Bayesian Event-Based Model for Disease Subtype and Stage Inference

Hongtao Hao, Joseph L. Austerweil

Introduction

Problem: Chronic diseases like Alzheimer's rarely follow a single pathway—patients exhibit distinct disease subtypes with different progression patterns. SuStaIn (Subtype and Stage Inference) is the current standard for discovering disease subtypes from cross-sectional data, but its robustness under model misspecification remains unclear.

Our Solution: We introduce bebms (Bayesian Event-Based Model for Subtyping), a principled Bayesian framework that iteratively estimates biomarker distributions, disease stages, and subtype assignments.

Key Improvements:

- **27% better** ordering accuracy
- **89% better** staging accuracy
- **56% better** subtype assignment
- **Faster** than SuStaIn

Methods

Model Specification

bebms models disease progression as a sequence of N biomarker events. Each participant j has:

- Disease stage: $k_j \in \{0, 1, ..., N-1\}$ (or $k_j = -1$ if healthy)
- Subtype assignment: $c_j \in \{1, ..., T\}$

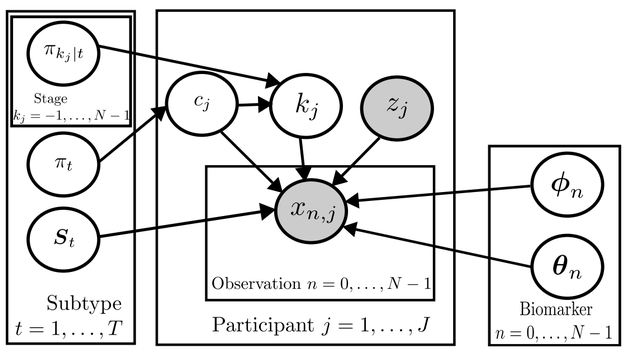

Figure 1: bebms as a graphical model

Key assumptions:

- Biomarkers follow Gaussian distributions (pre-event φ, post-event θ)
- Shared parameters across subtypes
- Mixture over subtypes and stages, i.e., a participant belongs to every subtype and every disease stage with different probabilities.

Data likelihood for progressing participant:

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

where $\boldsymbol{S}$ is a $T \times N$ matrix where row t is the order of disease progression over biomarkers for subtype t.

Inference Algorithm

Key innovation: Iterative estimation via Metropolis-Hastings MCMC

1. **Initialize** biomarker parameters (θ, φ) using K-Means + conjugate priors
2. **Propose** new subtype orderings $\boldsymbol{S}'$
3. **Compute** stage and subtype posteriors:

$$P_{\text{stage}(k \mid j,t)} \propto \pi_{k|t} \cdot p(\boldsymbol{X}_j \mid \boldsymbol{S}_t, k, \theta, \varphi) \tag{2}$$

$$P_{\text{subtype}(t \mid j)} \propto \pi_t \sum_k P_{\text{stage}(k \mid j,t)} \tag{3}$$

4. **Update** parameters using soft assignments
5. **Accept/reject** based on data likelihood

Model selection: 5-fold cross-validation with CVIC to choose optimal number of subtypes T.

Experimental Setup

Synthetic Data

Generation: 1,320 datasets with with data generated based on 12 ADNI biomarkers.

- Participant sizes (J): 300, 500, 1000, 1500
- Healthy ratios (R): 0.25, 0.5, 0.75
- Ground truth subtypes: 1-5 (randomly selected)
- Two generative models: EBM and Sigmoid

Model misspecification tests:

- Non-Gaussian biomarker distributions
- Continuous disease stages (vs. ordinal)
- Uneven event spacing

Baseline: SuStaIn (GMM and KDE variants)

Evaluation Metrics

- **Ordering:** Normalized Kendall's τ distance (lower = better)
- **Subtyping:** Adjusted Rand Index (higher = better)
- **Staging:** Mean stage assigned to healthy participants (0 is ground truth)
- **Runtime:** Processing time per dataset

Results

Synthetic Experiments

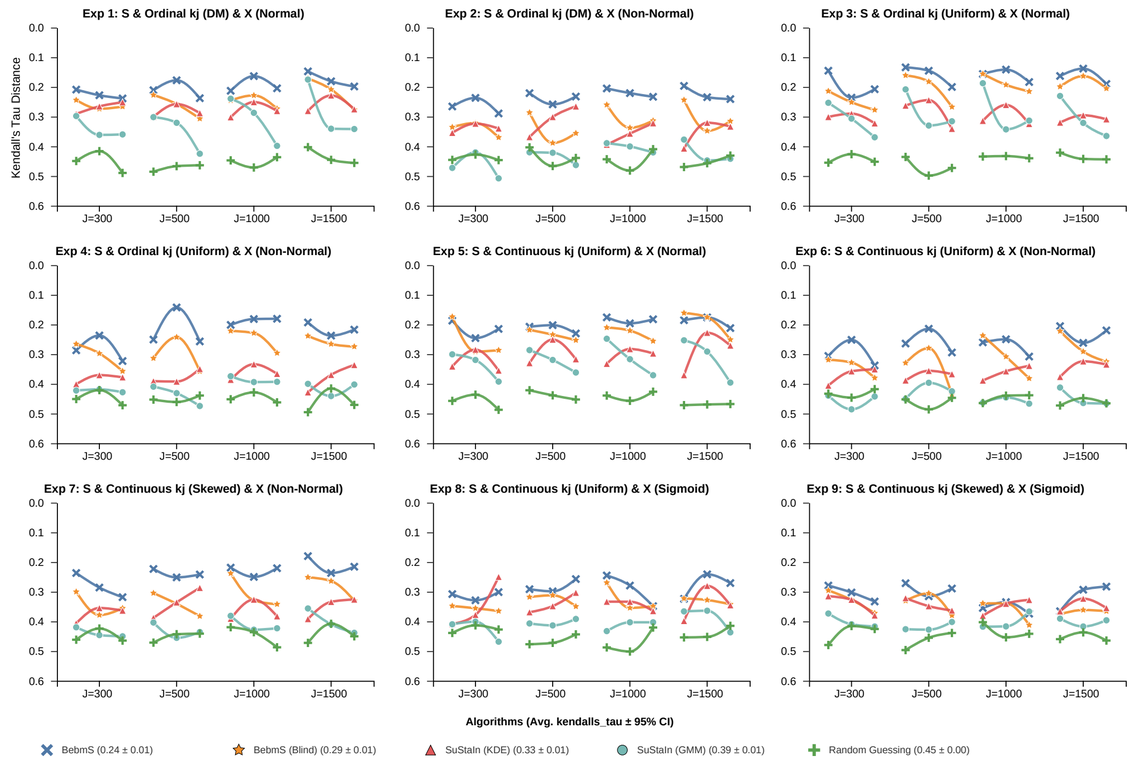

Figure 2: **Ordering accuracy:** bebms achieves normalized τ distance of 0.24 vs. SuStaIn's 0.39

Key findings:

Task	bebms	SuStaIn GMM	Improvement
Ordering (τ)	**0.24 ± 0.01**	0.39 ± 0.01	**27%**
Subtyping (ARI)	**0.25 ± 0.02**	0.16 ± 0.02	**56%**
Staging (mean)	**0.16 ± 0.03**	3.03 ± 0.24	**89%**
Runtime (min)	**2.37 ± 0.26**	6.88 ± 0.61	**2.9× faster**

- bebms performs well even with non-Gaussian data
- Performance saturates at 300 participants
- Robust to varying healthy ratios

ADNI Real-World Data

Dataset: 726 participants (153 AD, 236 LMCI, 182 EMCI, 155 CN)

Model selection: bebms selected 3 subtypes, SuStaIn selected 6

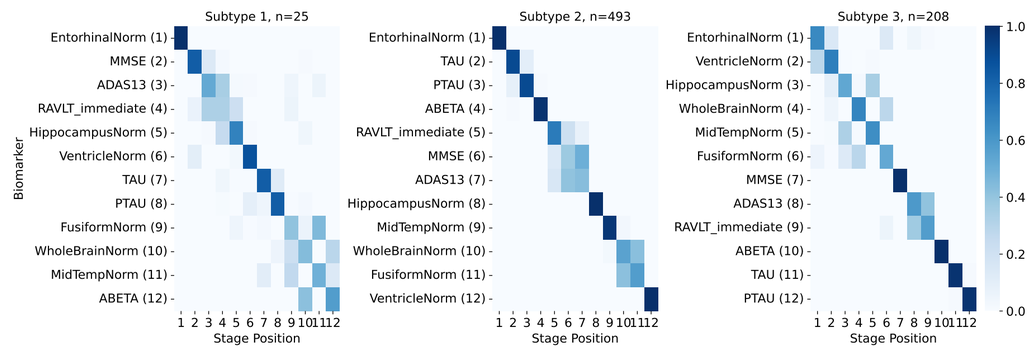

Figure 3: bebms AD progression subtype results on ADNI

bebms subtypes align with known AD pathology:

1. **Subtype 2 (67.9%):** CSF-first → **Typical AD**
2. **Subtype 3 (28.7%):** Entorhinal/hippocampal-first → **Limbic-predominant AD**
3. **Subtype 1 (3.4%):** Neocortical/cognitive-first → **Hippocampal-sparing AD**

These proportions closely match neuropathological evidence: TAD 75%, LPAD 14%, HSAD 11% (Murray et al., 2011).

Staging accuracy: bebms assigned healthy participants to stage 1.10 on average, vs. SuStaIn's 2.48.

Discussion & Conclusions

- Introduced bebms: a **principled Bayesian** approach to disease subtyping which consistently outperforms SuStaIn across all tasks
- More interpretable results aligning with scientific consensus
- Computationally more efficient

Current limitations:

- Assumes ordinal event sequences
- Struggles with continuous, unevenly-spaced events
- Validated on one real-world dataset (ADNI)

Future directions:

- Integrate continuous-time modeling (e.g., Temporal EBM)
- Validate on additional disease datasets
- Extend to longitudinal data

Code: `pip install bebms` | **Package:** github.com/jpcca/bebms_pkg
Data & Experiments: github.com/hongtaoh/bebms
Acknowledgement: ADNI/NACC for data, CHTC for computing resources, and JPCCA for funding

MACHINE LEARNING FOR HEALTH SYMPOSIUM 2025 hongtaoh@cs.wisc.edu