bioRxiv preprint doi: https://doi.org/10.1101/2022.07.10.499471; this version posted March 18, 2023. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

S-EBM: Generalising event-based modelling of disease progression for simultaneous events

Parker CS[1⊥], Oxtoby NP[1], Alexander DC[1], Zhang H[1] for the Alzheimer's Disease Neuroimaging Initiative[2]

[1] Centre for Medical Image Computing, Department of Computer Science, UCL, London, UK

[2] 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 at: http://adni.loni.usc.edu/wp-content/uploads/how_to_apply/ADNI_Acknowledgement_List.pdf

[⊥] Corresponding author: christopher.parker@ucl.ac.uk. University College London, Centre for Medical Image Computing, 90 High Holborn, Floor 1, London, UK, WC1V6LJ

bioRxiv preprint doi: https://doi.org/10.1101/2022.07.10.499471; this version posted March 18, 2023. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

Abstract

Estimating the temporal evolution of biomarker abnormalities in disease informs understanding of early disease processes and facilitates subject staging, which may augment the development of early therapeutic interventions and provide personalised treatment tools. Event-based modelling of disease progression (EBM) is a data-driven technique for inferring a sequence of biomarker abnormalities, or events, from cross-sectional or short-term longitudinal datasets and has been applied to a variety of different diseases, including Alzheimer's disease. Conventional EBM (C-EBM) assumes the sequence of biomarker abnormalities occurs in series, with one biomarker event per disease progression stage. However, events may occur simultaneously, for example due to the presence of shared causal factors, a property which cannot be inferred from C-EBM. Here we introduce simultaneous EBM (S-EBM), a generalisation of C-EBM to enable estimation of simultaneous events. S-EBM can estimate a wider range of sequence types than C-EBM while being fully backward compatible with the original model. Using simulated data, we firstly demonstrate the inability of C-EBM to infer simultaneous events. We next assess the accuracy of S-EBM against ground truth data and subsequently demonstrate a real-world example application to sequence disease progression in Alzheimer's disease. Simulations show that C-EBM can not discern serial events with high biomarker variance from simultaneous events, preventing its use for inferring simultaneous events. S-EBM has high estimation accuracy against ground truth for a range of sequence types (fully simultaneous, partially simultaneous, serial), number of biomarkers and biomarker variances. When applied to Alzheimer's disease biomarker data from ADNI, S-EBM estimated a sequence where events within sets of biomarker domains occur simultaneously. Accumulation of total and phosphorylated tau in cerebrospinal fluid; performance on RAVLT, ADAS-Cog and MMSE cognitive test scores; and volumetric decline in temporal regional brain volumes, were better described as groups of simultaneous events rather than a single set of serial events (likelihood ratio >> 1,000). Furthermore, C-EBM may be confidently incorrect regarding the serial ordering. S-EBM may be applied to prospective and retrospective biomarker data to refine understanding of disease progression and generate new hypotheses regarding disease aetiology and spread.

bioRxiv preprint doi: https://doi.org/10.1101/2022.07.10.499471; this version posted March 18, 2023. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

1. Introduction

Estimating the temporal progression of biomarker abnormalities throughout the course of a disease identifies biomarkers of early disease, which generates hypotheses regarding disease aetiology and spread; and facilitates subject staging, which may aid the development of therapeutic interventions and personalised treatment.

Disease progression has been previously estimated using hypothesis-driven approaches following literature review or post-mortem examination. For example, the Jack curves (Jack Jr. et al 2010) describe the evolution of biomarkers abnormalities in Alzheimer's disease (AD), and Braak stages were derived from post-mortem examination of AD patients (Braak, H & Braak, E 1991). Although these approaches are informative, they are qualitative in nature. Data-driven approaches are needed for objective assessment of disease spread. In the ideal scenario the temporal trajectory of different biomarkers is derived from longitudinal data acquired throughout the disease course. However, in practice cross-sectional or short-term longitudinal data are the predominant type of biomarker data available. There is therefore a need for approaches that estimate disease progression from such data.

Event-based modelling of disease progression (EBM) is a data-driven approach that estimates the evolution of biomarker abnormalities from cross-sectional or short-term longitudinal data (Fonteijn et al 2012). EBM has been applied to estimate progression of biomarker abnormality in a variety of diseases, including AD (Young et al 2014), Huntington's disease (Wijeratne et al 2018), multiple sclerosis (Eshaghi et al 2018) and amyotrophic lateral sclerosis (Gabel et al 2020).

Underlying the conventional EBM (C-EBM) approach (Fonteijn et al 2012), as well as its recent variants, is the assumption that biomarker abnormalities are ordered serially, i.e. no two biomarkers may become abnormal concurrently. However, biomarker abnormalities may occur simultaneously when they are driven by common causative factors, or be better approximated as simultaneous than as serial when the difference between their temporal trajectories is unresolvably small. Such simultaneous events cannot be inferred from C-EBM as they are excluded from the model by construction. The positional uncertainty that C-EBM estimates may suggest the presence of simultaneous events, but can also simply reflect high variance in biomarker measurements. By not accounting for simultaneous events, C-EBM may incorrectly estimate the sequence and patient staging, limiting its ability to impact disease understanding and therapeutic development.

To overcome this limitation, we introduce simultaneous EBM (S-EBM), a generalisation of C-EBM that can estimate a sequence containing simultaneous events. By allowing simultaneous events, a wider range of disease progression models can be

bioRxiv preprint doi: https://doi.org/10.1101/2022.07.10.499471; this version posted March 18, 2023. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

estimated from any given biomarker data input. In this study, we demonstrate C-EBM's inability to infer simultaneous events, describe the theory of S-EBM and sequence estimation, evaluate the performance of S-EBM against ground truth synthetic data, and provide an example application to sequence evolution of biomarker abnormalities in AD. We show that S-EBM can reliably estimate sequences containing simultaneous events and that such a sequence can better explain the evolution of AD biomarker abnormality.

2. Theory

2.1. Generalising the event-based model

2.1.1. Overview of the conventional event-based model

C-EBM represents the progression of biomarker abnormalities in disease by a sequence, which is an ordered list that encodes the temporal order in which each biomarker undergoes a transition from a normal state to an abnormal state. These transitions, termed events, demarcate the disease progression stages, from which subjects are assumed to be uniformly sampled.

A key assumption of C-EBM is monotonicity of biomarker evolution i.e. that biomarkers transition to an abnormal state but do not subsequently revert. Thus, in the first stage all biomarkers are in a normal state and at each subsequent stage a biomarker transitions to an abnormal state, until the final stage where all biomarkers are abnormal. A further key assumption of C-EBM is that all subjects are sampled from the same disease trajectory. In other words, the set of biomarker measurements for a given subject provides a snapshot of the disease at a particular stage. Furthermore, the subjects are assumed to be sampled from a single disease progression sequence.

C-EBM seeks the sequence with highest posterior probability given the observed biomarker measurements. By assuming an equal prior probability for all possible sequences, this becomes equivalent to the sequence likelihood i.e. the probability of the data given the sequence. As the sequence prescribes the set of events for each disease stage, then given the probability density functions associated with each biomarkers' possible event state (see section 2.4. Event distributions), then the likelihood of the sequence can be evaluated and subsequently maximised across sequence samples. A summary of sequence estimation is shown in Fig. 1.

bioRxiv preprint doi: https://doi.org/10.1101/2022.07.10.499471; this version posted March 18, 2023. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

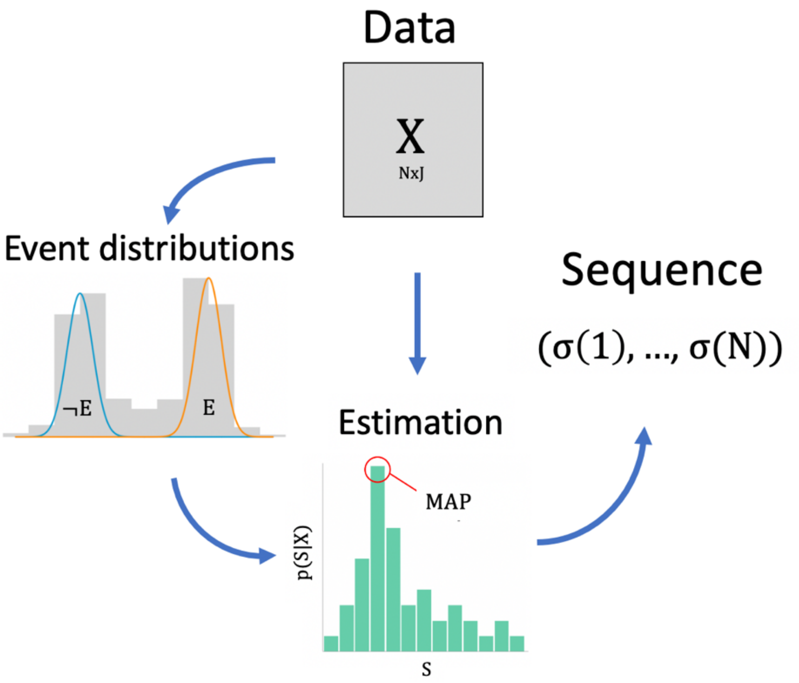


Figure 1. Overview of sequence estimation in C-EBM. C-EBM finds the sequence, S, with maximum posterior probability given the biomarker measurements, X. Given an equal prior probability of each sequence, this is equivalent to the maximum likelihood sequence.

The sequence likelihood is equal to the joint probability of observing the set of subjects' data. Given the input data matrix X, an N-by-J matrix containing N biomarker measurements for J subjects, and assuming that each subject is sampled independently, then the likelihood of the sequence, S, is the product of subject probabilities:

$$p(X|S) = \prod_{j=1}^{J} p(X_j|S) \tag{1}$$

where X_j is a column of X corresponding to the N biomarker measurements for subject j.

As described below, the formulation of $p(X_j|S)$ makes reference to the set of event states' distributions at each disease stage. As these events are defined by S, the formulation of $p(X_j|S)$ depends on the specific form of S. Next, we describe how the sequence is specified and likelihood formulation derived for C-EBM, which assumes the events occur in series, before describing the generalisation of the sequence and likelihood formulation for simultaneous events.

2.1.2. Conventional event-based model: sequence specification and likelihood function

C-EBM specifies the sequence as a permutation of the biomarker indices $1, \dots, N$. Each element of S, $s(i)$, holds the biomarker event occurring at the i'th disease progression stage. For example, for a sequence of four biomarkers a possible sequence is $S = (2,3,4,1)$,

bioRxiv preprint doi: https://doi.org/10.1101/2022.07.10.499471; this version posted March 18, 2023. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

which describes a disease progression where the first biomarker abnormality occurs in biomarker 2, followed by biomarker 3, then biomarker 4 and finally biomarker 1.

With each biomarkers' event states written as $\neg \mathrm{E}$ for normal and E for abnormal, then at a particular stage, k, of the sequence the events have occurred for $\mathrm{E}_{s(i)}, \ldots, \mathrm{E}_{s(k)}$ but have not yet occurred for $\mathrm{E}_{s(k+1)}, \ldots, \mathrm{E}_{s(N)}$. Given independence of biomarker measurements for the combination of events at each sequence position, the subjects' probability given the sequence and stage, k, is written as:

$$p(X_j|S,k) = \prod_{i=1}^{k} p\left(x_{s(i),j} \mid E_{s(i)}\right) \prod_{i=k+1}^{N} p\left(x_{s(i),j} \mid \neg E_{s(i)}\right) \quad (2)$$

Because each subjects' position in the sequence is considered unknown a priori, it is marginalised out over each possible position:

$$p(X_j|S) = \sum_{k=0}^{N} p(k) p(X_j|S,k) \quad (3)$$

The prior probability of each position, $p(k)$, is assumed to be constant and defined as $\frac{1}{N+1}$, where $N+1$ (or equivalently $|S|+1$) is the number of stages. By substituting Eq. 2 into 3, then the total likelihood defined in Eq. 1 is written as:

$$p(X|S) = \prod_{j=1}^{J} \left(\sum_{k=0}^{N} p(k) \left[\prod_{i=1}^{k} p\left(x_{s(i),j} \mid E_{s(i)}\right) \prod_{i=k+1}^{N} p\left(x_{s(i),j} \mid \neg E_{s(i)}\right) \right] \right) \quad (4)$$

Because the sequence can contain only one biomarker event at each position, it cannot represent simultaneous events.

2.1.3. Simultaneous event-based model: sequence specification and likelihood function

To generalise C-EBM for simultaneous events, the sequence specification is updated from an ordered list of biomarker indices to an ordered list of sets. Each set, $s(i)$, contains one or more biomarker indices corresponding to the events at position i in the sequence. For example, for four biomarkers a sequence containing only serial events is written $S = (\{2\},\{1\},\{3\},\{4\})$ and a sequence containing simultaneous events is written $S = (\{2\},\{1,3\},\{4\})$. Given the length of the sequence can vary, the number of positions in the sequence is now defined as $|S|+1$ instead of $N+1$. Therefore, the prior probability of each

bioRxiv preprint doi: https://doi.org/10.1101/2022.07.10.499471; this version posted March 18, 2023. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

position in the sequence is $p(k;\ S) = \frac{1}{|S|+1}$ and the likelihood of each subjects' data given their position is unknown a priori is written as:

$$p(X_j|S) = \sum_{k=0}^{|S|} p(k;\ S)p(X_j|S,k) \qquad (5)$$

As before, the likelihood of each subjects' data given their position k, $p(X_j|S,k)$ is the joint probability over the subjects' biomarker values given each biomarkers event state at that sequence position. For a position k in the sequence, the events have occurred for biomarkers $\cup_{1\le i\le k} s(i)$, whereas the events have not occurred for biomarkers $\cup_{k<i\le|S_m|} s(i)$. Hence, the likelihood of each subjects' data given their position is written as:

$$p(X_j|S,k) = \prod_{\substack{m\in \\ \cup_{1\le i\le k} s(i)}} p(x_{m,j} \mid E_m) \prod_{\substack{m\in \\ \cup_{k<i\le|S|} s(i)}} p(x_{m,j} \mid \neg E_m) \qquad (6)$$

By substituting Eq. 6 into 5, then the total likelihood defined by Eq. 1 is written as:

$$p(X|S) = \prod_{j=1}^{J}\left(\sum_{k=0}^{|S|} p(k;S)\left[\prod_{\substack{m\in \\ \cup_{1\le i\le k} s(i)}} p(x_{m,j}|E_m) \prod_{\substack{m\in \\ \cup_{k<i\le|S|} s(i)}} p(x_{m,j}|\neg E_m)\right]\right) \qquad (7)$$

This likelihood formulation is a fully generalised form of the C-EBM but can represent a wider range of sequence types. In the case of serial events, the likelihood defined in Eq. 7 becomes equal to the C-EBM likelihood defined in Eq. 4.

2.2. Sequence estimation

2.2.1. Conventional event-based model

In C-EBM (Fonteijn et al 2012), the sequence is estimated as the characteristic ordering of biomarker events, which is the average position of each event following Markov Chain Monte Carlo (MCMC) sampling of $p(S|X)$. In subsequent work (Young et al 2014), a stochastic greedy ascent was used to estimate the maximum likelihood sequence. As we aimed to compare the sequence obtained from (Young et al 2014) between C-EBM and S-EBM, this is the approach we adopt here.

bioRxiv preprint doi: https://doi.org/10.1101/2022.07.10.499471; this version posted March 18, 2023. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

The greedy ascent proceeds by iteratively perturbing the sequence and retaining those with higher likelihood for some given number of iterations. At each iteration, a perturbation of the sequence is generated by swapping the positions of two biomarker events. For example, the if the current sequence is $(2,3,4,1)$, then a perturbed sequence can be generated by swapping biomarkers 4 and 2, giving the sequence $(4,3,2,1)$. To prevent dependence of the greedy ascent on the initial random sequence, a number of initialisations are performed and the sequence with maximum likelihood over all ascents is the estimated sequence.

2.2.2. Simultaneous event-based model

To enable traversal of the full space of sequences that contain any combination of simultaneous events, we update the sequence perturbation method: a biomarker is chosen at random and is replaced at any other valid position in the sequence. For example, if the sequence is $(\{2\},\{1,3\},\{4\})$, then a perturbed sequence can be generated by randomly choosing biomarker 3 and replacing the biomarker at position 4 in the sequence, giving the sequence $(\{2\},\{1\},\{4\},\{3\})$. Other possible perturbations are shown in Supplementary Table 1. This perturbation method is compatible with the MCMC sampling method described in (Fonteijn 2012, Young et al 2014), as it retains the property of symmetric transition probability $p(S_{t+1}|S_t) = p(S_t|S_{t+1})$, which simplifies the formulation of the acceptance probability.

2.3. Event state probability density functions

Calculating the sequence likelihood requires the probability density functions of each biomarker under the condition that the event has or has not occurred, $p(x_{m=1,j}|E_{m=1}), \dots, p(x_{m=N,j}|E_{m=N})$ and $p(x_{m=1,j}|\neg E_{m=1}), \dots, p(x_{m=N,j}|\neg E_{m=N})$, respectively.

Hypothetically, if each subjects' position in the sequence is known, then the event state for each biomarker measurement is also known. For example, for a given biomarker i and its event state E_i, the probability density function $p(x_{i,j}|E_i)$ can be fitted to the measurements $\{x_{i,j} \mid k(j) \geq p, s(p) = i\}$ (i.e. the measurements for the subjects at a position greater or equal to the position of the event for biomarker i), where $k(j)$ is the position in the sequence of subject j.

However, as the subjects' sequence position is unknown a priori, then the assumption is made that the measurements are drawn from a mixture distribution $p(x_{i,j}) = w_i p(x_{i,j}|E_i) + (1 - w_i) p(x_{i,j}|\neg E_i)$, whose components are then recovered by fitting a mixture model to all measurements $\{x_{i,j}|j = 1, \dots N\}$.

bioRxiv preprint doi: https://doi.org/10.1101/2022.07.10.499471; this version posted March 18, 2023. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

3. Materials and Methods

3.1. Simulation experiments

3.1.1. Simultaneous event-based forward model

A forward model is used in this study to generate biomarker data for simulation experiments. The model generates data from a given ground truth sequence that can contain simultaneous events. The required inputs to the forward model are (i) the sequence (as described in 2.1.3. Simultaneous event-based model: sequence specification and likelihood function), (ii) the event distributions for each biomarker and (iii) the number of datapoints (i.e., subjects) to sample.

Firstly, a position k, of the subject within the disease progression sequence is sampled from the uniform prior distribution $\mathrm{Unif}\{0, |S|\}$. The biomarker data for this subject, indexed by j, is then generated by sampling from the event distributions corresponding to the position in the sequence: $\sim x_{m,j}|E_m \text{ if } m \in \bigcup_{1\le i\le k} s(i)$, or $\sim x_{m,j}|\neg E_m \text{ if } m \in \bigcup_{k\le i\le |S|} s(i)$. The process is then repeated for the specified number of subjects, returning a matrix X of size N-by-J containing the data samples for J subjects and N biomarkers.

3.1.2 Experiment 1: biomarker variance, simultaneous events and C-EBM uncertainty

To demonstrate that the uncertainty in event positions derived from C-EBM cannot be used to infer the presence of simultaneous events, we quantified the effect of both biomarker variance and simultaneous events on degree of sequence uncertainty. We hypothesised that both biomarker variance and simultaneous events can separately result in a high degree of uncertainty in event positions.

Data was simulated for two biomarkers sampled from either a serial event sequence $(\{1\},\{2\})$, or simultaneous event sequence $(\{1,2\})$, whose probability density functions were gaussian with a mean of zero for the normal event states (Eqs. 8 and 9) and one for abnormal event states (Eqs. 10 and 11). Standard deviation was varied from 0.05 to 2.00 and was equal for each biomarker and event state.

$$p(x_{1,j}|\neg E_1) = \frac{1}{\sigma\sqrt{2\pi}} \exp\left(\frac{(x_{1,j} - 0)^2}{4\pi^2}\right) \tag{8}$$

$$p(x_{2,j}|\neg E_2) = \frac{1}{\sigma\sqrt{2\pi}} \exp\left(\frac{(x_{2,j} - 0)^2}{4\pi^2}\right) \tag{9}$$

$$p(x_{1,j}|E_1) = \frac{1}{\sigma\sqrt{2\pi}} \exp\left(\frac{(x_{1,j} - 1)^2}{4\pi^2}\right) \tag{10}$$

bioRxiv preprint doi: https://doi.org/10.1101/2022.07.10.499471; this version posted March 18, 2023. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

$$p(x_{2,j}|E_2) = \frac{1}{\sigma\sqrt{2\pi}}\exp\left(\frac{(x_{2,j}-1)^2}{4\pi^2}\right) \tag{11}$$

For each sequence and standard deviation combination, one hundred datasets were simulated, each with ten ‘control’ subjects at position zero, where no events have yet occurred, ten ‘end-stage patients’ at the final sequence position $|S|$, where all events have occurred, and twenty ‘intermediate-stage patients’, which are sampled uniformly from the sequence positions i.e., $k \sim \text{Unif}\{0, |S|\}$. To remove the added variability in positional uncertainty due to the estimation of event distributions, these distributions were determined from their simulation definitions.

For each simulated dataset, the uncertainty was quantified in a positional variance matrix, P, whose $i, j'\text{th}$ entry gives the probability that biomarker i is at position j. This probability is defined as the frequency over MCMC samples where biomarker i is at position j (Fonteijn et al 2012) i.e. $P_{i,j} = (\sum_{S\in S_{ij}} 1)/N_{mcmc}$, where N_{mcmc} is the number of MCMC samples and S_{ij} is the set of sequences where biomarker i is at position j. In the case of a serial sequence containing only two biomarkers, this simplifies to $P_{i,j} = P(X|S_{i@j})$, where $S_{i@j}$ refers to the sequence with biomarker i at position j. A binary decision was then made as to whether each positional variance matrix has a significant level of uncertainty or not. A significant level of uncertainty was defined as the highest probability in the matrix being less than 0.95, which corresponds to the absence of certainty (with 0.95 probability of higher) in biomarker positions. The proportion of matrices containing significant levels of uncertainty for the serial or simultaneous sequences was then plotted as a function of biomarker standard deviation.

3.1.3. Experiment 2: Evaluation of simultaneous EBM performance

We evaluated simultaneous EBM performance against a known ground truth sequence by quantifying the percentage of correctly estimated sequences over a set of one hundred simulations of biomarker data. The set of one hundred simulations was repeated for each combination of sequence type (serial, partially simultaneous and fully simultaneous), number of biomarkers (2, 4 and 10), number of subjects (40, 80 and 160) and biomarker variance (s.d’s of 0.1, 0.2 and 0.3).

For each number of subjects, the subject types were split in a 1:2:1 ratio between control, intermediate and end-stage. As in Experiment 1 (section 3.1.2.), the means of the event states used to generate the simulated data were zero and one for normal and abnormal event states, respectively, and the standard deviations were equal for the biomarker event states for each s.d. value.

bioRxiv preprint doi: https://doi.org/10.1101/2022.07.10.499471; this version posted March 18, 2023. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

To sufficiently sample the set of possible sequences during sequence estimation, the number of initialisations and iterations of the greedy ascent was adjusted for each number of biomarkers: 1 and 2 respectively for two biomarkers, 10 and 100 respectively for 4 biomarkers, and 50 and 1000 respectively for 10 biomarkers. For all sequence estimations, the event distributions were fitted using the data from the control and end-stage subjects.

3.1.4. Experiment 3: Comparison to conventional EBM for serial events

To evaluate the ability of S-EBM to correctly identify a sequence containing serial events in the case where C-EBM reports high uncertainty, we quantify the percentage of correctly estimated sequences as a function of biomarker variance for the range of biomarker variance that resulted in a high proportion of positional uncertainty, as determined from section 3.1.2. The simulation conditions are as described in 3.1.2. except with the sequence estimation being performed by either C-EBM or S-EBM on the serial sequence.

3.2. Application to Alzheimer's disease progression

We applied S-EBM to sequence the evolution of biomarker abnormalities in AD while accounting for simultaneous events and compared it to the serial sequence estimated by C-EBM. Our pipeline for data selection follows that of (Young et al 2014) but utilises existing sources of pre-compiled AD data.

3.2.1. AD biomarker source

Biomarkers of cerebrospinal fluid (CSF) (total tau, phosphorylated tau, amyloid-β_{1-42}), cognitive test scores (RAVLT, ADAS-Cog, MMSE) and regional brain volumes (hippocampus, entorhinal cortex, mid-temporal gyrus, fusiform and ventricles) were obtained from the TADPOLE dataset, which is available for download from the Alzheimer's disease Neuroimaging Initiative (ADNI) database (adni.loni.usc.edu) (Mueller et al 2005). TADPOLE is a pre-compiled source of ADNI biomarker data that includes data from phases 1, GO and 2 of ADNI. TADPOLE datasets D1 and D2, which contain biomarker data from every individual that has participated in in at least two separate visits, were used in this study. The image processing steps used by ADNI to generate the biomarkers later compiled in the TADPOLE dataset are described in 3.2.2. ADNI processing.

3.2.2. ADNI processing

CSF measurements of total tau, phosphorylated tau and amyloid-β were obtained via lumbar puncture (Shaw et al 2009). Cognitive test scores were obtained via specialist clinical assessment (Crane et al 2012). Structural magnetic resonance (MR) images were acquired

bioRxiv preprint doi: https://doi.org/10.1101/2022.07.10.499471; this version posted March 18, 2023. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

and underwent pre-processing with standard ADNI pipelines (Jack Jr. et al 2008), which involved correction for gradient non-linearity, B1 non-uniformity correction and peak sharpening. Regional volumes were extracted using Freesurfer cross-sectional and longitudinal pipelines (Reuter et al 2012).

3.2.3. Biomarker processing

Following (Young et al 2014), we included subjects with available biomarker data acquired at baseline up to 5th February 2013 from those subjects scanned at 1.5T. Brain volumes were averaged over hemispheres and normalised by intracranial volume to control for individual differences in head size. CSF total tau and phosphorylated tau were log-transformed to improve event distribution estimation. Cognitively normal subjects who were positive for CSF amyloid-β (<992 pg/ml) or phosphorylated tau (>25 pg/ml) were removed to improve the estimation of event distributions, which are presumed to be predominantly normal in this group.

3.2.4. Event distributions

For each biomarker, probability density functions corresponding to the event having occurred or having not occurred, were fitted to the cognitively normal and AD patients' biomarker data using a constrained gaussian mixture model implemented in MATLAB, as described in (Young et al 2014). The standard deviations of each event component (E and $\neg\mathrm{E}$) are constrained to be less than or equal to that of the cognitively normal or AD group, respectively, and the means are constrained to be no less extreme than the cognitively normal or AD groups. These constraints ensure a robust fit in the case where the distributions of healthy and patient population overlap significantly.

3.2.5. Sequencing allowing simultaneous events

The maximum likelihood S-EBM sequence was estimated from 1,000,000 MCMC samples. MCMC was initialised using the sequence estimated from a greedy ascent performed with 200 initialisations each with 2,000 iterations.

3.2.6. Sequencing of serial events

The maximum likelihood C-EBM sequence was estimated using greedy ascent with 200 random initialisations, each with 2,000 iterations. 1,000,000 MCMC samples were taken to estimate the uncertainty in each biomarkers position.

bioRxiv preprint doi: https://doi.org/10.1101/2022.07.10.499471; this version posted March 18, 2023. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

4. Results and Discussion

4.1. Simulation experiments

4.1.1. Experiment 1: biomarker variance, simultaneous events and C-EBM uncertainty

A serial sequence with high biomarker variance can produce data which is interpreted by C-EBM as having high positional uncertainty (Fig. 2, green line). This uncertainty arises from the relative smoothness of the likelihood function across the sequence space due to overlapping event probability distributions. However, the same degree of uncertainty is also apparent in data produced from sequences containing simultaneous events (Fig. 2, blue line). This many-to-one mapping between sequence features (biomarker variance, simultaneous events) and positional uncertainty suggests that the presence of positional uncertainty in a particular dataset does not imply that the sequence contains simultaneous events. This prevents the use of C-EBM's positional uncertainty for detecting sequences containing simultaneous events.

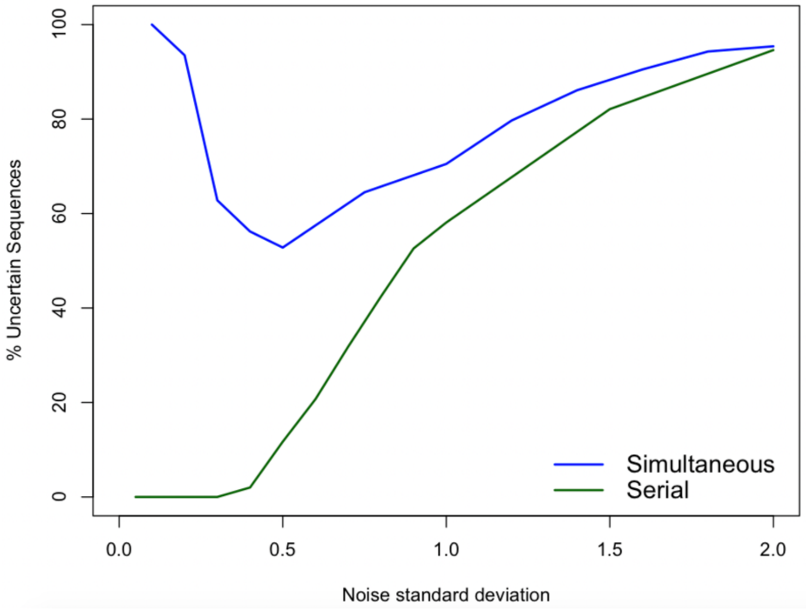


Figure 2. The relation between biomarker standard deviation (x-axis) and uncertainty in the serial sequence estimated by C-EBM (y-axis) for simultaneous events (blue line) and serial events (green line). Both high biomarker variance in serial sequences, and sequences containing simultaneous events, result in a high percentage of uncertain sequences.

4.1.2. Experiment 2: Evaluation of simultaneous EBM performance

S-EBM accurately estimated sequences containing serial events, simultaneous events or both, under a range of experimental conditions (Fig. 3). Sequence estimation accuracy was high for sequences of 10 biomarkers and high biomarker variance when a sufficiently high number of datapoints was sampled. When fewer than 10 datapoints were

bioRxiv preprint doi: https://doi.org/10.1101/2022.07.10.499471; this version posted March 18, 2023. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

sampled per sequence position, accuracy tended to decrease for biomarker standard deviations exceeding 0.1 for both serial and partially simultaneous sequences. Accuracy was high for sequences containing simultaneous events under all conditions.

These results suggest that for moderately sized cohorts of individuals, S-EBM will produce accurate estimates of sequences containing serial events, simultaneous events or both. Given the increasing availability of large prospective and retrospective repositories of cross-sectional or short-term longitudinal biomarker data, this technique has the potential to inform on disease spread patterns for a range of disease. Of particular interest is using retrospective data to provide a refined understanding of disease progression previously estimated using C-EBM.

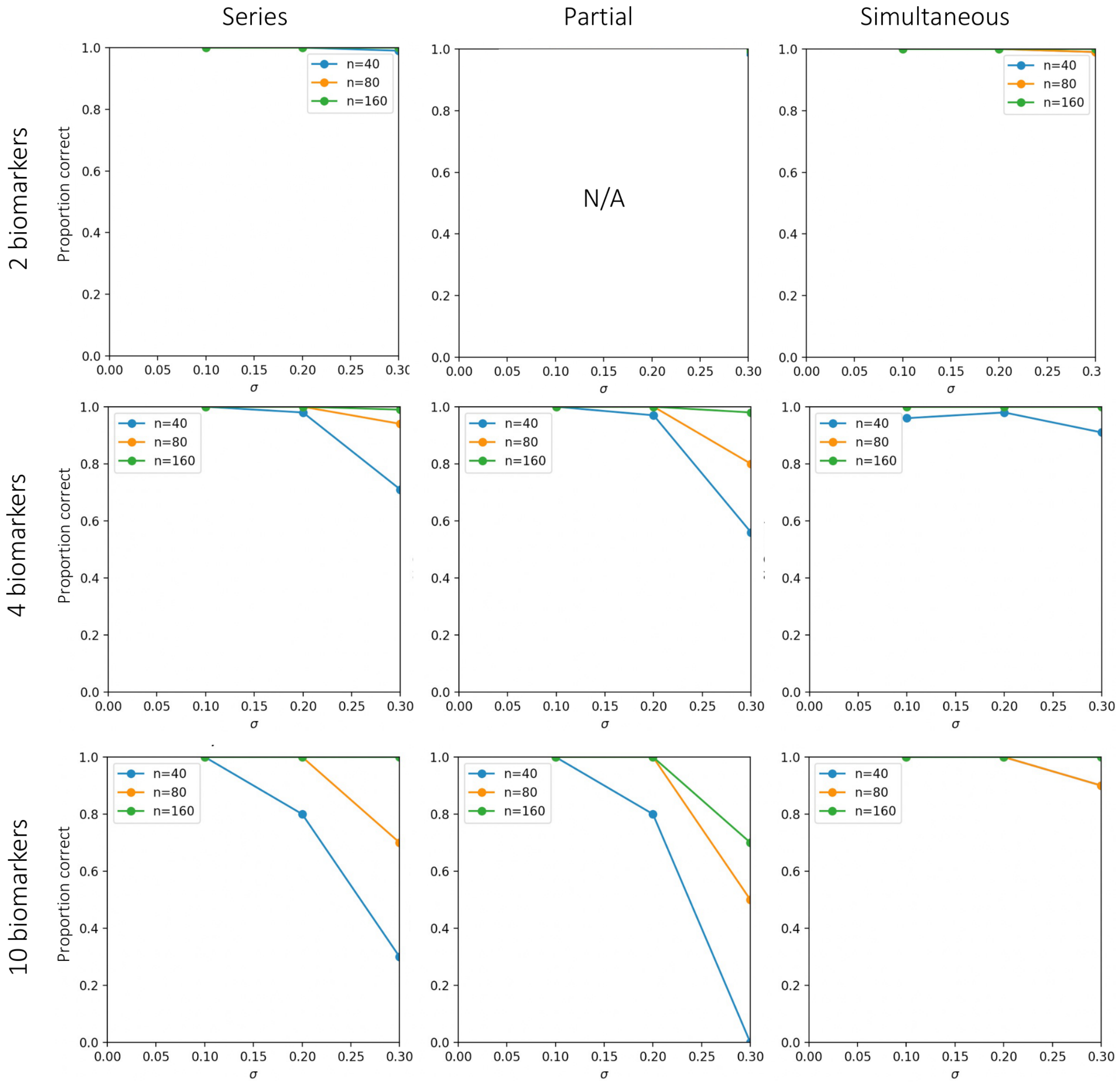


Figure 3. Accuracy of S-EBM sequence estimation for different sequence types (columns), numbers of biomarkers (rows), noise standard deviations (x-axis) and number of subjects

bioRxiv preprint doi: https://doi.org/10.1101/2022.07.10.499471; this version posted March 18, 2023. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

(coloured lines). Accuracy was high for almost all simulations but tended to decrease with fewer subjects, higher noise standard deviation and more biomarkers.

4.1.3. Experiment 3: Comparison to conventional EBM for serial events

C-EBM had higher sequence estimation accuracy than S-EBM for noisy serial sequences which had high C-EBM positional uncertainty (Fig. 4). This suggests that when C-EBM is uncertain on the positional orderings, its maximum likelihood sequence is nevertheless more likely to be correct than the maximum likelihood sequence estimated by S-EBM. This may be expected given that the size of the sequence space of simultaneous events is greater than that for serial sequences, which leaves more scope for false positives. Despite this, without a priori knowledge of the sequence type, S-EBM offers the opportunity to correctly identify a far wider range of types of sequences beyond those restricted by serial order.

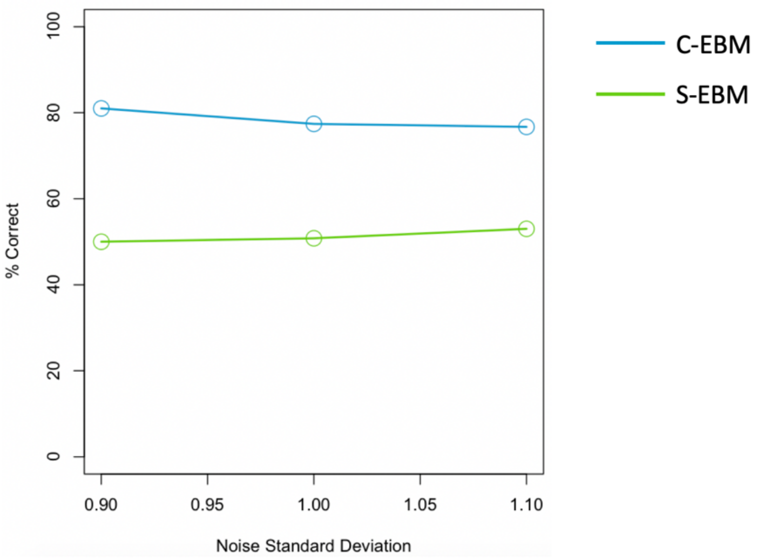


Figure 4. A comparison between C-EBM and S-EBM of serial sequence estimation accuracy in the case where C-EBM reports high positional uncertainty. In this case C-EBM's performance is superior to S-EBM due to the smaller sequence search space.

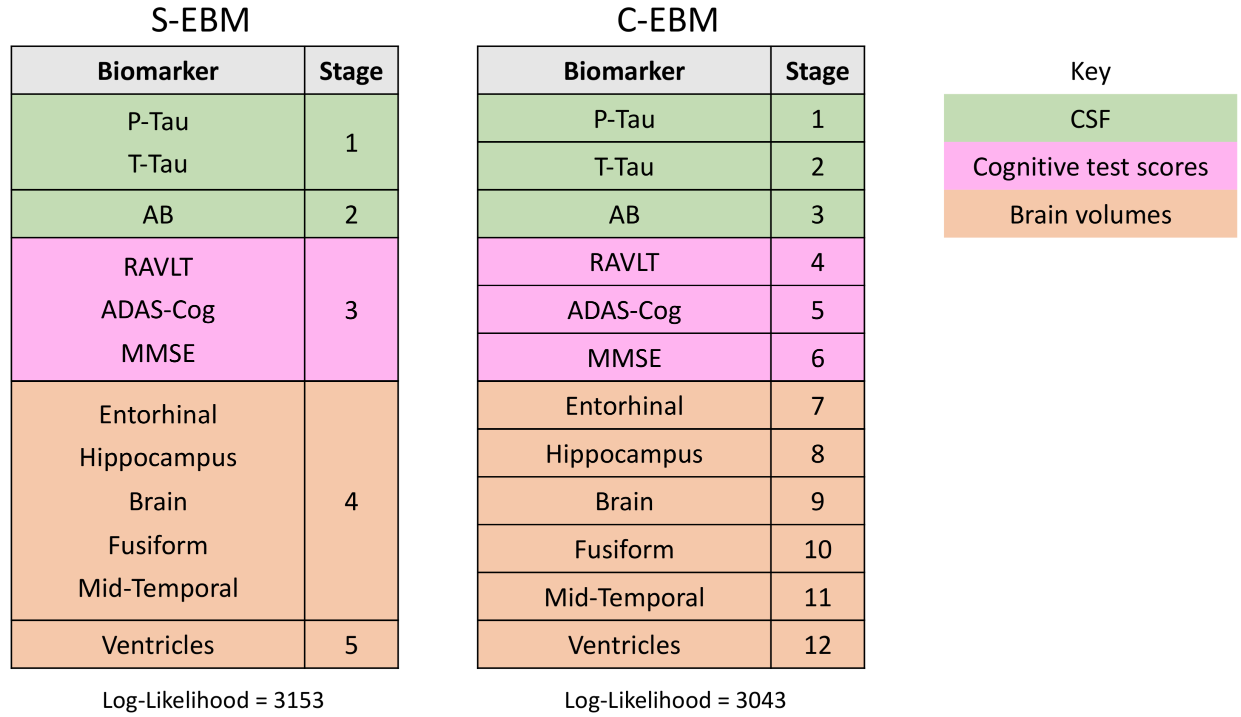
bioRxiv preprint doi: https://doi.org/10.1101/2022.07.10.499471; this version posted March 18, 2023. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

4.2. Application to Alzheimer's disease progression

4.2.1. S-EBM: estimated sequence allowing simultaneous events

The sequence of AD biomarker progression estimated by S-EBM is shown in Fig. 5. S-EBM identified a sequence containing simultaneous events which had a substantially higher log-likelihood compared to the serial sequence estimated by C-EBM.

Simultaneous events were estimated for biomarkers within common biomarker marker domains - CSF, cognitive test scores and brain volumes. Increased CSF total tau and phosphorylated tau were the first events in the sequence, occurring simultaneously, and were followed by high CSF amyloid-β. At disease stage three, low-scoring performance on cognitive test scores RAVLT, ADAS-Cog and MMSE were estimated as simultaneous events. Following cognitive events, the next disease stage consisted of simultaneous volumetric decline in temporal lobe brain regions. The final event in the sequence was increased ventricular volume.



S-EBM

Biomarker	Stage
P-Tau T-Tau	1
AB	2
RAVLT ADAS-Cog MMSE	3
Entorhinal Hippocampus Brain Fusiform Mid-Temporal	4
Ventricles	5

Log-Likelihood = 3153

C-EBM

Biomarker	Stage
P-Tau	1
T-Tau	2
AB	3
RAVLT	4
ADAS-Cog	5
MMSE	6
Entorhinal	7
Hippocampus	8
Brain	9
Fusiform	10
Mid-Temporal	11
Ventricles	12

Log-Likelihood = 3043

Key
CSF
Cognitive test scores
Brain volumes

Figure 5. The sequence of abnormality in biomarkers of CSF, cognitive test scores and brain volumes in AD estimated using S-EBM (left) and C-EBM (right). S-EBM estimates a sequence with substantially higher log-likelihood than C-EBM, by grouping certain biomarkers within domains into the same disease stage. In contrast, S-EBM assumes each biomarker abnormality occurs in series.

bioRxiv preprint doi: https://doi.org/10.1101/2022.07.10.499471; this version posted March 18, 2023. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

4.2.2. C-EBM: estimated serial sequence

The serial sequence estimated by C-EBM (Fig. 6) identified a lower log-likelihood sequence that, by design, assumed all events occur in series. However, it was consistent with the S-EBM sequence in finding a positional separation between groups of biomarker events belonging to different biomarker domains.

The C-EBM positional variance diagram (Fig. 6) however shows a heterogeneous distribution of positional uncertainty for the groups of simultaneous events, highlighting that positional uncertainty cannot be used to infer simultaneous events. Interstingly, interpreting the blocks of positional uncertainty as simultaneous events derives a sequence ({T-Tau}, {P-Tau}, {Abeta}, {RAVLT, ADAS-Cog}, {MMSE}, {Entorhinal, Hippocampus}, {Brain, Fusiform, Mid-Temporal}, {Ventricles}) with lower log-likelihood (log(L)=3108) than that estimated by S-EBM (log(L)=3153) but which nevertheless more closely matches the data than the serial sequence estimated by C-EBM (log(L) = 3043).

Furthermore, the C-EBM positional uncertainty can be low for groups of simultaneous events, such as T-Tau and P-Tau, demonstrating that C-EBM can be confidently incorrect regarding serial event ordering.

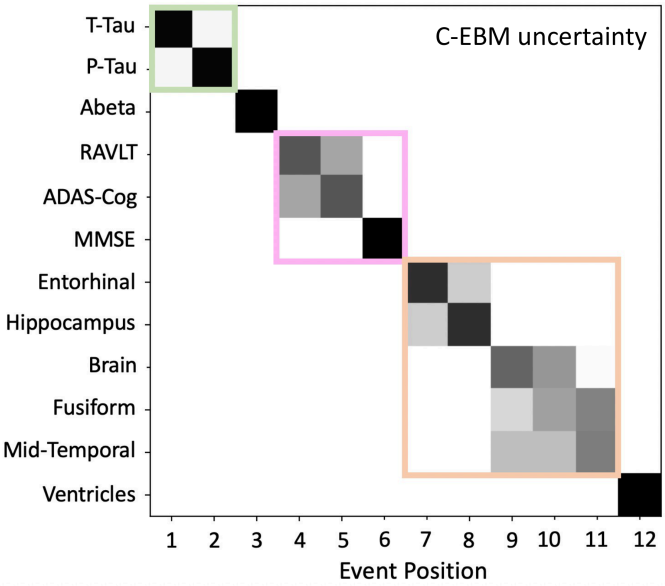


Figure 6. Positional variance diagram showing the positional uncertainty in the serial sequence estimated by C-EBM. Boxes depict the biomarkers grouped into the same stage by S-EBM. The heterogeneity within boxes indicates that C-EBM uncertainty does not infer the same information about simultaneous events as S-EBM. Furthermore, C-EBM can be confidently incorrect regarding serial orderings.

bioRxiv preprint doi: https://doi.org/10.1101/2022.07.10.499471; this version posted March 18, 2023. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

5. Conclusion

This study introduces the simultaneous event-based model. S-EBM is a generalisation of the conventional event-based model for estimating disease progression patterns that contain simultaneous events. With moderate sample sizes, S-EBM produces highly accurate sequence estimates for a range of different sequence types, including serial sequences, thereby broadening the scope of event-based modelling. By removing the requirement that the number of disease progression stages correlates linearly with the number of input biomarkers, the approach suggests a simpler explanation of AD progression, with biomarker abnormality occurring simultaneously within biomarker domains. S-EBM may provide new insights into disease evolution and more accurate subject staging, facilitating the development of therapeutic interventions targeting early disease.

bioRxiv preprint doi: https://doi.org/10.1101/2022.07.10.499471; this version posted March 18, 2023. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

Acknowledgments

CSP, DCA and HZ are supported by the Medical Research Council (MR/T046473/1). CSP is further supported by the EPSRC CMIC Platform Grant (EP/M020533/1). NPO is a UKRI Future Leaders Fellow (MR/S03546X/1) and acknowledges funding from the E-DADS project (EU JPND 2019; UK MRC MR/T046422/1), and the National Institute for Health Research University College London Hospitals Biomedical Research Centre.

6. References

1. Braak, H. and Braak, E. 1991. Neuropathological stageing of Alzheimer-related changes. *Acta neuropathologica*, 82(4)
2. Crane, P.K., Carle, A., Gibbons, L.E., Insel, P., Mackin, R.S., Gross, A., Jones, R.N., Mukherjee, S., Curtis, S.M., Harvey, D. and Weiner, M. 2012. Development and assessment of a composite score for memory in the Alzheimer's Disease Neuroimaging Initiative (ADNI). *Brain imaging and behavior*, *6*(4)
3. Eshaghi, A., Marinescu, R.V., Young, A.L., Firth, N.C., Prados, F., Jorge Cardoso, M., Tur, C., De Angelis, F., Cawley, N., Brownlee, W.J. and De Stefano, N. 2018. Progression of regional grey matter atrophy in multiple sclerosis. *Brain*, 141(6)
4. Fonteijn, H.M., Modat, M., Clarkson, M.J., Barnes, J., Lehmann, M., Hobbs, N.Z., Scahill, R.I., Tabrizi, S.J., Ourselin, S., Fox, N.C. and Alexander, D.C. 2012. An event-based model for disease progression and its application in familial Alzheimer's disease and Huntington's disease. *NeuroImage*, 60(3)
5. Gabel, M.C., Broad, R.J., Young, A.L., Abrahams, S., Bastin, M.E., Menke, R.A., Al-Chalabi, A., Goldstein, L.H., Tsermentseli, S., Alexander, D.C. and Turner, M.R. 2020. Evolution of white matter damage in amyotrophic lateral sclerosis. *Annals of clinical and translational neurology*, 7(5)
6. Jack Jr, C.R., Bernstein, M.A., Fox, N.C., Thompson, P., Alexander, G., Harvey, D., Borowski, B., Britson, P.J., L. Whitwell, J., Ward, C. and Dale, A.M. 2008. The Alzheimer's disease neuroimaging initiative (ADNI): MRI methods. *Journal of Magnetic Resonance Imaging: An Official Journal of the International Society for Magnetic Resonance in Medicine*, *27*(4)
7. Jack Jr, C.R., Knopman, D.S., Jagust, W.J., Shaw, L.M., Aisen, P.S., Weiner, M.W., Petersen, R.C. and Trojanowski, J.Q. 2010. Hypothetical model of dynamic biomarkers of the Alzheimer's pathological cascade. *The Lancet Neurology*, 9(1)), 119-128
8. Marinescu, R.V., Oxtoby, N.P., Young, A.L., Bron, E.E., Toga, A.W., Weiner, M.W., Barkhof, F., Fox, N.C., Klein, S. and Alexander, D.C. 2018. Tadpole challenge: Prediction of longitudinal evolution in Alzheimer's disease. arXiv preprint arXiv:1805.03909.

bioRxiv preprint doi: https://doi.org/10.1101/2022.07.10.499471; this version posted March 18, 2023. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

9. Mueller, S.G., Weiner, M.W., Thal, L.J., Petersen, R.C., Jack, C.R., Jagust, W., Trojanowski, J.Q., Toga, A.W. and Beckett, L. 2005. Ways toward an early diagnosis in Alzheimer’s disease: the Alzheimer’s Disease Neuroimaging Initiative (ADNI). *Alzheimer's & Dementia*, 1(1)
10. Reuter, M., Schmansky, N.J., Rosas, H.D. and Fischl, B. 2012. Within-subject template estimation for unbiased longitudinal image analysis. *Neuroimage*, 61(4)
11. Shaw, L.M., Vanderstichele, H., Knapik-Czajka, M., Clark, C.M., Aisen, P.S., Petersen, R.C., Blennow, K., Soares, H., Simon, A., Lewczuk, P. and Dean, R. 2009. Cerebrospinal fluid biomarker signature in Alzheimer's disease neuroimaging initiative subjects. *Annals of neurology*, *65*(4)
12. Wijeratne, P.A., Young, A.L., Oxtoby, N.P., Marinescu, R.V., Firth, N.C., Johnson, E.B., Mohan, A., Sampaio, C., Scahill, R.I., Tabrizi, S.J. and Alexander, D.C. 2018. An image-based model of brain volume biomarker changes in Huntington's disease. *Annals of clinical and translational neurology*, 5(5)
13. Young, A.L., Oxtoby, N.P., Daga, P., Cash, D.M., Fox, N.C., Ourselin, S., Schott, J.M. and Alexander, D.C., 2014. A data-driven model of biomarker changes in sporadic Alzheimer's disease. *Brain*, 137(9)