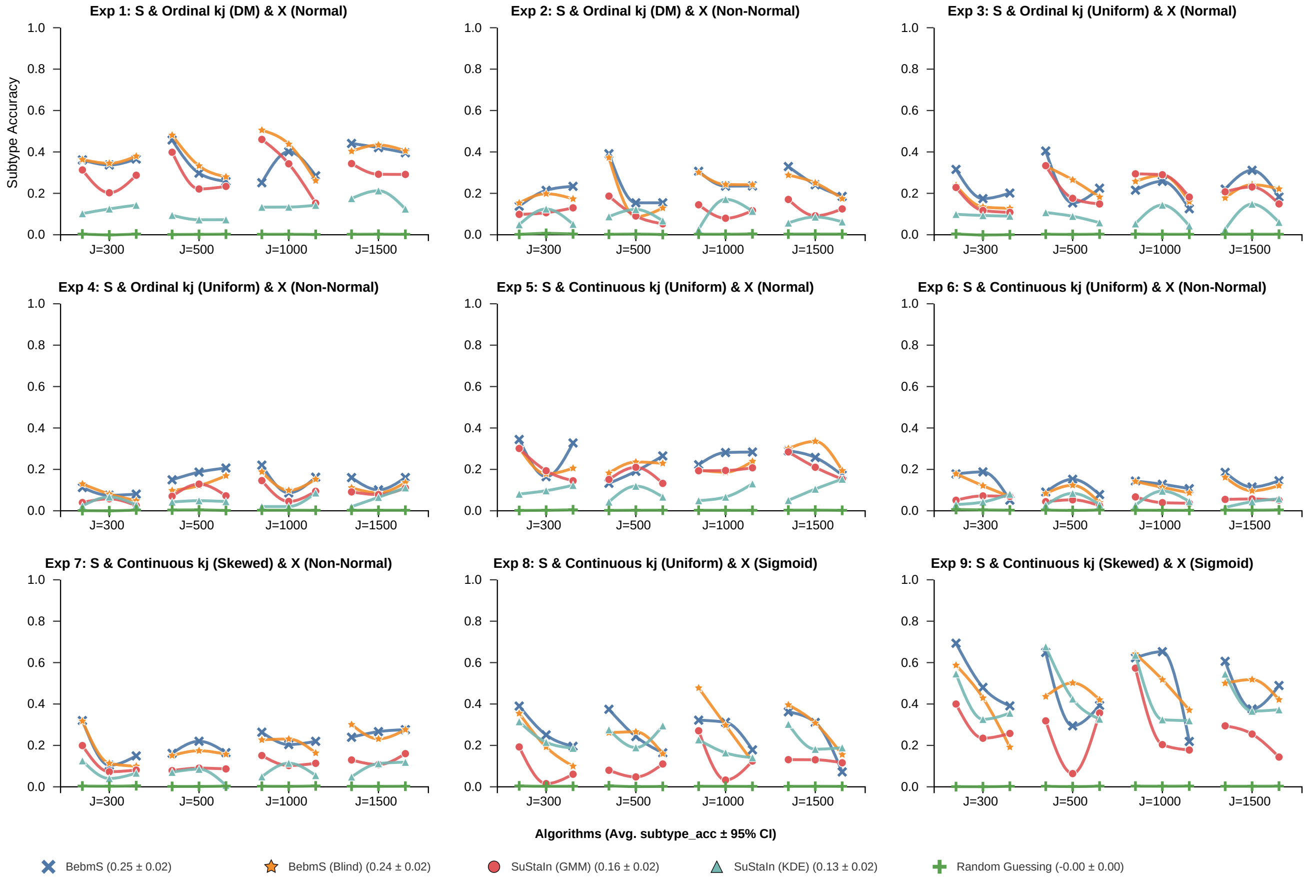

Figure 6: Subtyping results, measured with Adjusted Rand Index (ARI). Larger values reflect better performance. Subtyping is hard for both BEBMS and SuStaIn.

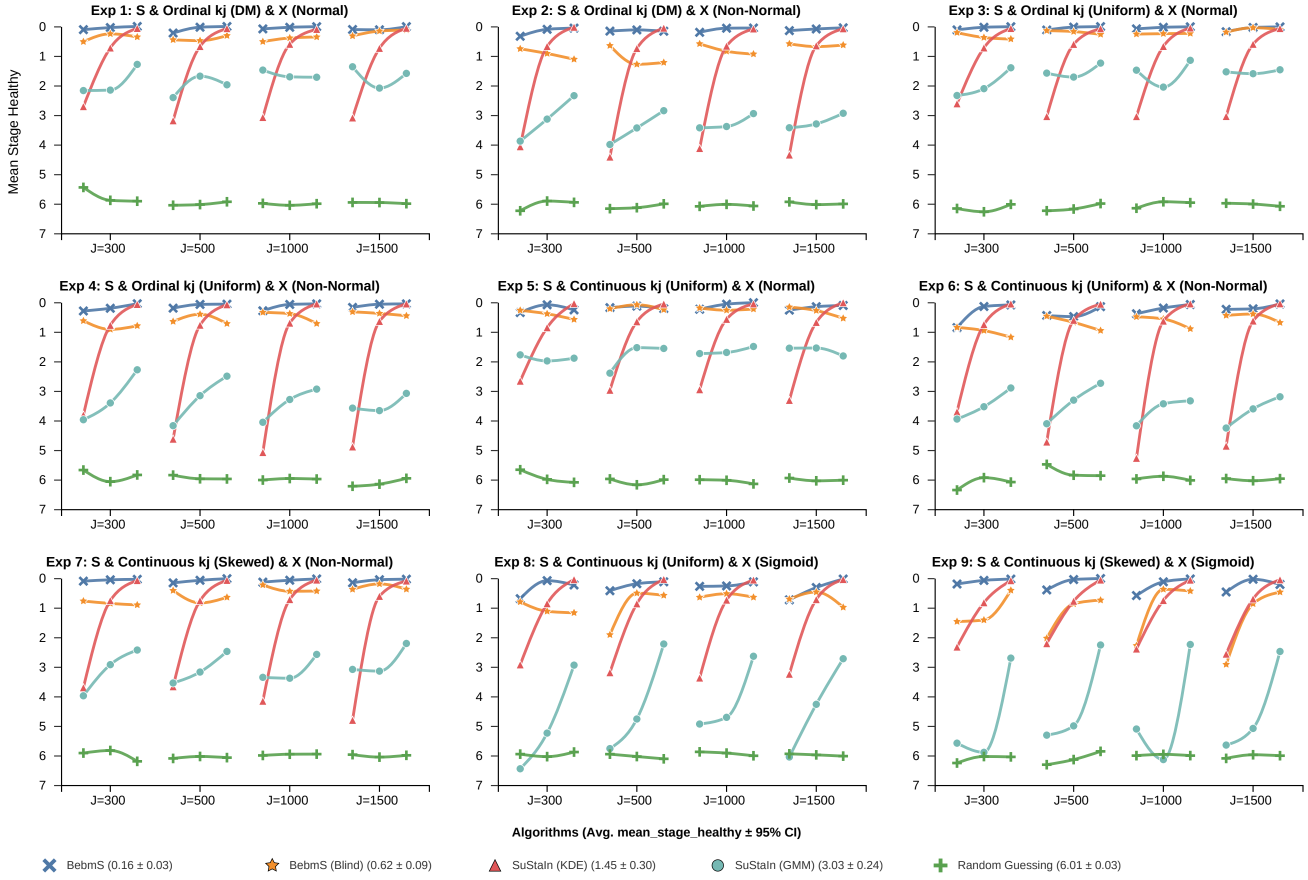

Figure 7: Staging results. BEBMS assigned lower disease stages to control participants than SuStaIn. 0 is the ground truth.

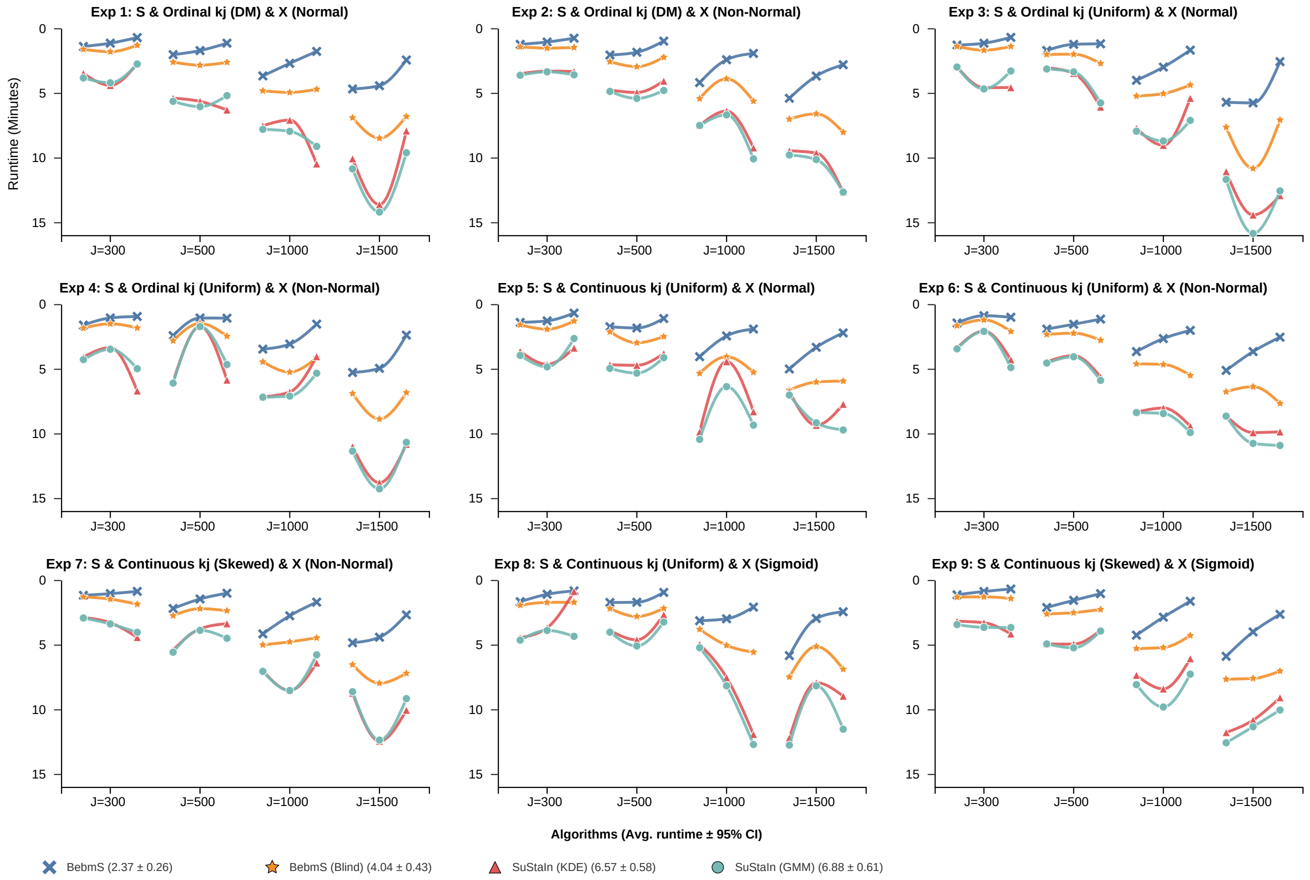

Figure 8: Runtime analysis. BEBMS is much faster than SuStaIn.

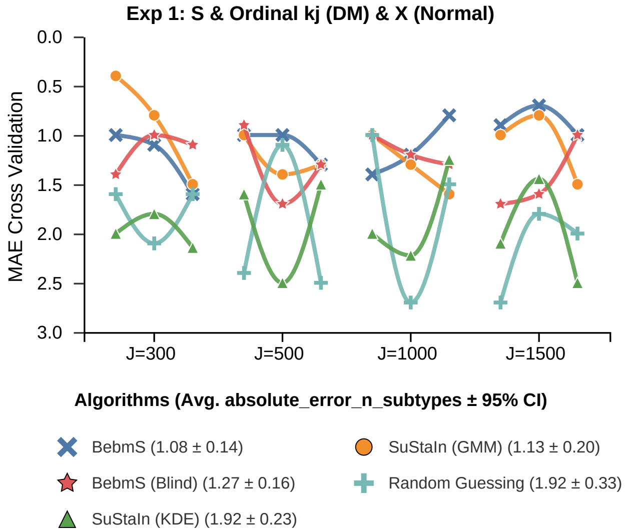

Figure 9: Mean Absolute Error (MAE) for the estimation of ground-truth number of subtypes. Smaller values reflect better performance. BEBMS's performance is similar to that of SuStaIn GMM.

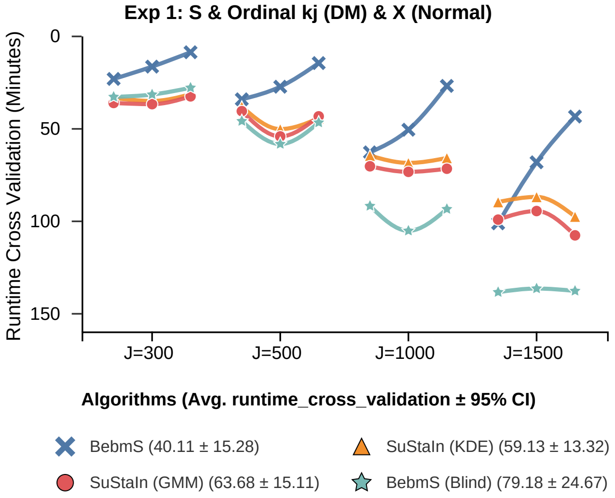

Figure 10: Runtime analysis for the estimation of ground-truth number of subtypes. BEBMS achieved similar accuracy to SuStaIn GMM but with a faster speed.

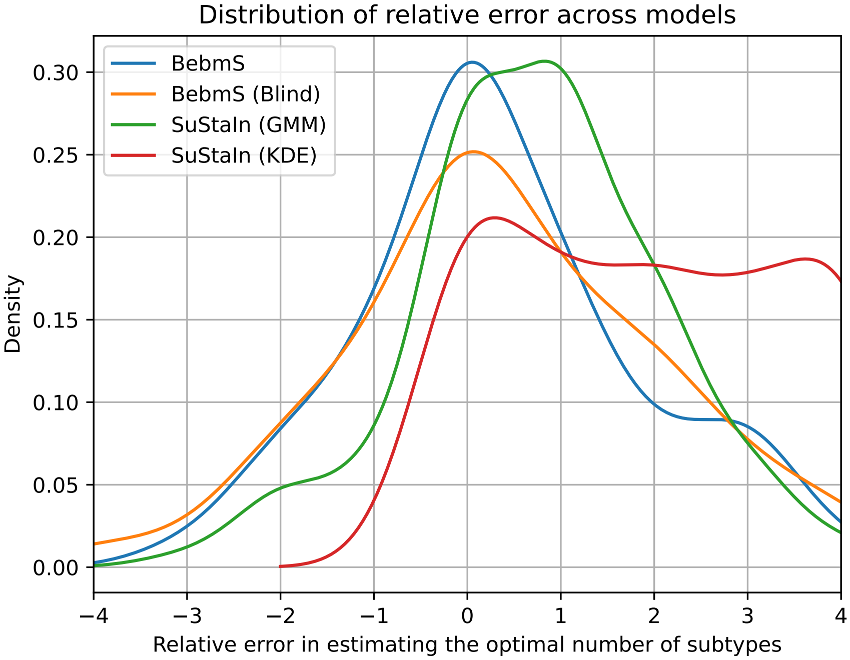

Figure 11: Relative error distributions for estimating the optimal subtype count. The BEBMS variants produce symmetric, zero-centered errors, while SuStaIn consistently overestimates (overfits) the number of subtypes.

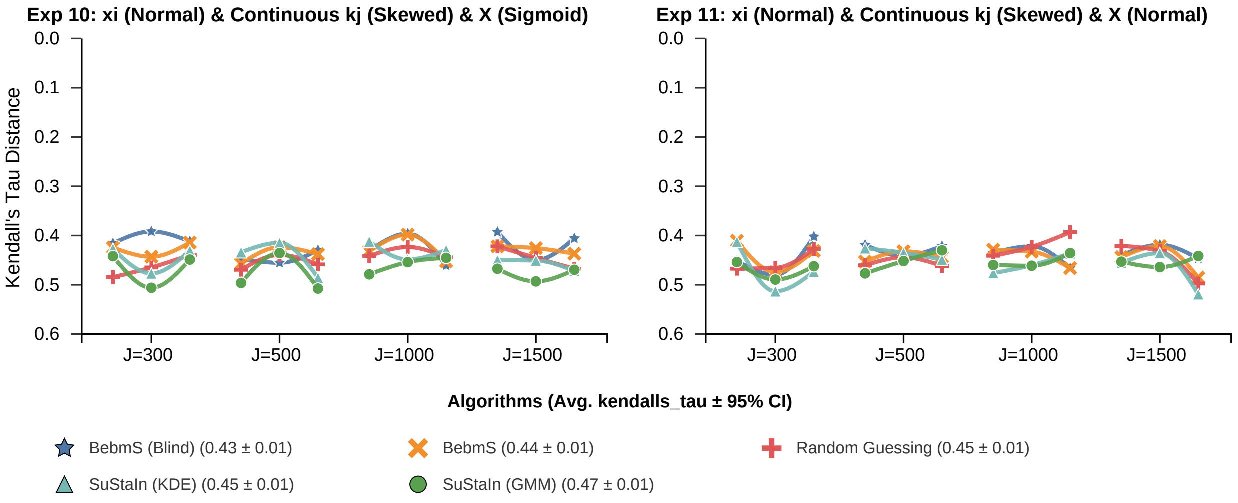

Figure 12: Ordering results for stress-test experiments (Kendall's tau distance).

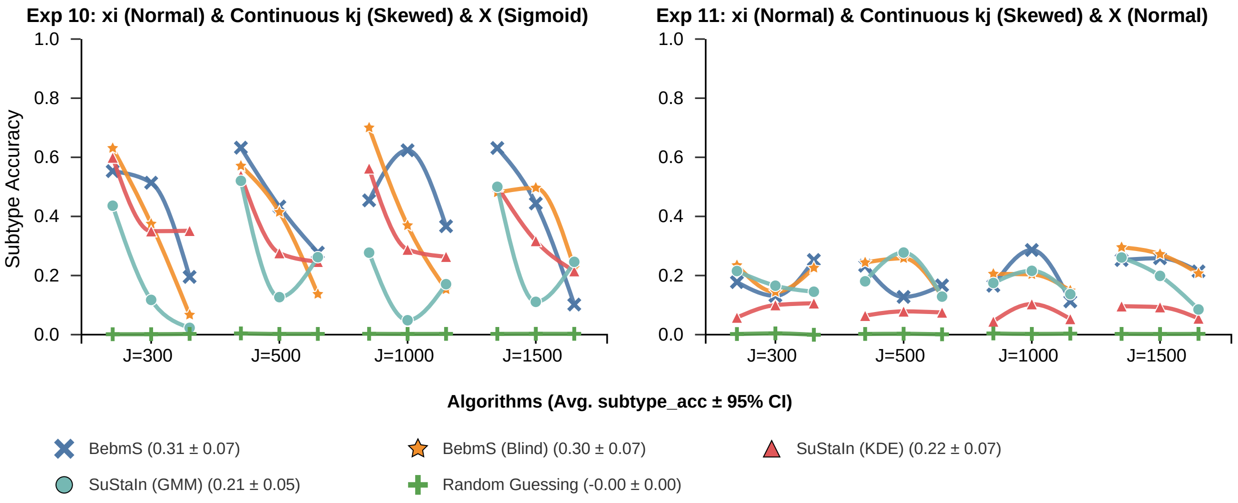

Figure 13: Subtype assignment results for stress-test experiments. .

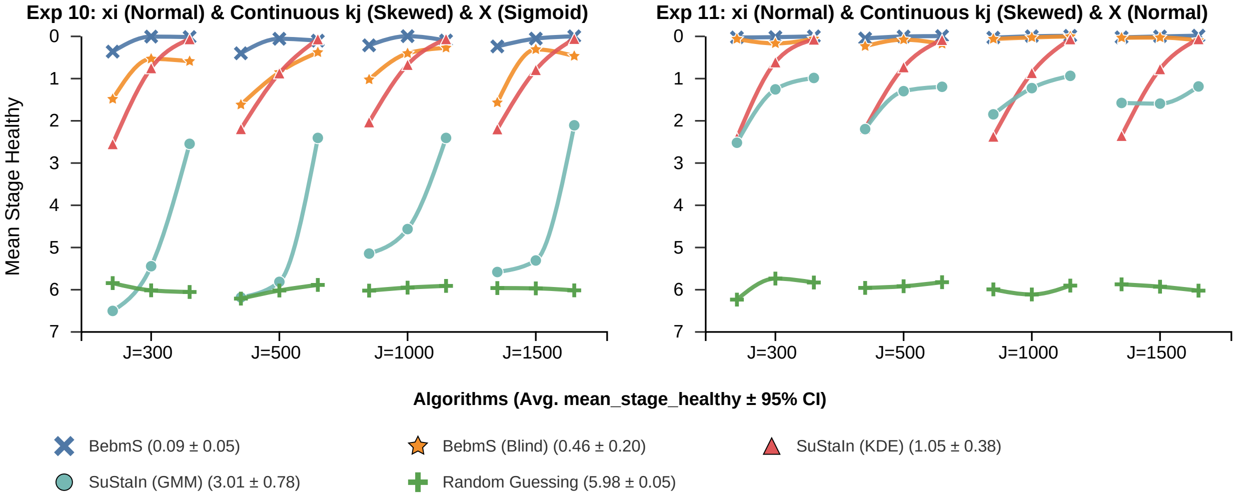

Figure 14: Staging on stress-test experiments.

Appendix J. ADNI Results

The results vary because of randomness. We tested BEBMS and SuStaIn GMM several times. BEBMS favored 3 or 4 as the optimal number of subtypes, and SuStaIn favored 5 or 6. We picked 6 for SuStaIn because that was our last attempt. For BEBMS, when the number of subtypes was 4, one of the resulting subtypes only had 6 participants, a clear indication of overfitting. Therefore, we chose 3.

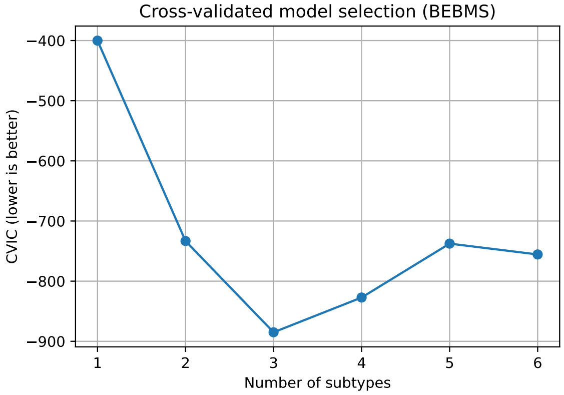

Figure 15: Cross-validation on ADNI using BEBMS

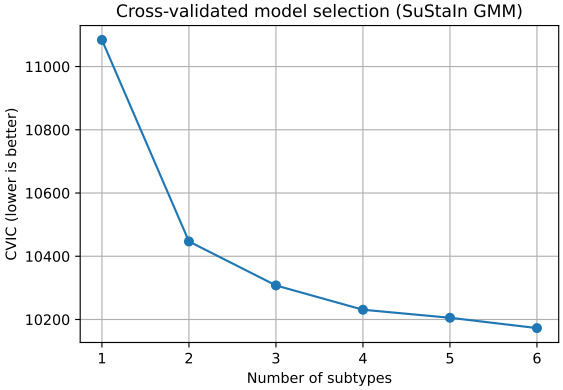

Figure 16: Cross-validation on ADNI using SuStaIn (GMM)

Table 3: Comparison of CVIC across number of subtypes for BEBMS and SuStaIn (GMM).

Number of Subtypes	BEBMS CVIC	SuStaIn (GMM) CVIC
1	−400.11	11084.15
2	−733.31	10446.74
3	**−885.00**	10307.71
4	−827.21	10230.84
5	−737.59	10205.47
6	−755.60	**10172.81**

Note. CVIC = Cross-Validation Information Criterion. Lower values indicate better model fit.

Table 4: Diagnostic composition (proportions) across subtypes (SuStaIn GMM)

Subtype	Total	AD	CN	EMCI	LMCI
1	342	0.31	0.16	0.21	0.32
2	124	0.03	0.30	0.32	0.35
3	148	0.27	0.14	0.25	0.34
4	54	0.00	0.35	0.28	0.37
5	46	0.02	0.39	0.28	0.30
6	12	0.17	0.42	0.33	0.08

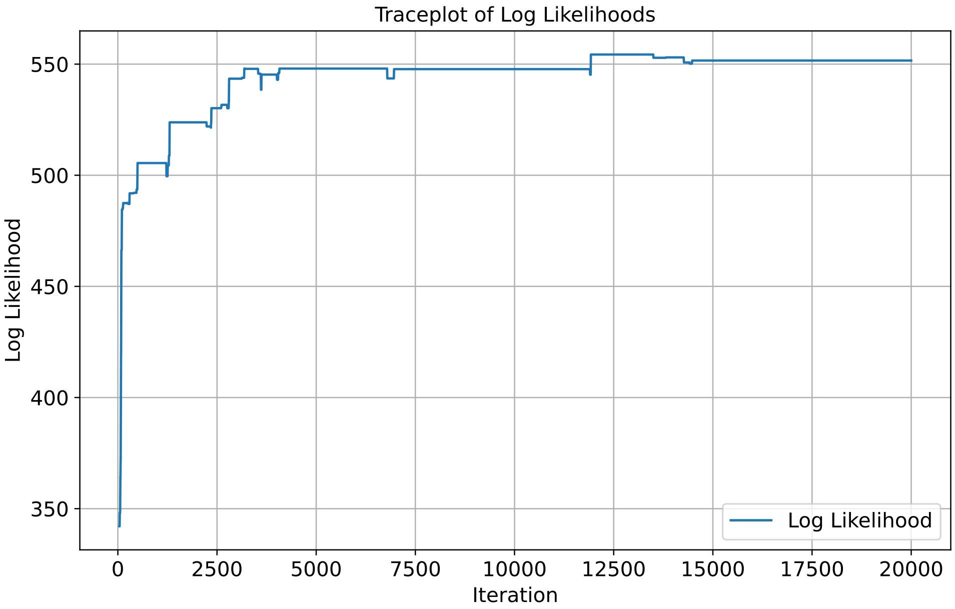

Figure 17: The trace plot of running BEBMS on ADNI. The starting point is iteration 40. Qualitatively, this suggests good convergence.

Table 5: Diagnostic composition (proportions) across subtypes (BEBMS)

Subtype	**Total**	**AD**	**CN**	**EMCI**	**LMCI**
1	25	0.28	0.04	0.28	0.40
2	493	0.17	0.26	0.30	0.28
3	208	0.30	0.13	0.13	0.43

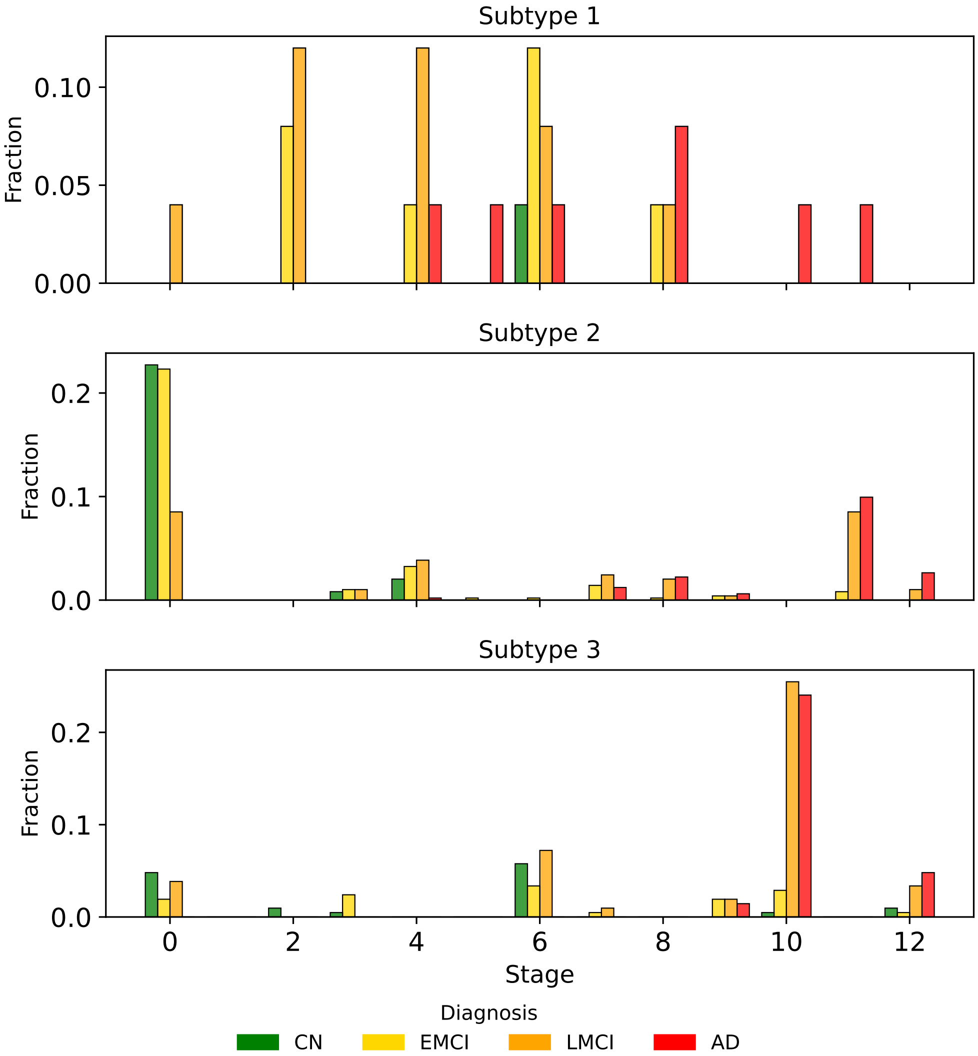

Figure 18: BEBMS ADNI staging and subtyping

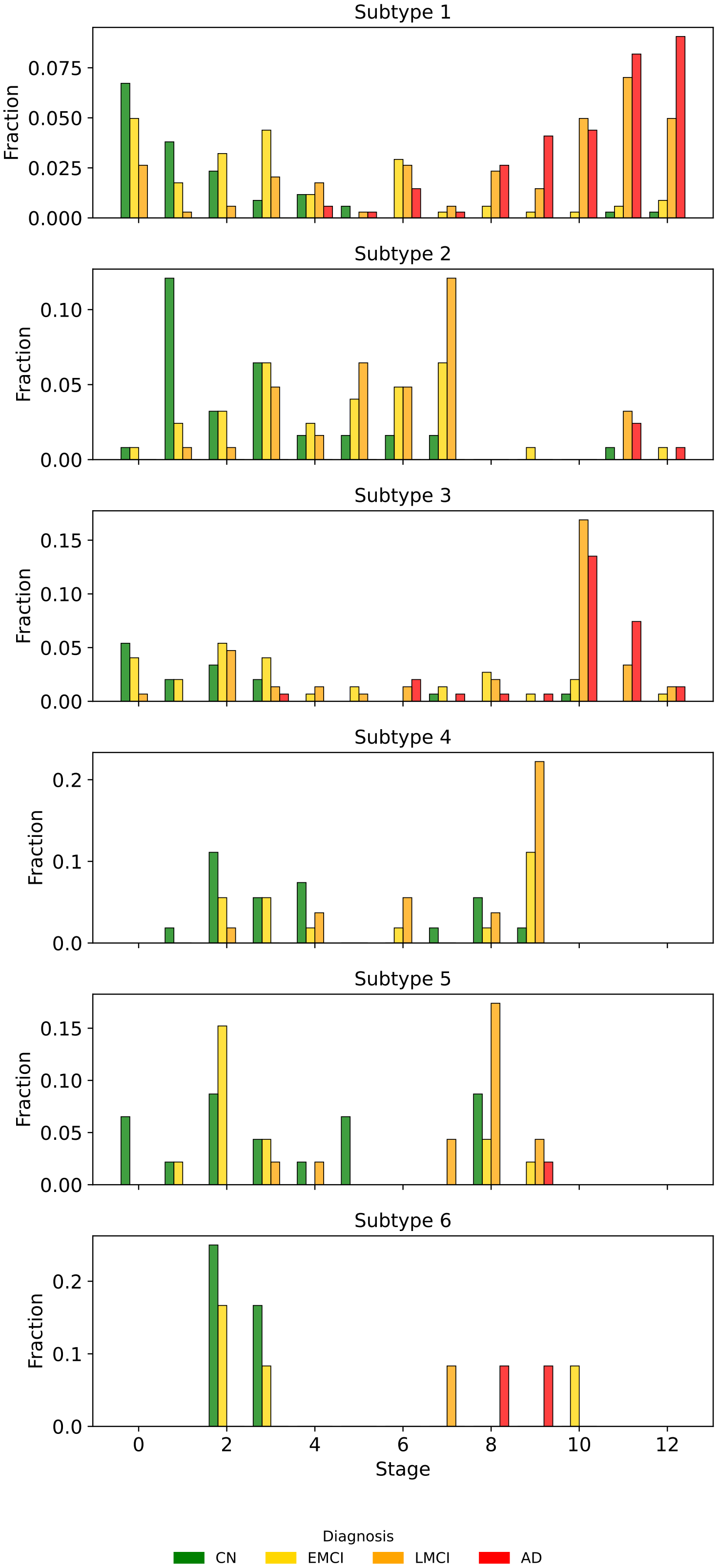

Figure 19: SuStaIn ADNI staging and subtyping

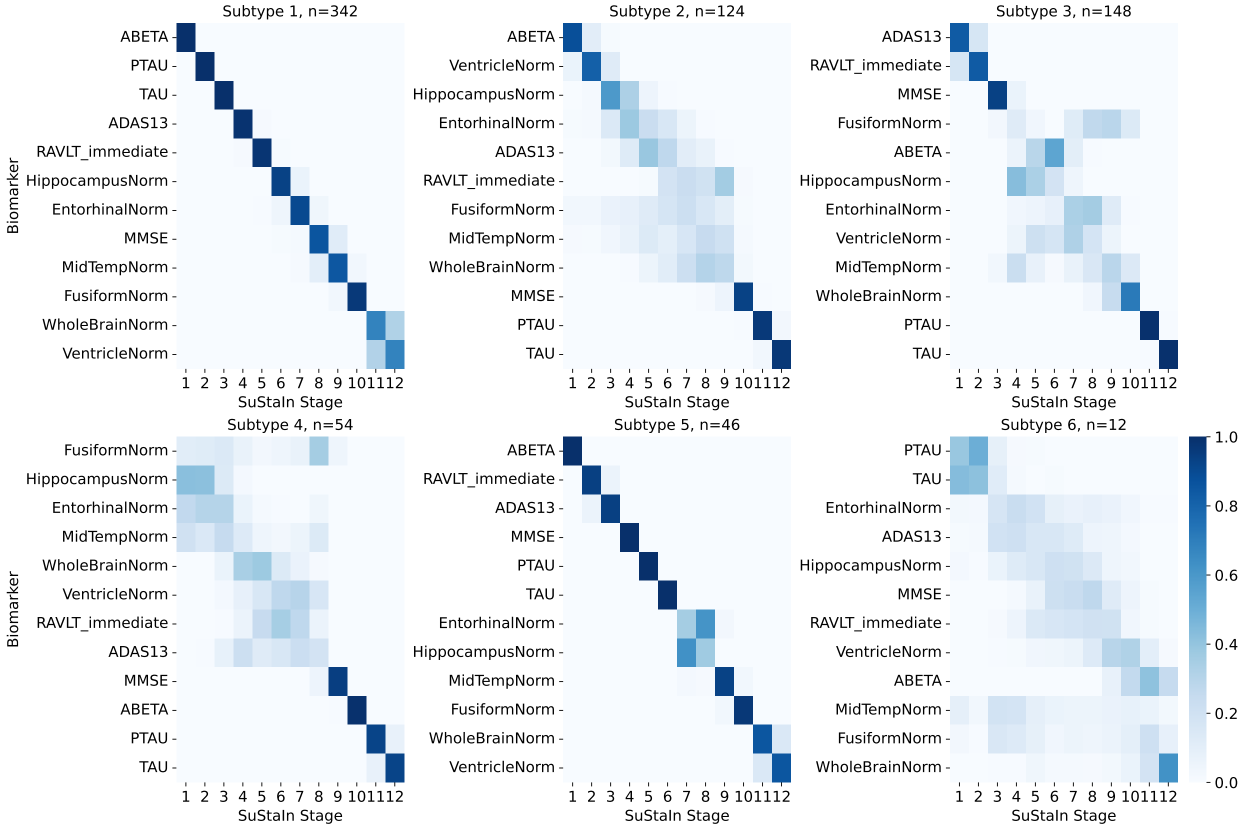

Figure 20: SuStaIn ADNI ordering

Appendix K. Incomplete & High Dimensional Data

K.1. Handling of Incomplete Data

BEBMS is designed to accommodate datasets with missing values. To achieve this, missing entries (represented as `NaN`) are systematically excluded during two key processes: (1) the initialization and iterative updating of biomarker distribution parameters, and (2) the calculation of the data likelihood.

On the other hand, although missing data SuStaIn (Estarellas et al., 2024) has been proposed, the currently available implementation of SuStaIn, whose source code was reported by Estarellas et al. (2024) in the data availability section, is not able to process data with missing entries.

K.2. Performance on High-dimensional Data

To assess the scalability and performance of BEBMS on high-dimensional data, we conducted a synthetic experiment. First, distribution parameters $(\boldsymbol{\theta}, \boldsymbol{\phi})$ for 100 biomarkers were synthetically generated using AI. More specifically, we used the following prompt on ChatGPT5:

```
[Paste data from the JSON file into the prompt:
https://raw.githubusercontent.com/hongtaoh/bebms/refs/heads/main/4highdim_params_ucl_gmm.json]

This is for Alzheimer's disease.

Based on this, could you give me 100 synthetic biomarkers' parameters?
```

The resulting JSON can be found at https://raw.githubusercontent.com/hongtaoh/bebms/refs/heads/main/high_dimensional/high_dimensional.json.

Based on these parameters, five distinct datasets were created following the configurations of Experiment 1 (see Appendix F). The experiment was configured with $J = 300$ participants and $R = 0.25$, and the model was run for 3,000 MCMC iterations. All other experimental settings were identical to those described in the synthetic experiments of Section 4.

The five high-dimensional datasets are available at https://github.com/hongtaoh/bebms/tree/main/high_dimensional/data.

The results, summarized in Table 6, indicate that BEBMS required an average processing time of approximately 10 minutes for each dataset. Considering the dimensionality of the data (100 biomarkers, 300 participants, and on average 3.4 subtypes), the model demonstrated robust performance on the ordering, subtyping, and staging tasks.

This computational speed is comparable to, though slower than, that of s-SuStaIn (Tandon et al., 2024), which reportedly processes a dataset of 200 participants and 100 biomarkers (with 3-4 subtypes) in approximately 2 minutes. In contrast, SuStaIn GMM (without parallel start points) failed to complete processing on the 5th dataset within a 3-hour time limit. This limitation of SuStaIn aligns with the findings previously reported in the Figure 2 of Tandon et al. (2024).

Table 6: Comparison of runtime, ordering accuracy, subtype assignment accuracy, and healthy mean stage.

Dataset	# subtypes	Runtime (s)	Kendall's τ	Subtype Acc.	Mean Stage (Healthy)
1	1	226.58	0.293	1.000	0.173
2	5	822.69	0.419	0.012	0.014
3	4	719.95	0.378	0.678	2.164
4	1	229.38	0.349	1.000	1.440
5	5	837.29	0.482	0.016	1.068
Average	3.2	**567.18**	**0.384**	**0.541**	**0.972**