sEBM: scaling Event Based Models to predict disease progression via implicit biomarker selection and clustering

Raghav Tandon[1,2][0000−0003−2603−4930], Anna Kirkpatrick[1,3][0000−0002−8737−1132], and Cassie S. Mitchell[1,2][0000−0002−5472−6355]

[1] Laboratory for Pathology Dynamics, Department of Biomedical Engineering, Georgia Institute of Technology and Emory University School of Medicine, Atlanta, GA 30332, USA
[2] Center for Machine Learning, Georgia Institute of Technology, Atlanta, GA 30332, USA
[3] School of Mathematics, Georgia Institute of Technology, Atlanta, GA 30332, USA
{raghav.tandon@gatech.edu, akirkpatrick3@gatech.edu, cassie.mitchell@bme.gatech.edu}

Abstract. The Event Based Model (EBM) is a probabilistic generative model to explore biomarker changes occurring as a disease progresses. Disease progression is hypothesized to occur through a sequence of biomarker dysregulation "events". The EBM estimates the biomarker dysregulation event sequence. It computes the data likelihood for a given dysregulation sequence, and subsequently evaluates the posterior distribution on the dysregulation sequence. Since the posterior distribution is intractable, Markov Chain Monte-Carlo is employed to generate samples under the posterior distribution. However, the set of possible sequences increases as $N!$ where N is the number of biomarkers (data dimension) and quickly becomes prohibitively large for effective sampling via MCMC. This work proposes the "scaled EBM" (sEBM) to enable event based modeling on large biomarker sets (e.g. high-dimensional data). First, sEBM implicitly selects a subset of biomarkers useful for modeling disease progression and infers the event sequence only for that subset. Second, sEBM clusters biomarkers with similar positions in the event sequence and only orders the "clusters", with each successive cluster corresponding to the next stage in disease progression. These two modifications used to construct the sEBM method provably reduces the possible space of event sequences by multiple orders of magnitude. The novel modifications are supported by theory and experiments on synthetic and real clinical data provides validation for sEBM to work in higher dimensional settings. Results on synthetic data with known ground truth shows that sEBM outperforms previous EBM variants as data dimensions increase. sEBM was successfully implemented with up to 300 biomarkers, which is a 6-fold increase over previous EBM applications. A real-world clinical application of sEBM is performed using 119 neuroimaging markers from publicly available Alzheimer's Disease Neuroimaging Initiative (ADNI) data to stratify subjects into 6 stages of disease progression.

Subjects included cognitively normal (CN), mild cognitive impairment (MCI), and Alzheimer's Disease (AD). sEBM stage is differentiated for the 3 groups ($\chi^2 p - value <$ 4.6e-32). Increased sEBM stage is a strong predictor of conversion risk to AD ($p - value < 2.3e - 14$) for MCI subjects, as verified with a Cox proportional-hazards model adjusted for age, sex, education and APOE4 status. Like EBM, sEBM does not rely on apriori defined diagnostic labels and only uses cross-sectional data.

Keywords: disease progression modeling · bayesian learning · prognostic biomarker selection · biomarker clustering

1 Introduction

A popular approach to disease progression modeling is the event-based model (EBM) [1, 2]. EBM hypothesizes disease progression to occur through a sequence of discrete events, which correspond to biomarker abnormalities without reliance on a priori diagnostic labels or explicit biomarker cut-off or threshold values. The model infers the sequence of events most consistent with data measured from clinical subjects. It has been applied to cross-sectional data from sporadic and familial Alzheimer's disease (AD) [3] , Huntington's disease [4], epilepsy [5] and progressive supranuclear palsy [6] to name a few.

The ability to work with cross-sectional data (single measurement per individual) and ease of integrating multiple biomarker types (imaging volumes, fluid, cognitive, etc.) makes EBM a useful model to study disease progression. However, a key challenge to the EBM approach is its scalability to larger biomarker sets. The possible number of event sequence increase as $N!$ (N being the number of biomarkers).

This work presents scaled EBM (sEBM) as an improved event based model that overcomes the challenges of factorial increases in possible event sequences, which inherently occurs in data sets with a large number of biomarkers (high dimensionality data). Throughout this work the classic event-based model in [2] is referred to as EBM, and the modified EBM model based on this work as sEBM (scaled EBM). Permutation complexity refers to the possible number of distinct event-ordering sequences for either of these models.

2 Background

2.1 Event-based model (EBM)

EBM for familial cases of Alzheimer's Disease was proposed in [1] and later generalized to sporadic cases in [2], estimates the ordering of biomarker dysregulation events. The model consists of a set of events $E_1, \dots E_N$ and an ordering $S = (s(1) \dots s(N))$ which is a permutation of the integers $1, \dots N$ determining the event ordering $E_{s(1)}, \dots E_{s(N)}$ representing the sequence of biomarker dysregulation.

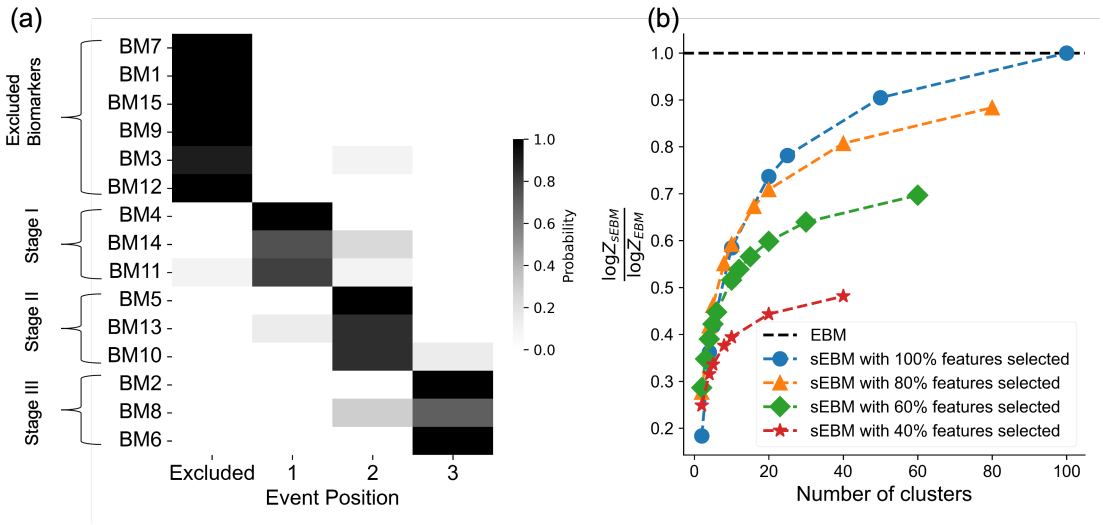

Fig. 1: (a) schematic output from sEBM and (b) reduction in permutation complexity (for the case with equal sized clusters) which can be used to guide the choice of features to be included and number of clusters.

If patient j (whose biomarker measurements are denoted by $X_j = \{x_{1j}, x_{2j} \ldots x_{Nj}\}$ and the full data $X \in \mathbb{R}^{J \times N}$) is at stage k in the progression model, events $E_{s(1)}, \ldots, E_{s(k)}$ have occurred, while events $E_{s(k+1)}, \ldots, E_{s(N)}$ have not. Data likelihood for that patient given an event-ordering S can be written as

$$p(X_j|k,S) = \prod_{i=1}^{k} p(x_{s(i)j}|E_{s(i)}) \times \prod_{i=k+1}^{N} p(x_{s(i)j}|\neg E_{s(i)}) \tag{1}$$

Since $X_j's$ are independent, the complete data likelihood can be written while marginalizing out the stage k of individual subjects as -

$$p(X|S) = \prod_{j=1}^{J} \sum_{k=0}^{N} p(k) \prod_{i=1}^{k} p(x_{s(i)j}|E_{s(i)}) \times \prod_{i=k+1}^{N} p(x_{s(i)j}|\neg E_{s(i)}) \tag{2}$$

Bayes' rule can be used to derive a posterior on the event ordering S as

$$p(S|X) = \frac{p(X|S)p(S)}{p(X)} \tag{3}$$

Model Assumptions The EBM model in [1, 2] makes strong assumptions that are inherited by sEBM. These are - 1) All subjects follow the same disease progression trajectory. Note this assumption has been relaxed in [14, 15]. Similar frameworks can be coupled with the proposed sEBM to overcome the rigidity

of this assumption and its impact on modeling heterogeneous disease. 2) Independence of likelihood of each biomarker measurement is conditional on event occurrence. 3) Biomarker changes are monotonic and 4) population under study is uniformly sampled across disease stages.

Mixture models for computing data likelihood conditioned on event occurrence Computing eq. (2) requires separate models for $p(x_{s(i)j}|E_{s(i)})$ and $p(x_{s(i)j}|\neg E_{s(i)})$ which is obtained by fitting a two component Gaussian mixture model [2] to all observations of biomarker i in the data ($\{x_{ij}|j = 1 \dots J\}$). The components of the mixture model correspond to $p(x_{ij}|E_i)$ and $p(x_{ij}|\neg E_i)$.

Sampling event ordering S using MCMC Since the posterior $p(S|X)$ is intractable, MCMC (Metropolis-Hastings) can be used to generate samples. The MCMC algorithm proceeds as follows - at each iteration, the current ordering S_t swaps two randomly chosen biomarkers to generate a proposed ordering S'. This proposed ordering will be accepted ($S_{t+1} = S^{'}$) with a probability $\min(1, \frac{p(X|S^{'})}{p(X|S_t)})$. The initial state is selected using a greedy-ascent algorithm which accepts a proposed ordering only when $\frac{p(X|S^{'})}{p(X|S_t)} > 1$.

2.2 Challenges due to data dimensionality and proposed solutions

The support of the distribution for $P(S|X)$ increases as $N!$ where N is the dimensionality of X. This can lead to severe challenges for effective MCMC sampling of the underlying distribution. Two solutions are proposed within sEBM -

1. Implicit feature selection is performed within sEBM model to subset useful biomarkers and inferring an event ordering S only over them.
2. sEBM associates multiple biomarkers with a single event, instead of a unique event per biomarker. Such clustering of biomarkers into fewer events makes permutations within a cluster no longer count, thereby removing a large number of closely related but distinct sequences.

3 Method

3.1 Implicit biomarker selection within sEBM

Let $C_S \in \{0, 1\}^N$ such that $||C_S||^2 = f$ (f out of the N biomarkers are chosen for event ordering). The event ordering S will be a permutation of the integers $1, \dots, f$. Since both C_S and S are to be inferred, the posterior $p(S|X)$ in eq. (3) is modified to include C_S. The new posterior becomes $p(C_S, S|X)$ in eq. (4).

$$p(C_s, S|X) = \frac{p(X|C_s, S)p(C_s, S)}{p(X)} \tag{4}$$

Similar to [1, 2], we assume a uniform prior on $P(C_S, S)$. In eq. (4), $p(X|C_S, S)$ can be written as -

$$\begin{aligned} p(X|C_S, S) = p(X_{\setminus S}, X_S|C_S, S) =& p(X_S|C_S, S)p(X_{\setminus S}|X_S, C_S, S) \\ =& p(X_S|S)p(X_{\setminus S}|X_S) \end{aligned} \tag{5}$$

In eq. (5), $X_S \in \mathbb{R}^{J\times f}$ refers to the data subset corresponding to the selected f biomarkers whereas $X_{\setminus S} \in \mathbb{R}^{J\times(N-f)}$ corresponds to the data subset of $N - f$ excluded biomarkers (from J individuals). $p(X_S|C_S, S)$ can be simplified to $p(X_S|S)$ (knowing S fixes C_S) and $p(X_{\setminus S}|X_S, C_S, S)$ can be simplified to $p(X_{\setminus S}|X_S)$ since $X_{\setminus S} \perp\!\!\!\perp C_S, S|X_S$. A simple way to look at this independence relation is that no additional information for $X_{\setminus S}$ is derived from the choice of biomarkers (C_S) or their event ordering (S), since they only pertain to X_S.

MCMC samples for $p(C_S, S|X)$ can be generated from eq. (4) by substituting the expression for $p(X|C_S, S)$ from eq. (5) and proceeding as follows. At each iteration t, the current state (C_{S_t}, S_t) is perturbed to $(C_{S'}, S')$ by randomly swapping two biomarkers. This swap can either change the event ordering alone or lead to a new set of chosen biomarkers with an event ordering similar to S_t but the included biomarker replacing the excluded biomarker in the proposed event ordering S'. This perturbation is accepted ($S_{t+1} = S'$) with a probability $\min(1, a)$, where $a = \frac{p(X_{S'}|S')p(X_{\setminus S'}|X_{S'})}{p(X_{S_t}|S_t)p(X_{\setminus S_t}|X_{S_t})}$. Each iteration requires two terms to be computed - $p(X_{S'}|S')$ and $p(X_{\setminus S'}|X_{S'})$. These terms can be interpreted as follows -

- $p(X_{S'}|S')$ - The data likelihood over included biomarkers $X_{S'}$ given the chosen event ordering S'. This can be computed as in eq. (2) (or eq. (6) which will be introduced in 3.2)
- $p(X_{\setminus S'}|X_{S'})$ - The data likelihood for the excluded biomarkers given the included biomarkers. It is assumed here that $p(X_{\setminus S_t}|X_{S_t})$ and $p(X_{\setminus S'}|X_{S'})$ are approximately equal since they may differ at most by one out of N data features.

3.2 Implicit biomarker clustering within EBM

The biomarkers used for event ordering can be clustered together into sets with pre-defined sizes using a simple modification on the original EBM. Equation (2) computes the data likelihood by marginalizing all possible disease stages which could be any stage between no biomarker dysregulated ($k = 0$) and all biomarkers dysregulated (final disease stage, $k = N$). Instead of ordering all biomarkers by giving them a unique position along the event cascade, biomarkers could be clustered together. All biomarkers within the same cluster share the same position in the ordering. Positional differences only occur across clusters. This can be introduced in the EBM framework (eqs. (1) and (2)), by constraining k to only take on specific values which are a function of the individual cluster

sizes. Let c_i denote the i^{th} cluster which consists of $|c_i|$ biomarkers. Let the total number of clusters be n. Hence the vector of cluster sizes can be written as $[|c_1|\dots|c_n|]$ and $\sum_{i=1}^{n}|c_i| = N$. Let $u_z = \sum_{i=1}^{z}|c_i|$ which represents the cumulative sum of cluster sizes until cluster c_z $(1 \leq z \leq n)$ and $U = [0, u_1 \dots u_n]$. Equation (2) can be constrained to assign events to the clusters c_i instead of finding an event ordering over all biomarkers as shown in eq. (6) -

$$p(X|S) = \prod_{j=1}^{J} \sum_{k \in U} p(k) \prod_{i=1}^{k} p(x_{s(i)j}|E_{s(i)}) \prod_{i=k+1}^{N} p(x_{s(i)j}|\neg E_{s(i)}) \tag{6}$$

$k \in U$ represents further discretization of EBM where disease stage advances from no biomarker abnormality ($k = 0$) to all biomarkers abnormal (k=N) in steps of $|c_1|, |c_2| \dots |c_{n-1}|, |c_n|$ biomarkers. This results in fewer possible states to be explored by the MCMC sampler, fewer terms in the marginalization over individual's stage (k), and faster run-times per iteration. The only constraint on c_i is that $1 \leq |c_i| \leq N$, $\sum_{i=1}^{n}|c_i| = N$, and c_i are disjoint.

3.3 Analysis of reduction in permutational complexity for MCMC sampling via biomarker subset selection and clustering

Equation (4) requires MCMC sampling over (C_S, S) instead of just S which is the case in eq. (3). The joint inference over two variables can be easier under certain conditions which are explained here. The possible number of orderings for N biomarkers is $N!$ which can be significantly reduced by clustering biomarkers and only looking at orderings across clusters. This can be easily seen in the case where the N biomarkers are divided equally among c clusters so that each cluster gets m biomarkers ($c \times m = N$ and $N, c, m \in \mathbb{Z}^+$). Equal sized clusters is not a requirement but helps in presenting our case. The possible number of orderings with equal sized clusters is -

$$\underbrace{\binom{N}{m}\binom{N-m}{m}\binom{N-2m}{m}\cdots\binom{m}{m}}_{c \text{ clusters}} = \frac{N!}{m!^c} \tag{7}$$

Further, if a subset of f biomarkers can be selected from N biomarkers such that only those f are used for event ordering, the number of total possible orderings will be $\binom{N}{f}\frac{f!}{m'!^c}$. Assuming the same number of c clusters as in eq. (7), the number of biomarkers in this case is given by m' $(c \times m' = f)$, and $f, c, m' \in \mathbb{Z}^+$. The feature selection step increases the complexity by a factor of $\binom{N}{f}$ but simultaneously reduces the complexity for ordering the selected biomarkers to $\frac{f!}{m'!^c}$ due to clustering. By choosing f and c appropriately for a given N, it is possible to significantly reduce the overall permutation complexity as shown in fig. 1b. The y-axis shows the log of permutation complexity under the new model $(Z_{sEBM} = \binom{N}{f}\frac{f!}{m'!^c})$ divided by the log of permutation complexity from EBM $(Z_{EBM} = N!)$, at different f and c (shown for $N = 100$).

3.4 Hyperparameter selection

There are three important hyperparameters in the sEBM model. 1) Fraction of biomarkers to be included by the model, is recommended to be 0.5 or upwards. The selected biomarkers encode for the excluded biomarkers, thereby allowing their exclusion. Including sufficient number of biomarkers helps in simplifying the acceptance ratio for the MCMC associated with eqs. (4) and (5). 2) The number of clusters implies the number of disease stages. Popular disease staging procedures can be used as motivation for cluster selection. For example the Braak criterion [11] divides AD into 6 stages. This motivates 5 clusters, as was used in this work in section 4.2. 3) Size of individual clusters. Making individual clusters larger while keeping the number of clusters constant will reduce the underlying permutational complexity. However, the associated risk is lesser information about disease stages with smaller clusters associated to them. Future work will focus on a well formed methodology for hyperparameter selection.

4 Experiments

4.1 Synthetic data with known ground truth

sEBM is tested on synthetic data generated using the data simulation framework in [7]. The simulation framework generates cross-sectional biomarker data sets from neurodegenerative disease cohorts that reflect the temporal evolution of the disease. The simulation allows for generating data with known ground-truth event ordering which can be compared against an inferred event ordering. The simulation framework assumes that the temporal evolution of the biomarkers is sigmoidal and can be modeled with the equation-

$$z(t, \theta_i) = a_i + \frac{r_i}{1 + \exp\left(-\frac{4}{\tau_i}(t - c_i)\right)} \tag{8}$$

with parameters $\theta_i = (a_i, r_i, \tau_i, c_i)$ for the i^{th} biomarker. a_i is the trajectory minimum for the i^{th} biomarker and $a \sim \mathcal{N}(\mathbf{0}, \Sigma_\alpha)$ with $a_i = a[i]$. Σ_α is a symmetric positive semi-definite matrix to model covariance between a_i (α represents the fraction of non-zero entries in Σ_α and is set to 0.5). $r_i \in [1, 3]$ is the range, $\tau_i \in [3, 6]$ is the gradient, $c_i \in [2, 18]$ is the inflection point. c_i, r_i, τ_i are sampled uniformly from their respective ranges. t represents the time point in the progression trajectory of the individual, and is sampled uniformly from the range [0, 20]. A small value for c_i implies that the biomarker changes early on in disease progression and vice-versa. t indicates how far the subject has progressed into the disease and is sampled uniformly across the disease timeline. The simulation dataset are generated using eq. (8) for 3 different sample sizes (200, 400, 600) and 6 different dimensions (50, 100, 150, 200, 250, 300). At each setting of sample size and dimensions, 6 different dataset are generated using random seeds resulting in 108 ($3 \times 6 \times 6$) different datasets.

Model Baseline Comparisons sEBM is compared to three other methods on 108 datasets - EBM [2], discriminative EBM [12] and EBM run on data with 50% features removed (largest p-values) using a Mann-Whitney U test. The dEBM model is fit to data using the pyebm package in python, using default parameter choices (the gaussian mixture model algorithm it uses has been shown to be more stable in [12]).

Hyperparameters In all settings, sEBM is set to implicitly select half of the biomarkers. The number of clusters are set to 5, and cluster sizes are set to be equal in order to facilitate normalization of the Kendall-Tau distance metric (explained below). The underlying MCMC is initialized using the best ordering (highest likelihood) from among 100 greedy initialization run for 800 iterations each. MCMC is then used to generate 2e6 samples of the underlying ordering, of which the first 1.5e6 are discarded as burn-in for mixing of the markov chain.

Comparison to ground-truth ordering Since ground-truth ordering cannot be known with real cross-sectional data, simulating cross-sectional data with a chosen ground-truth and inferring this ordering using sEBM for comparison is an important exercise. The inferred ordering and the ground-truth ordering are compared by using the Kendall-Tau metric for partial rankings as specified in [8]. It is similar to the general Kendall-Tau distance for full-rankings except that it also accounts for pairs of biomarkers which were assigned same rank in one ordering, but different ranks in the other ordering.

$$d_k^{(p)}(\pi_*, \pi_{gt}) = \sum_{l \prec_{\pi_{gt}} j} \mathbb{1}_{l \succ_{\pi_*} j} + \; p \sum_{l \cong_{\pi_{gt}} j} \mathbb{1}_{l \not\cong_{\pi_*} j} + \; p \sum_{l \not\cong \pi_{gt} j} \mathbb{1}_{l \cong_{\pi_*} j} \tag{9}$$

A few things are to be noted about eq. (9). It is a distance metric for $p \geq 0.5$ [8], hence in this work we fix $p = 0.5$. $d_k^{(p)}$ can be normalized by finding the maximum distance from the ground-truth sequence, which is obtained by the reverse of the ground-truth sequence ($\pi_{gt}^R(i) = n + 1 - \pi_{gt}(i)$, where $\pi_{gt}^R(i)$ is the position of element i in π_{gt}^R and n is the number of unique positions in the rankings). However, finding π_{gt}^R is not straightforward when there are unequal number of biomarkers at each position. This provides a motivation to use equal sized clusters in our experiments. Last, we convert any full-rankings into partial rankings using specified cluster sizes (for e.g. EBM ordering). This is to facilitate direct comparison with the partial-ranked outputs of sEBM. While eq. (9) can be easily generalized to full-ranks where it is the same as the adjacent swap distance, that no longer holds true for partial ranks [9]. Hence in all cases, only normalized Kendall-Tau distance for partial-ranks is used. Last, the normalized Kendall-Tau distance is computed using the selected sEBM biomarkers from the maximum likelihood ordering among all MCMC generated samples.

4.2 Real Clinical Data to Assess sEBM

Data previously made available in the TADPOLE benchmarking challenge [10] is analysed using sEBM. Specifically, file *dfMri_D12.csv* is used. The first recorded observation of all subjects is divided into a training set (CN and AD, n=327), and validation set (MCI, n=551). The data had 123 features, of which 119 were used by removing 4 which had a constant zero value for all subjects in the training and validation sets. The longitudinal follow-up data from these subjects is separately used to study their conversion risk.

Hyperparameters Number of clusters are set to 5 (as in section 4.1), resulting in 6 stages which aligns with Braak staging [11]. sEBM is set to implicitly select approximately half of the data features (60/119) and divide them equally across 5 clusters (12 features per clusters). The MCMC settings are the same as the one used in experiments with synthetic data (section 4.1).

5 Results

5.1 sEBM outperforms baseline EBM models with synthetic data

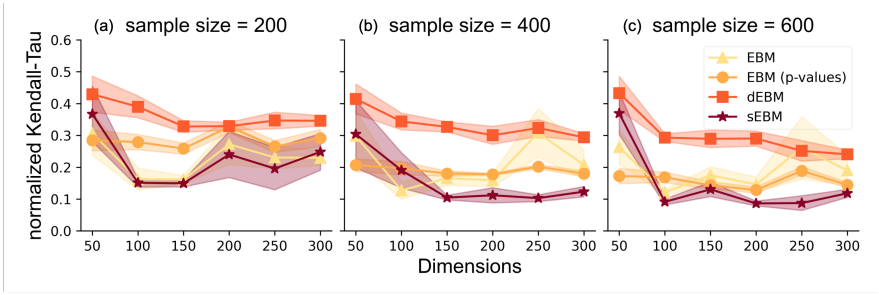

Fig. 2: **Performance on simulated data with known ground truth -** sEBM is compared to three different methods- EBM, discriminative EBM (dEBM), and EBM (p-values) wherein half the features are pre-selected using the Mann Whitney U test. sEBM shows better scaling ability with data dimensions.

Figure 2 shows sEBM scales better with increasing data dimensions, in comparison to EBM [2], discriminative EBM (dEBM) [12], and EBM (p-values) which only used half the features to begin with. The shaded regions show the standard error on the mean from 6 datasets for each sample $\times$ dimension setting. sEBM and all other methods also outperformed [13] (not shown).

5.2 sEBM shows distinct stages for CN, MCI, and AD subjects.

sEBM is used to infer stages on all training subjects (CN and AD subjects, n=327), and validation subjects (MCI subjects, n=551) with stages showing difference across the three classes in **fig. 3a** (χ^2 p-value $< 4.6e-32$).

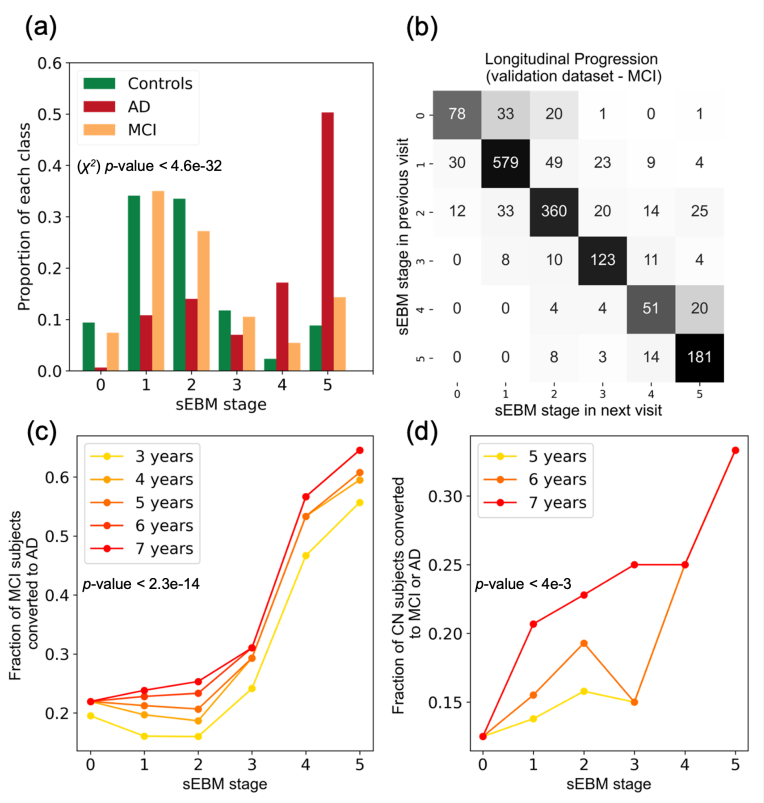

Fig. 3: (a) sEBM assigned stages for CN, MCI and AD subjects (b) show consistency in predicting stages for successive visits (c) For MCI subjects, are strongly associated with risk of converting to AD (d) Same as c, but for CN subjects.

5.3 Longitudinal validation shows consistency of assigned staging

The maximum-likelihood event ordering derived from sEBM is used to stage the follow-up visits from all subjects. Subject staging shows longitudinal consistency with most follow-up visits showing same stages or gradual progression to neighboring stages (**fig. 3b**). While sEBM works with cross-sectional data, it can be extended to longitudinal data using a temporal version of EBM presented in [16].

5.4 sEBM accurately predicts progression, conversion to AD

sEBM requires a single clinical baseline measurement from subjects to assign them a stage, which corresponds to disease severity. As expected, CN dominated early stages, AD dominated later stages, and MCI were spread between these two (**fig. 3a**). Longitudinal follow-up data shows if the subject's condition worsens at a later visit (e.g. CN to MCI or MCI to AD). Despite not using longitudinal data, sEBM accurately predicts conversion risk to MCI or AD (**fig. 3c**). Conversion (3-year) from MCI to AD is modeled using a Cox proportional hazards model using sEBM stage, age, gender, education and APOE4 status (p-value for sEBM stage $<$ 2.3e-14 and Hazard ratio = 1.57). Similar patterns are seen for conversion (5-year) of CN subjects to MCI or AD in **fig. 3d** (p-value for sEBM stage $<$ 4e-3 and Hazard ratio = 1.53).

5.5 sEBM inferred event ordering is clinically meaningful

Stages 1 and 2 indicates early changes to emotional and reward processing regions (nucleus accumbens), cardiovascular risk (right, left vessel), and autonomic and endocrine projections (cingulate), and the classic onset of semantic and non-verbal language processing (pars triangularis, right frontal pole). Stage 3 illustrates changes to cerebrospinal fluid homeostasis (chloroid plexus, third ventricle, CSF) and deeper language processing issues (pars orbitalis). In stage 4, overt memory losses correlate with decreased temporal, entorhinal, and hippocampal volumes, which are commonly associated with AD. Stage 5 indicates stronger ties to reward, emotion, and speech processing via further degeneration of orbitofrontal, temporal lobe structures, amygdala, and the putamen. The excluded features largely contain either highly correlated features functionally represented in the included feature set (e.g. other ventricles) or features that are less associated with AD (white matter, brain stem). Overall, the sEBM predictions align with clinical intuition of AD progression [17]. The exclusion of the cortex volumes was interesting and unexpected given loss of cortex is common in aging and changes of cognition [18].

6 Conclusions and Future Work

sEBM comprised two modifications to EBM to make it scalabale with increasing data dimensions - 1) implicit feature selection and; 2) clustering biomarkers into fewer event positions. This drastically reduced the permutation complexity of the underlying MCMC sampling, improved its mixing, and expedited sample generation from the posterior distribution of event orderings. sEBM showed superior performance on synthetic data with known ground-truth in comparison to other similar methods [2, 12, 13]. On real clinical biomarkers derived from brain MRI, the assigned sEBM stages were well-separated for CN and AD, predictive of conversion risk in MCI subjects, and stable across follow-up clinical visits. While sEBM has some restrictive assumptions such as homogeneity across subjects in disease progression trajectory, these assumptions are inherited from the previous work [1, 2] and can be addressed using methods presented in [14, 15].

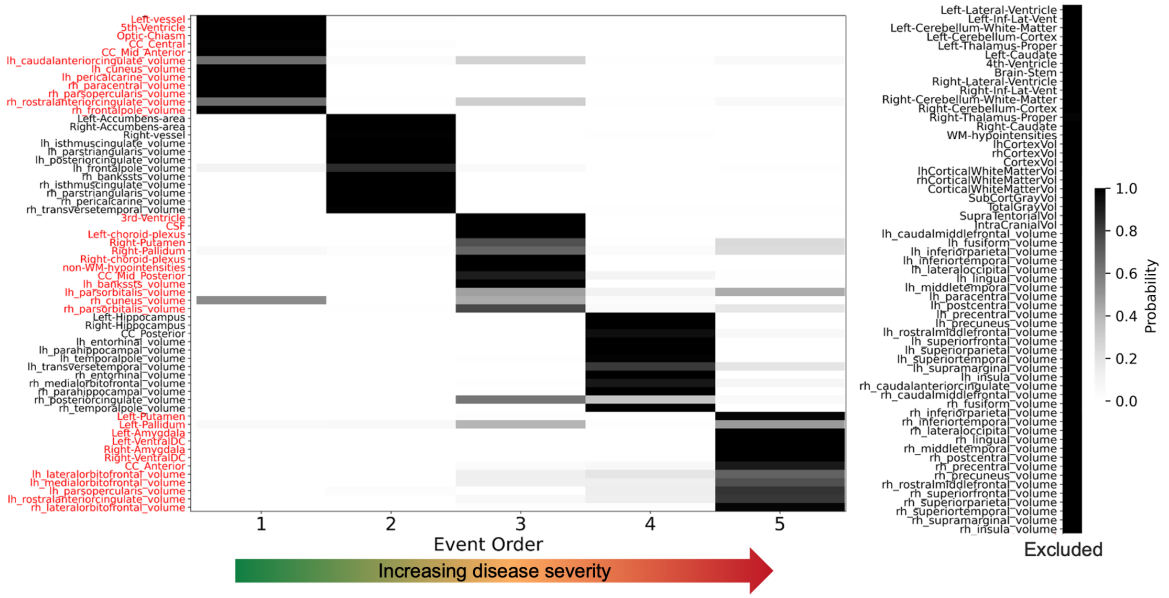

Fig. 4: maximum likelihood event ordering by sEBM

Acknowledgements Support for this work was provided by National Science Foundation grant 1944247 and National Institutes of Health grants U19AG056169 and 5R01AG070937 to C.M.

References

1. Fonteijn, Hubert M., Marc Modat, Matthew J. Clarkson, Josephine Barnes, Manja Lehmann, Nicola Z. Hobbs, Rachael I. Scahill et al. "An event-based model for disease progression and its application in familial Alzheimer's disease and Huntington's disease." NeuroImage 60, no. 3 (2012): 1880-1889.
2. Young, Alexandra L., 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, no. 9 (2014): 2564-2577.
3. Oxtoby, Neil P., Alexandra L. Young, David M. Cash, Tammie LS Benzinger, Anne M. Fagan, John C. Morris, Randall J. Bateman, Nick C. Fox, Jonathan M. Schott, and Daniel C. Alexander. "Data-driven models of dominantly-inherited Alzheimer's disease progression." Brain 141, no. 5 (2018): 1529-1544.
4. Byrne, Lauren M., Filipe B. Rodrigues, Eileanor B. Johnson, Peter A. Wijeratne, Enrico De Vita, Daniel C. Alexander, Giuseppe Palermo et al. "Evaluation of mutant huntingtin and neurofilament proteins as potential markers in Huntington's disease." Science Translational Medicine 10, no. 458 (2018): eaat7108.

5. Lopez, Seymour M., M. Aksman, Neil P. Oxtoby, Sjoerd B. Vos, Jun Rao, Erik Kaestner, Saud Alhusaini et al. "Event-based modeling in temporal lobe epilepsy demonstrates progressive atrophy from cross-sectional data." Epilepsia (2022).
6. Scotton, W. J., M. Bocchetta, E. Todd, D. M. Cash, N. Oxtoby, L. VandeVrede, H. Heuer et al. "A data-driven model of brain volume changes in progressive supranuclear palsy." Brain communications 4, no. 3 (2022): fcac098.
7. Young, Alexandra L., Neil P. Oxtoby, Sebastien Ourselin, Jonathan M. Schott, Daniel C. Alexander, and Alzheimer's Disease Neuroimaging Initiative. "A simulation system for biomarker evolution in neurodegenerative disease." Medical image analysis 26, no. 1 (2015): 47-56.
8. Fagin, Ronald, Ravi Kumar, Mohammad Mahdian, D. Sivakumar, and Erik Vee. "Comparing partial rankings." SIAM Journal on Discrete Mathematics 20, no. 3 (2006): 628-648.
9. Cicirello, Vincent A. "Kendall tau sequence distance: Extending Kendall tau from ranks to sequences." arXiv preprint arXiv:1905.02752 (2019).
10. Marinescu, Razvan V., Neil P. Oxtoby, Alexandra L. Young, Esther E. Bron, Arthur W. Toga, Michael W. Weiner, Frederik Barkhof et al. "The alzheimer's disease prediction of longitudinal evolution (TADPOLE) challenge: Results after 1 year follow-up." arXiv preprint arXiv:2002.03419 (2020).
11. Braak, Heiko, and Eva Braak. "Neuropathological stageing of Alzheimer-related changes." Acta neuropathologica 82, no. 4 (1991): 239-259.
12. Venkatraghavan, Vikram, et al. "Disease progression timeline estimation for Alzheimer's disease using discriminative event based modeling." NeuroImage 186 (2019): 518-532.
13. Firth, Nicholas C., 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, no. 7 (2020): 965-973.
14. Young, Alexandra L., Neil P. Oxtoby, Jonathan Huang, Razvan V. Marinescu, Pankaj Daga, David M. Cash, Nick C. Fox et al. "Multiple orderings of events in disease progression." In International Conference on Information Processing in Medical Imaging, pp. 711-722. Springer, Cham, 2015.
15. Young, Alexandra L., Razvan V. Marinescu, Neil P. Oxtoby, Martina Bocchetta, Keir Yong, Nicholas C. Firth, David M. Cash et al. "Uncovering the heterogeneity and temporal complexity of neurodegenerative diseases with Subtype and Stage Inference." Nature communications 9, no. 1 (2018): 1-16.
16. Wijeratne, Peter A., Daniel C. Alexander, and Alzheimer's Disease Neuroimaging Initiative. "Learning transition times in event sequences: The temporal event-based model of disease progression." Information Processing in Medical Imaging: 27th International Conference, IPMI 2021, Virtual Event, June 28–June 30, 2021, Proceedings. Cham: Springer International Publishing, 2021.
17. Salvatore, Christian, Antonio Cerasa, and Isabella Castiglioni. "MRI characterizes the progressive course of AD and predicts conversion to Alzheimer's dementia 24 months before probable diagnosis." Frontiers in aging neuroscience 10 (2018): 135.
18. Roe JM, Vidal-Piñeiro D, Sørensen Ø, Brandmaier AM, Düzel S, Gonzalez HA, Kievit RA, Knights E, Kühn S, Lindenberger U, Mowinckel AM. "Asymmetric thinning of the cerebral cortex across the adult lifespan is accelerated in Alzheimer's disease." Nature communications. 2021 Feb 1;12(1):1-1.