Stage-Aware Event-Based Modeling (SA-EBM) for Disease Progression

Hongtao Hao [1] Vivek Prabhakaran [1] Veena A Nair [1] Nagesh Adluru [1] Joseph Austerweil [2,1] and for the Alzheimer's Disease Neuroimaging Initiative

[1]University of Wisconsin-Madison [2]Chiba Institute of Technology

Introduction

Understanding how diseases progress over time is central to early diagnosis, prognosis, and intervention. This is especially the case for chronic and neurodegenerative conditions such as Alzheimer's and Parkinson's diseases. While longitudinal studies are ideal, they are often expensive, time-consuming, and logistically challenging, resulting in limited availability. As a result, there is increasing interest in using cross-sectional data.

Table 1. Participant biomarker measurements

ID	Impacted	FUS-FCI	P-Tau	MMSE	AB	HIP-FCI	PCC-FCI	⋯
1	Yes	27.16	-6.14	24.49	147.99	1.59	2.80	⋯
2	Yes	17.20	57.89	24.43	157.13	-4.06	8.84	⋯
3	Yes	13.99	62.51	20.87	158.12	6.48	4.42	⋯
⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋱

Methods

In the EBM framework, each biomarker exists in a "pre-event" or "post-event" state, with the "event" signifying the point at which the biomarker becomes pathological. Assuming a set of N biomarkers, we have N possible disease stages. Let J denote the total number of participants, j index participants, and k_j be their current disease stage, where $k_j = 0$ for healthy participants and $k_j > 0$ for progressing participants. Let n be a biomarker and $S(n)$ be its index (1-based) of the disease progression order $\boldsymbol{S}$. EBM assumes biomarker n becomes pathological when $k_j \geq S(n)$, with pre-event and post-event states modeled by separate distributions parameterized by $\boldsymbol{\phi}$ and $\boldsymbol{\theta}$ respectively.

Let $\boldsymbol{X}$ denote the full data, $\boldsymbol{X}_j$ be the biomarker measurements for participant j, and $x_{j,n}$ be biomarker n's measurement of participant j. The likelihood of $\boldsymbol{X}_j$ for a progressing participant with $k_j > 0$ is:

$$P(\boldsymbol{X}_j \mid \boldsymbol{S}, z_j = 1) = \sum_{k_j=1}^{N} P(k_j) p(\boldsymbol{X}_j \mid \boldsymbol{S}, z_j = 1, k_j) \tag{1}$$

where $P(k_j)$ is the prior probability of stage k_j, and z_j indicates this is a progression subject (otherwise $z_j = 0$). $p(\boldsymbol{X}_j \mid \boldsymbol{S}, k_j)$ is computed as:

$$p(\boldsymbol{X}_j \mid \boldsymbol{S}, z_j = 1, k_j) = \underbrace{\prod_{i=1}^{k_j} p(x_{j,S_i} \mid \boldsymbol{\theta}_{S_i})}_{\text{post-event likelihood}} \underbrace{\prod_{i=k_j+1}^{N} p(x_{j,S_i} \mid \boldsymbol{\phi}_{S_i})}_{\text{pre-event likelihood}} \tag{2}$$

where S_i is the i-th (1-based) biomarker to become pathological according to $\boldsymbol{S}$, and x_{j,S_i} is its measurement for participant j. The likelihood for a healthy participant is:

$$p(\boldsymbol{X}_j \mid \boldsymbol{S}, z_j = 0) = \prod_{i=1}^{N} p(x_{j,S_i} \mid \boldsymbol{\phi}_{S_i}) \tag{3}$$

The total likelihood of the dataset is:

$$P(\boldsymbol{X} \mid \boldsymbol{S}, \boldsymbol{z}) = \prod_{j=1}^{J} P(\boldsymbol{X}_j \mid \boldsymbol{S}, z_j) \tag{4}$$

where $\boldsymbol{z} = (z_1, z_2, ..., z_J)$. EBM is a generative model, and can be used to generate synthetic data:

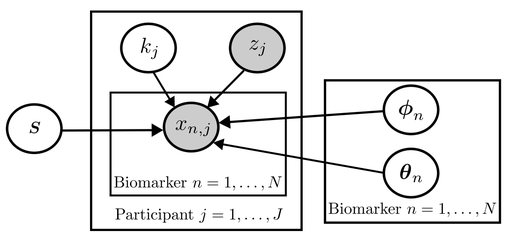


Figure 1. **Graphical model of EBM**

$$x_{n,j} \mid \boldsymbol{S}, k_j, \boldsymbol{\theta}_n, \boldsymbol{\phi}_n, z_j \sim I(z_j = 1)\Big[I(S(n) \leq k_j)\, p(x_{n,j} \mid \boldsymbol{\theta}_n) + I(S(n) > k_j)\, p(x_{n,j} \mid \boldsymbol{\phi}_n)\Big] + \big(1 - I(z_j = 1)\big)\, p(x_{n,j} \mid \boldsymbol{\phi}_n) \tag{5}$$

Algorithm 1 Stage-Aware Event-Based Model (SA-EBM) Algorithm

$\boldsymbol{\pi} = \big(P(k_j)\big)_{k_j=1}^{N} \sim \text{Dirichlet}(\boldsymbol{\alpha}_0)$, where $\boldsymbol{\alpha}_0 = \mathbf{1}_N$
$\boldsymbol{\theta} = (\boldsymbol{\theta}_n)_{n=1}^{N}$ (post-event state) and $\boldsymbol{\phi} = (\boldsymbol{\phi}_n)_{n=1}^{N}$ (pre-event state) using K-Means clustering and conjugate prior updates on the biomarker data
Initialize $\boldsymbol{S}$ as sampled uniformly from all permutations: $\boldsymbol{S} \sim \text{Uniform}(N!)$.
$\ell = -\infty$
for $i = 1$ to M (number of MCMC iterations) **do**
  Propose $\boldsymbol{S}'$ by randomly swapping two biomarkers in $\boldsymbol{S}$.
  $\mathbf{A} = \big(P(k_j \mid \boldsymbol{X}_j, \boldsymbol{S}', \boldsymbol{\theta}, \boldsymbol{\phi}, \boldsymbol{\pi}) \quad \forall k_j \in \{1, 2, ..., N\}\big)_{j=1}^{J}$.
  Compute $\boldsymbol{\theta}', \boldsymbol{\phi}'$ based on $\boldsymbol{S}'$ and $\mathbf{A}$
  $\ell' = \mathcal{L}(\boldsymbol{X} \mid \boldsymbol{S}', \boldsymbol{\theta}', \boldsymbol{\phi}', \boldsymbol{\pi})$ using Equation 4
  $p = \min\big(1, \exp(\ell' - \ell)\big)$
  $U \sim \text{Uniform}(0, 1)$
  if $U < p$ **then**
    $\boldsymbol{S} \leftarrow \boldsymbol{S}'$
    $\ell \leftarrow \ell'$
    $\boldsymbol{\theta} \leftarrow \boldsymbol{\theta}'$ and $\boldsymbol{\phi} \leftarrow \boldsymbol{\phi}'$
    $\mathbf{A} \leftarrow \big(P(k_j \mid \boldsymbol{X}_j, \boldsymbol{S}', \boldsymbol{\theta}, \boldsymbol{\phi}, \boldsymbol{\pi}) : k_j \in \{1, 2, ..., N\}\big)_{j=1}^{J}$
    $\boldsymbol{\pi} \leftarrow \boldsymbol{\pi}' \sim \text{Dirichlet}\left(\left[\boldsymbol{\alpha}_{0k_j} + \sum_{j=1}^{J} A_{j,k_j}\right]_{k_j=1}^{N}\right)$
  end if
end for
Return $\boldsymbol{\theta}$, $\boldsymbol{\phi}$, $\boldsymbol{\pi}$, $\mathbf{A}$, $(\ell_m)_{m=1}^{M}$, and $(\boldsymbol{S}_m)_{m=1}^{M}$

Synthetic Experiments & Results

We designed 9 synthetic experiments using two generative models: (1) EBM-native with ordinal stages and (2) a sigmoid model with continuous stages. Experiments varied stage distributions (uniform or skewed), biomarker distributions (normal, non-normal, or sigmoid), participant sizes ($J = 50, 200, 500, 1000$) and healthy ratios ($r = 0.1, 0.25, 0.5, 0.75, 0.9$). In total, we generated 9,000 datasets (9 experiments × 4 participant sizes × 5 healthy ratios × 50 repetitions). We evaluated algorithm performance using two main metrics: (1) the accuracy of biomarker ordering and (2) the accuracy of patient staging.

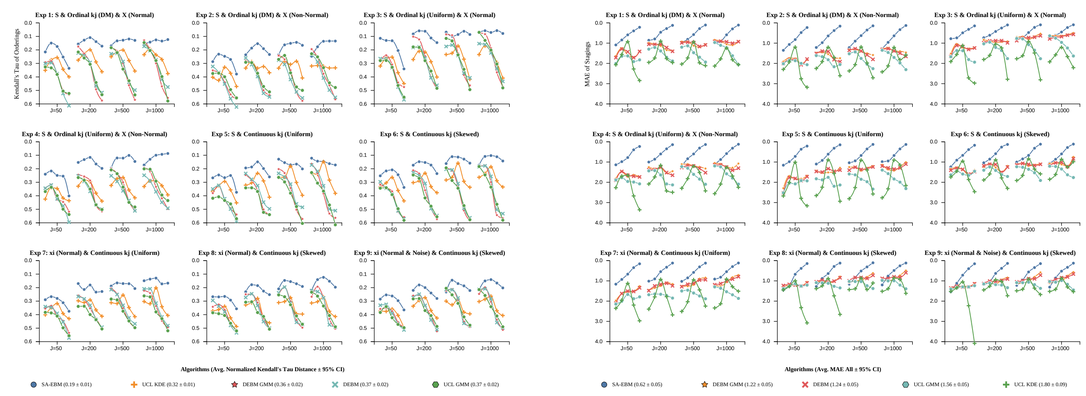


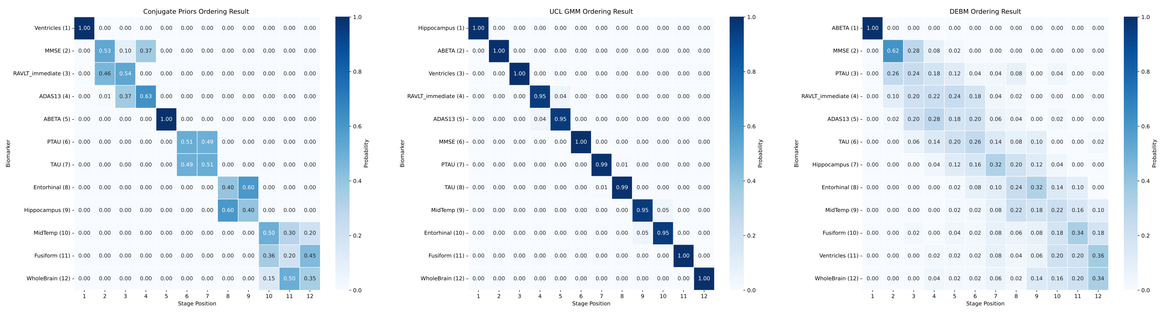
Figure 2. **Left:** Average normalized Kendall's Tau distance values (±95% CI) for ordering tasks. **Right:** Mean average errors (±95% CI) for staging accuracy. Results shown for all algorithms across nine synthetic experiments.

- SA-EBM shows superior robustness across challenging scenarios compared to prior EBM.
- Greatest advantages with high healthy participant ratios - crucial for rare disease studies where recruiting affected participants is difficult.
- Reliable modeling possible with limited sample sizes.
- Gaussian assumptions outperform KDE in limited data scenarios due to bias-variance tradeoff.
- SA-EBM performs well even with continuous event times and subject-specific variations.

ADNI

We applied our SA-EBM algorithms to real-world data from the Alzheimer's Disease Neuroimaging Initiative (ADNI).

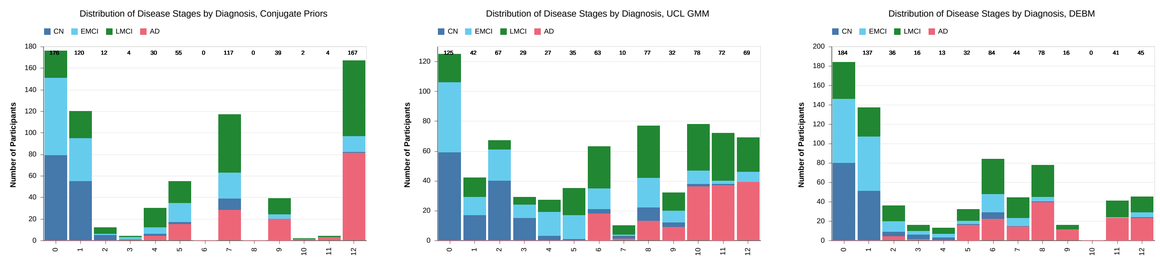
The result indicates that ventricular enlargement occurs first, followed by cognitive decline (MMSE, RAVLT Immediate, and ADAS). Next, pathology in $A\beta_{1-42}$ protein and the two Tau-related biomarkers. Neurodegeneration in brain regions—including the Entorhinal cortex, Hippocampus, MidTemporal area, Fusiform gyrus, and WholeBrain—occurs last. This is essentially Ventricles → Cognition (C) → Amyloid (A) → Tau (T) → Neurodegeneration (N), partially aligns with ATNC ordering from NIA-AA Research Framework (Jack CR Jr, et al., 2018).



(a) SA-EBM (b) UCL GMM (c) DEBM

Figure 3. Ordering results on ADNI by SA-EBM, and two benchmark algorithms

The staging outcome supports that our result is realistic:



(a) SA-EBM (b) UCL GMM (c) DEBM

Figure 4. Staging results on ADNI by SA-EBM, and two benchmark algorithms

Limitations

- Single global progression assumption: Assumes one shared biomarker sequence across all participants, missing real-world disease complexity.
- Ordinal focus only: Models event order without temporal intervals between events.
- Limited subject-level variation: May not fully account for individual differences in disease progression patterns.

Future Work

Mixed Pathology EBM: We are exploring an EBM model for the mixed pathology, for example, when there are more than one diseases affecting patients simultaneously. Essentially, we are looking for a mathematical formulation of $P(\sigma|\sigma^{(1)}, \sigma^{(2)}, ..., \sigma^{(K)})$:

$$P(\sigma|\sigma^{(1)}, \sigma^{(2)}, ..., \sigma^{(K)}) \propto \exp(-E(\sigma))$$

Subtype EBM: We are interested in applying SA-EBM to the subtype problem, i.e., when a disease has different progressions.