Subtype and stage inference with timescales

Alexandra L. Young[1 [0000−0002−7772−781X]], Leon M. Aksman[2 ⋆[0000−0003−2342−0780]], Daniel C. Alexander[3 [0000−0003−2439−350X]], Peter A. Wijeratne[3,4 [0000−0002−4885−6241]], and for the Alzheimer's Disease Neuroimaging Initiative⋆⋆

[1] Department of Neuroimaging, Institute of Psychiatry, Psychology and Neuroscience, King's College London, UK alexandra.young@kcl.ac.uk
[2] Stevens Neuroimaging and Informatics Institute, Keck School of Medicine, University of Southern California, Los Angeles, CA, USA Leon.Aksman@loni.usc.edu
[3] Centre for Medical Image Computing, Department of Computer Science, University College London, London, UK d.alexander@ucl.ac.uk
[4] Department of Informatics, University of Sussex, Brighton, UK p.wijeratne@sussex.ac.uk

Abstract. Neurodegenerative conditions typically have highly heterogeneous trajectories, with variability in both the spatial and temporal progression of neurological changes. Disentangling the variability in spatiotemporal progression patterns offers major benefits for patient stratification and disease understanding but is a complex methodological challenge. Here we present Temporal Subtype and Stage Inference (T-SuStaIn), a technique that uniquely integrates distinct ideas from unsupervised learning: disease progression modelling, clustering, and hidden Markov modelling. We formulate T-SuStaIn mathematically and devise an algorithm for inferring the model parameters and uncertainty. We demonstrate that the combination of disease progression modelling, clustering, and hidden Markov modelling uniquely enables the discovery of subtypes distinguished not just by ordering of abnormality accumulation, but also timescale. We apply T-SuStaIn to longitudinal volumetric imaging data from the Alzheimer's Disease Neuroimaging Initiative, deriving spatiotemporal Alzheimer's disease subtypes together with their timelines of evolution and associated uncertainty. T-SuStaIn has broad utility across a range of longitudinal clustering problems, both in neurodegenerative conditions and more widely in progressive diseases.

Keywords: Longitudinal clustering · Markov jump process · Disease progression model · Subtyping · Prognosis · Dementia

⋆ Joint first author and corresponding author

⋆⋆ 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

1 Introduction

Characterising the natural history of a disease is a crucial step towards understanding the underlying disease biology and predicting disease outcomes. In neurodegenerative diseases, such as Alzheimer's disease (AD), this has proven challenging due to the decades-long disease timescale [1], which makes it impractical to chart the disease from start to end in an individual. This problem is exacerbated by the heterogeneity between individuals, with many individuals following an atypical trajectory [2]. Quantitative models of neurodegenerative disease subtypes and their progression timescales provide insights into the heterogeneous patterns of disease changes and enable patient stratification and prediction of future outcomes in clinical trials and healthcare.

Disease progression modelling (e.g., [3–7]) is a form of unsupervised learning that enables the estimation of long-term trajectories of disease change from cross-sectional and short-term longitudinal datasets [8]. The event-based model (EBM) [3] of disease progression describes a disease as a series of events, where each event corresponds to a new biomarker becoming abnormal, enabling the identification of population-level disease progression patterns from cross-sectional datasets. However, the EBM (a) assumes that all individuals follow a common progression pattern; (b) estimates only the sequence, not the timescale, of disease changes; and (c) does not appropriately exploit longitudinal data. The Subtype and Stage Inference (SuStaIn) algorithm [9] places the EBM in a clustering framework, enabling the identification of disease subtypes with distinct disease progression patterns. The temporal event-based model (T-EBM) [10] places the EBM in a hidden Markov modelling framework, appropriately modelling longitudinal data to enable the estimation of disease timescales from short-term longitudinal data.

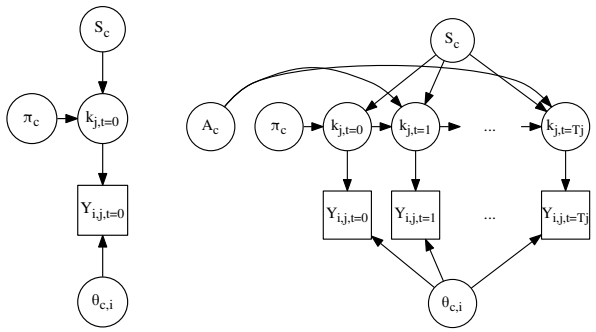
However, SuStaIn cannot estimate disease timescales and the T-EBM cannot account for disease subtypes. Moreover, SuStaIn inherits the EBM's inability to appropriately exploit longitudinal data, which can hinder model identifiability when degenerate solutions exist, e.g., when trajectories from two subtypes cross over (*crossing subtype trajectories*). The majority of previous techniques for longitudinal clustering of disease subtypes (e.g. [11]) fail to consider heterogeneity in disease stage at baseline (i.e. they do not perform disease progression modelling). Those that do account for disease stage heterogeneity have typically required a large number of observations (approximately 1000 individuals with five time-points each) and have high model complexity [12, 13], hindering their utility in medical datasets and in identifying subtypes with low prevelance.

Here we present Temporal Subtype and Stage Inference (T-SuStaIn), a technique that uniquely enables the estimation of disease subtypes with distinct progression patterns and their timescales from short-term longitudinal datasets. We formulate T-SuStaIn mathematically and derive an algorithm for simultaneously inferring disease subtypes, progression patterns, and timescales of progression. Harnessing the added complexity of combining disease progression modelling, clustering, and hidden Markov modelling in a single framework necessitates the development of a novel constrained transition matrix that restricts the dimensionality of the parameter space by encoding common assumptions of disease

progression models. This further enables extension of the inference to use Markov Chain Monte Carlo (MCMC) sampling, providing a joint estimate of the uncertainty in the subtype progression patterns, proportion of individuals belonging to each subtype, and the subtype progression timescales. We show that T-SuStaIn can successfully exploit longitudinal data to recover event sequences that are not identifiable by the cross-sectional SuStaIn algorithm. We demonstrate T-SuStaIn using volumetric structural imaging data from the Alzheimer's Disease Neuroimaging Initiative, identifying AD subtypes with distinct spatiotemporal progression patterns and their timelines.

2 Theory

2.1 Mathematical model for T-SuStaIn



(a) SuStaIn (b) T-SuStaIn

Fig. 1. Graphical models for SuStaIn and T-SuStaIn. Hidden variables are denoted by circles and observations by squares.

The mathematical model underlying T-SuStaIn is formulated as a mixture of temporal disease progression models, combining ideas from SuStaIn [9] and the T-EBM [10]. In this work we use a piecewise linear z-score model of disease progression [9], which we adapt to have a hidden Markov model formulation that leverages longitudinal datasets and estimates timelines. As such, T-SuStaIn makes the same assumptions as both the T-EBM and SuStaIn, namely: *i*) monotonic biomarker dynamics at the group level; *ii*) Markov stage transitions at the individual level; and *iii*) a mixture of event sequences across the population. Graphical models of SuStaIn and T-SuStain are shown in Figure 1. We denote the data for each biomarker i and individual j observed at time t as $Y_{i,j,t}$; the initial probability distribution for cluster c as π_c; the transition probability matrix for cluster c with elements $a_{a,b}$ as A_c; the distribution parameters for biomarker i in cluster c as $\theta_{c,i}$; the overall set of model parameters as $\Theta_c = [\pi_c, A_c, \theta_{c,i}]$;

the event sequence for cluster c as S_c. Here $S_c = \{s_c(1), ..., s_c(N)\}$ is a permutation of N events that represents the hidden sequence of events defining the state space for a discrete time Markov jump process, where an event is the transition of a biomarker from one z-score to another. We denote the hidden stage for individual j at time t as $k_{j,t}$ and define $0 \leq x \leq 1$ as a dimensionless variable that spans the event space. Following [10], under assumptions (i)-(iii), we can write the equation for the total data likelihood of the T-EBM for cluster c as,

$$P(Y_j|\Theta_c, S_c) = \sum_{k_{j,t}=0}^{N} \int_{x=\frac{k}{N+1}}^{x=\frac{k+1}{N+1}} P(k_{j,t=0}|S_c, \pi_c) \prod_{t=1}^{T_j} P(k_{j,t}|k_{j,t-1}, S_c, A_c) \prod_{t=0}^{T_j} \prod_{i=1}^{I} P(Y_{i,j,t}|k_{j,t}, \theta_{c,i}, S_c) dx, \tag{1}$$

where,

$$P(Y_{i,j,t}|k_{j,t}, \theta_{c,i}, S_c) = \text{NormPDF}(Y_{i,j,t}, \theta_{c,i}). \tag{2}$$

Following [9], we assume a univariate normal distribution for the data, $Y_i \sim \mathcal{N}(\mu_i, \sigma_i)$, and choose $\theta_{c,i}(x) = [\mu_{c,i}(x), \sigma_{c,i}]$, where $\mu_{c,i}$ and $\sigma_{c,i}$ are the mean and standard deviations of distribution i in cluster c. In the following we drop the c index for notational simplicity. We define $\mu_i(x)$ as a piece-wise linear function,

$$\mu_i(x) = \begin{cases} \frac{z_1}{x_{Ez_1}} x, \ 0 < x \leq x_{Ez_1} \\ z_1 + \frac{z_2 - z_1}{x_{Ez_2} - x_{Ez_1}} \left(x - x_{Ez_1}\right), \ x_{Ez_1} < x \leq x_{Ez_2} \\ \vdots \\ z_M + \frac{z_{max} - z_{M_i}}{1 - x_{Ez_{M_i}}} \left(x - x_{Ez_{M_i}}\right), \ x_{Ez_{M_i}} < x \leq 1. \end{cases} \tag{3}$$

Here $z_i = z_1, ..., z_{M_i}$ is the set of z-scores for biomarker i such that $N = \sum M_i$, z_{max} is the maximum z-score for biomarker i; and E_{z_i} is the z-score event at $x_{Ez_i} = (k+1)/(N+1)$. Here we define the z-scores with respect to the control population, and hence set the standard deviation $\sigma_i = 1$, i.e., the z-scores correspond to the number of standard deviations from the control population.

Following [14], the elements of the transition matrix A_c are defined as,

$$a_{a,b} \equiv P(k_{j,t} = b|k_{j,t-1} = a, S_c, A_c), \tag{4}$$

the elements of the initial stage probability vector π_a are defined as,

$$\pi_a = P(k_{j,t=0} = a|S_c, \pi_c), \tag{5}$$

and the expected duration of each stage (sojourn time) δ_k as,

$$\delta_k = \sum_{\delta=1}^{\infty} \delta P_k(\delta) = 1/(1 - a_{kk}), \tag{6}$$

where a_{kk} are the diagonal elements of A_c.

The mathematical model underlying T-SuStaIn is defined as a mixture of temporal event-based models (Equation 1),

$$P(Y|\Theta, S) = \prod_{i=1}^{I} \left[\sum_{c=1}^{C} f_c P(Y_j|\Theta_c, S_c) \right], \tag{7}$$

where f_c is the fraction of individuals in subtype c out of a total C clusters.

2.2 Constrained transition matrix

In this work we propose a novel constrained form of the transition matrix A_c that aligns with the assumption of sequential transition through disease stages made by disease progression models. In a traditional hidden Markov model, states may be transitioned between in any order. However, in a disease progression model, the states are instead thought of as stages and have a strict order, with individuals transitioning sequentially through each stage. For example, if an individual started at stage 0 at time $t = 0$ and then transitioned to stage 2 at time $t = 1$, a disease progression model would assume that they transition through stage 1 at some point between $t = 0$ and $t = 1$, whereas in a traditional hidden Markov model there is no such assumption (individuals can move instantaneously from any state to any other state). We encode this idea by assuming that, for $i < j < k$, the probability of transitioning from stage i to stage k is equal to the probability of transitioning from stage i to stage j and then from stage j to stage k, i.e. $a_{ik} = a_{ij}a_{jk}$ for $i < j < k$ (assuming a_{ij} and a_{jk} are independent). This generalises to $a_{ik} = \prod_{j=i+1}^{k} a_{j-1,j}$. Disease progression models also assume monotonic progression, which we enforce by using an upper triangular transition matrix to only allow forward transitions between stages. Following on from these two assumptions, we can derive an analytical solution to the transition matrix A_c that depends only on a transition probability vector $\alpha_c = (a_{00}, \ldots, a_{NN})$ encoding the diagonal of the transition matrix.

To do this we derive analytical solutions for the first off-diagonal elements $(a_{0,1}, a_{1,2}, \ldots, a_{N-2,N-1}, a_{N-1,N})$ that depend only on the elements of the transition probability vector $\alpha_c = (a_{00}, \ldots, a_{NN})$. The first off-diagonal elements can then be used to compute the rest of the elements in the upper triangle of the transition matrix using $a_{ik} = \prod_{j=i+1}^{k} a_{j-1,j}$. From this and the monotonicity assumption we have:

$$A_c = \begin{pmatrix} a_{0,0} & a_{0,1} & \prod_{j=1}^{2} a_{j-1,j} & \ldots & \prod_{j=1}^{N-2} a_{j-1,j} & \prod_{j=1}^{N-1} a_{j-1,j} & \prod_{j=1}^{N} a_{j-1,j} \\ 0 & a_{1,1} & a_{1,2} & \ldots & \prod_{j=2}^{N-2} a_{j-1,j} & \prod_{j=2}^{N-1} a_{j-1,j} & \prod_{j=2}^{N} a_{j-1,j} \\ \vdots & \vdots & \vdots & \ddots & \vdots & \vdots & \vdots \\ 0 & 0 & 0 & \ldots & a_{N-2,N-2} & a_{N-2,N-1} & \prod_{j=N-1}^{N} a_{j-1,j} \\ 0 & 0 & 0 & \ldots & 0 & a_{N-1,N-1} & a_{N-1,N} \\ 0 & 0 & 0 & \ldots & 0 & 0 & a_{N,N} \end{pmatrix}.$$

As each row must sum to 1 to give a valid transition matrix, in the last row we have $a_{N,N} = 1$. In the second to last row we have $a_{N-1,N-1} + a_{N-1,N} = 1$, which can be rearranged to give $a_{N-1,N} = 1 - a_{N-1,N-1}$. In the third to last row we have $a_{N-2,N-2} + a_{N-2,N-1} + a_{N-2,N} = 1$. Substituting in $a_{N-2,N} = a_{N-2,N-1}a_{N-1,N}$ and $a_{N-1,N} = 1 - a_{N-1,N-1}$ gives $a_{N-2,N-2} + a_{N-2,N-1} + a_{N-2,N-1}(1-a_{N-1,N-1}) = 1$, which rearranges to give $a_{N-2,N-1} = \frac{1-a_{N-2,N-2}}{2-a_{N-1,N-1}}$. Following the same logic and substitutions, the fourth to last row rearranges to $a_{N-3,N-2} = \frac{1-a_{N-3,N-3}}{2-a_{N-2,N-2}}$. In general we have $a_{i,i+1} = \frac{1-a_{i,i}}{2-a_{i+1,i+1}}$. So under our assumptions the transition matrix A_c only depends on the diagonal of the transition matrix (the transition probability vector α_c) and we have

$$A_c = \begin{pmatrix} a_{0,0} & \frac{1-a_{0,0}}{2-a_{1,1}} \prod_{j=1}^{2} a_{j-1,j} & \dots & \prod_{j=1}^{N-2} a_{j-1,j} \prod_{j=1}^{N-1} a_{j-1,j} & \prod_{j=1}^{N} a_{j-1,j} \\ 0 & a_{1,1} & \frac{1-a_{1,1}}{2-a_{2,2}} & \dots & \prod_{j=2}^{N-2} a_{j-1,j} \prod_{j=2}^{N-1} a_{j-1,j} & \prod_{j=2}^{N} a_{j-1,j} \\ \vdots & \vdots & \vdots & \ddots & \vdots & \vdots & \vdots \\ 0 & 0 & 0 & \dots & a_{N-2,N-2} & \frac{1-a_{N-2,N-2}}{2-a_{N-1,N-1}} & \prod_{j=N-1}^{N} a_{j-1,j} \\ 0 & 0 & 0 & \dots & 0 & a_{N-1,N-1} & 1 - a_{N-1,N-1} \\ 0 & 0 & 0 & \dots & 0 & 0 & 1 \end{pmatrix}.$$

2.3 Inference

Similarly to [9], we devise a hierarchical framework that sequentially fits an increasing numbers of clusters from a number of randomly chosen initial progression patterns, choosing the optimal number of clusters using cross-validation. We use MCMC sampling to estimate the model parameters and their uncertainty. There are a number of parameters to infer for each cluster $c = 1, ..., C$: the sequence $\overline{S}_c$, fraction of the population in the cluster $\overline{f}_c$, initial probability vector $\overline{\pi}_c$, and the transition probability matrix $\overline{A}_c$ that maximise the total log likelihood, $\mathcal{L}(S_c, f_c, \pi_c, A_c) = \log P(Y; S_c, f_c, \pi_c, A_c)$. We make two assumptions to simplify inference: we use the constrained transition probability matrix described in Section 2.2 (parameterised by the vector α_c) and, following the event-based model [3], we assume a uniform initial probability vector π_c. To speed up convergence of the MCMC, we obtain initial estimates of the set of sequences $\hat{S}_c$ and fractions $\hat{f}_c$ using the SuStaIn algorithm with the modified likelihood function $\mathcal{L}(S_c, f_c, \pi_c, A_c) = \log P(Y; S_c, f_c, \pi_c, A_c)$ and α_c set to $\alpha_c = (0.5, \dots, 0.5)$. We initialise the MCMC sampling procedure using $\hat{S}_c$, $\hat{f}_c$, and $\alpha_c = (0.5, \dots, 0.5)$, sampling the full distribution of the parameters S_c, f_c, and α_c for each cluster.

2.4 Subtyping and staging

T-SuStaIn can output subtypes and stages of individuals using either cross-sectional or longitudinal observations. In either case, an individual is first assigned to a subtype, then a stage given that subtype. In the case where an individual only has a single (i.e., cross-sectional) observation, their subtype is assigned according to their maximium likelihood subtype. In the case where an

individual has multiple (i.e., longitudinal) observations, their subtype is assigned according to their maximium likelihood subtype across all observations. This ensures that individuals stay in the same subtype longitudinally, which is a benefit over cross-sectional SuStaIn.

3 Experiments

3.1 Synthetic data

We first verify that T-SuStaIn can recover subtype progression patterns and timelines in synthetic datasets with a similar size and number of visits to the ADNI dataset. We simulate data directly from the mathematical model underlying T-SuStaIn to enable a direct comparison of parameter estimates and thus perform a sanity check that the algorithm can recover trajectories under idealised conditions. We generate 10 synthetic datasets, setting the number of subtypes to two and randomly generating a progression pattern for each subtype, with 75% of individuals belonging to the first subtype, and 25% belonging to the second. We set the number of subjects to 250, with three visits per subject, giving a total number of data points of 750. We set the number of biomarkers to three, the number of z-scores to three (1, 2, and 3), the maximum z-score to 4, and assume the z-scores evolve from a minimum of 0 with a standard deviation of 1. We set the transition probability for each z-score event to $a = 0.2$ for all biomarkers in each subtype, corresponding to an average transition time of 1.25 years.

3.2 Crossing subtype trajectories

Crossing subtype trajectories have a stage in the middle of each trajectory where the two subtypes look identical (see example in Figure 2). In this case cross-sectional SuStaIn cannot disentangle which beginning and end of each trajectory belong to which subtype. However, T-SuStaIn should be able to disentangle the trajectories by observing the trajectories of individuals before and after the cross-over stage. We run a set of simulations that specifically test the performance of T-SuStaIn for inferring crossing subtype trajectories compared to the SuStaIn algorithm in [9], which only handles cross-sectional data. To ensure any improvements are not simply due to an increase in the number of data points, we use the same number of data points in each case. Specifically, we run SuStaIn on five simulated datasets of 2500 subjects with cross-sectional data only, and T-SuStaIn on five simulated datasets of 500 subjects with 5 time-points. We simulate data from two subtypes across three biomarkers with the progression patterns shown in Figure 2, assuming that the first subtype has a 60% prevalance, and the second a 40% prevalence. We simulate two z-score events per biomarker (z=1 and z=2), a maximum z-score of 3, and assume the z-scores evolve from a minimum of 0 with a standard deviation of 1. The transition probability is set to $a = \frac{1}{3}$ for all biomarkers in each subtype, corresponding to an expected transition time of 1.5 years per stage. Consequently an individual with five time points would be expected to transition three stages on average over the course of their five follow-ups.

3.3 Alzheimer's disease dataset

Data used in the preparation of this article were obtained from the Alzheimer's Disease Neuroimaging Initiative (ADNI) database (adni.loni.usc.edu)[5]. We selected 308 participants from ADNI [15], including individuals who have complete data for two or more consecutive yearly visits, an MCI or AD diagnosis at baseline, and are APOE4 positive (one or more APOE4 alleles). This gave a set of 808 data points in total (average of 2.62 visits per person) with follow-ups spaced at one year intervals. We applied T-SuStaIn to regional MRI volumes from the hippocampus, temporal lobe and other cortical regions (sum of all cortical regions except those in the temporal lobe), correcting for age, sex, years of education, scanner field strength (1.5T vs. 3T) and intracranial volume (ICV) using a control population of 220 APOE4-negative controls. To do this we built a linear regression model for each region, with regional volume as the dependent variable and the above covariates as the independent variables. We then residualized each region (true value minus predicted value from regression) and z-scored the residuals using the controls' means and standard deviations. We used these z-scored residuals as the biomarker inputs to T-SuStaIn. We ran T-SuStaIn using 4 startpoints and 1E5 MCMC iterations, and following [9] identified the optimal number of clusters using the cross-validation information criterion (CVIC).

3.4 Positional variance diagrams

Subtype progression patterns are plotted using positional variance diagrams, which visualise the sequence for each subtype and the uncertainty in that sequence. The colours represent different z-score events, with red corresponding to a z-score of 1, magenta a z-score of 2, and blue a z-score of 3. Each square visualises the probability a particular z-score event occurs at that particular stage, ranging from 0 in white to 1 in red (z=1), magenta (z=2) or blue (z=3).

3.5 Event timelines

The most probable timelines and their uncertainty are obtained from the MCMC samples of the transition matrix for each subtype by using kernel density estimation to fit a non-parametric distribution to the samples and hence obtain descriptive statistics of the mode and full width at half maximum (FWHM).

4 Results

4.1 T-SuStaIn can recover event sequences and timelines in synthetic datasets of similar size to ADNI

We simulated 10 datasets of a similar size to ADNI (250 subjects with three time points), with two subtypes with a prevalence of 75% and 25% and randomly chosen progression patterns. The kendall tau distance between the ground truth and

[5] For further information see: www.adni-info.org

estimated subtype progression patterns was 0.91 (sd=0.20) for Subtype 1 and 0.57 (sd=0.20) for Subtype 2 (kendall tau of 1 indicates maximum similarity, and -1 indicates maximum dissimilarity). T-SuStaIn estimated an average transition probability of 0.25 (sd=0.06) for Subtype 1 and 0.25 (sd=0.04) for Subtype 2.

4.2 T-SuStaIn can infer crossing trajectories in synthetic datasets

Figure 2 illustrates an example dataset in which T-SuStaIn infers crossing subtype trajectories when cross-sectional SuStaIn fails. Across five simulated datasets, T-SuStaIn estimated the correct subtype progression patterns in all five simulations. Cross-sectional SuStaIn estimated the correct subtype progression patterns in only two of five simulations, consistent with our expectation that cross-sectional SuStaIn should estimate the correct progression by chance 50% of the time given that there are two possible stage 4-6 progression patterns that could be randomly appended to either of the stage 1-3 progression patterns. T-SuStaIn can leverage longitudinal data from individuals who move from stages 1-3 to stages 4-6 to link the patterns together correctly.

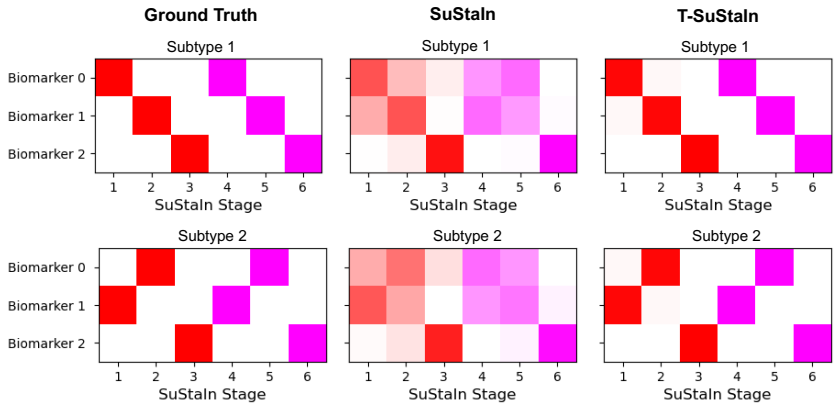


Fig. 2. Example outputs of SuStaIn vs. T-SuStaIn for recovering crossing trajectories using synthetic data. Each progression pattern is visualised using a positional variance diagram (see Section 3.4). In the ground truth the subtypes cross at stage 3, at which point all the biomarkers have reached a z-score of 1. However, the two subtypes have distinct progression patterns before and after this cross-over point. These progression patterns can be inferred by T-SuStaIn, but not SuStaIn.

4.3 T-SuStaIn identifies two subtypes with distinct progression patterns in ADNI

Figure 3 shows the two subtypes inferred by T-SuStaIn in ADNI. As with SuStaIn, each subtype has a distinct progression pattern, but T-SuStaIn further es-

timates an event transition matrix (and therefore a distinct timeline). The first subtype (86% prevalence) has early hippocampal atrophy, followed by temporal lobe atrophy and then widespread cortical atrophy. The second subtype (14% prevalence) has temporal and cortical atrophy at earlier stages. We hypothesise that the first subtype reflects previously described 'typical' AD subtypes and the second 'cortical' AD subtypes [9, 2, 11].

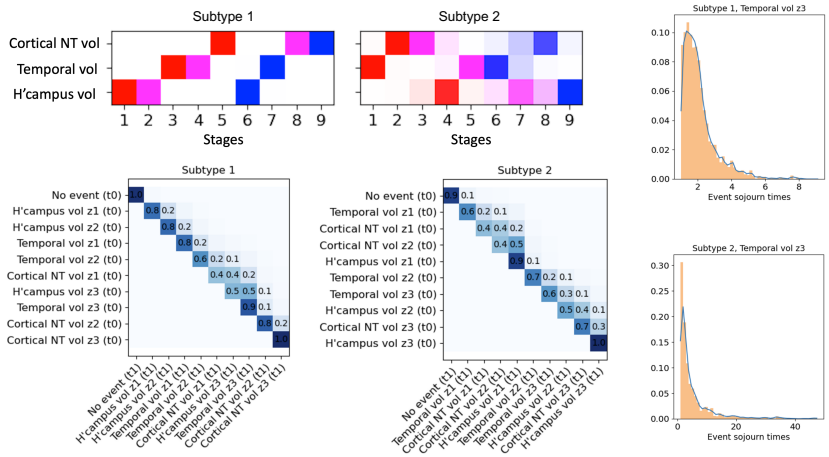


Fig. 3. T-SuStaIn identifies two subtypes with distinct progression patterns and timelines in ADNI. Top row: positional variance diagrams (explained in Section 3.4) of the progression patterns estimated for each subtype. Bottom row: transition matrices for each subtype. Right column: example MCMC samples of event sojourn times in each subtype (blue line indicates the kernel density estimates described in Section 3.5). Cortical NT: all regions in cortex excluding the temporal lobe.

4.4 Each ADNI subtype has a distinct timeline

Figure 4 shows the event timelines inferred by T-SuStaIn in ADNI. T-SuStaIn infers that the overall timeline of Subtype 1 is longer than Subtype 2, consistent with previous studies indicating faster progression of cortical AD subtypes [16].

5 Discussion

We introduced T-SuStaIn, a longitudinal discrete clustering technique that disentangles spatial and temporal heterogeneity in progressive diseases. The strengths of T-SuStaIn lie in its ability to infer interpretable temporal subtypes from reasonably sized datasets of order 100 individuals, and to provide improved identi-

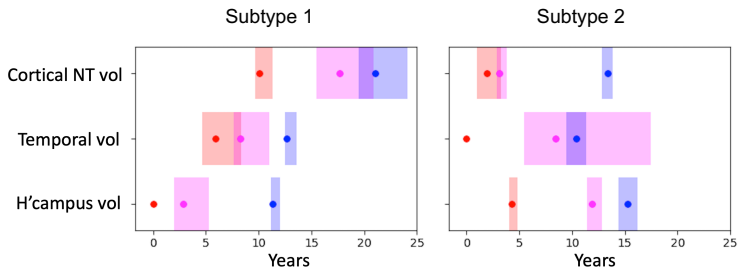


Fig. 4. Visualisation of ADNI subtype timelines. The colours represent different z-score events (red: z=1, magenta: z=2, blue: z=3), with the dots representing when the event occurs (mode described in Section 3.5), and the boxes representing the uncertainty in the timing of the event (FWHM described in Section 3.5).

fiability over cross-sectional SuStaIn, for example in the case of crossing trajectories. Whilst our results support T-SuStaIn's broad potential clinical utility, we acknowledge that its current formulation limits its use to data with fixed-time intervals; future work will allow for variable-time intervals [17]. Although we focused on complete structural MRI data in this work, T-SuStaIn can readily use any type of dynamic biomarker data and accounting for missing data is straightforward, e.g., [10]. As with SuStaIn, various disease progression models can be used with T-SuStaIn to model alternative data types [18, 3, 9, 19]. As such, T-SuStaIn will be able to infer longitudinal subtypes from short-term multi-modal datasets with irregular sampling and missing data, further extending its use in improving disease understanding and patient stratification.

Acknowledgements

ALY is supported by an MRC Skills Development Fellowship (MR/T027800/1). Data collection and sharing for this project was funded by the Alzheimer's Disease Neuroimaging Initiative (ADNI) (National Institutes of Health Grant U01 AG024904) and DOD ADNI (Department of Defense award number W81XWH-12-2-0012)[6].

References

1. V. L. Villemagne, S. Burnham, P. Bourgeat, et al. Amyloid deposition, neurodegeneration, and cognitive decline in sporadic alzheimer's disease: a prospective cohort study. *Lancet Neurol.*, 12:357–67, 2013.
2. J. W. Vogel, A. L. Young, N. P. Oxtoby, et al. Four distinct trajectories of tau deposition identified in alzheimer's disease. *Nat Med.*, 27:871–881, 2021.

[6] For a full list of ADNI funders see: https://adni.loni.usc.edu/wp-content/uploads/how_to_apply/ADNI_Data_Use_Agreement.pdf

3. H. M. Fonteijn, M. Modat, M. J. Clarkson, et al. An event-based model for disease progression and its application in familial alzheimer's disease and huntington's disease. *NeuroImage*, 60:1880–1889, 2012.
4. B. M. Jedynak, A. Lang, B. Liu, et al. A computational neurodegenerative disease progression score: method and results with the alzheimer's disease neuroimaging initiative cohort. *Neuroimage*, 15:1478–86, 2012.
5. M. C Donohue, H. Jacqmin-Gadda, M. Le Goff, et al. Estimating long-term multivariate progression from short-term data. *Alzheimer's & Dementia*, 10:S400–410, 2014.
6. J. B. Schiratti, S. Allassonnière, O. Colliot, et al. A bayesian mixed-effects model to learn trajectories of changes from repeated manifold-valued observations. *J Machine Learning Research*, 18:1–33, 2017.
7. M. Lorenzi, M. Filippone, G. B. Frisoni, et al. Probabilistic disease progression modeling to characterize diagnostic uncertainty: Application to staging and prediction in alzheimer's disease. *NeuroImage*, 190:56–68, 2019.
8. N. P. Oxtoby and D. C. Alexander. Imaging plus x: multimodal models of neurodegenerative disease. *Curr Opin Neurol*, 30(4):371–379, 2019.
9. A. L. Young, R. V. Marinescu, N. P. Oxtoby, et al. Uncovering the heterogeneity and temporal complexity of neurodegenerative diseases with subtype and stage inference. *Nature Communications*, 9, 2018.
10. P. A. Wijeratne and D. C. Alexander. Learning transition times in event sequences: The temporal event-based model of disease progression. *IPMI LNCS*, 12729:583–595, 2021.
11. K. Poulakis, J.B. Pereira, J.S. Muehlboeck, et al. Multi-cohort and longitudinal bayesian clustering study of stage and subtype in alzheimer's disease. *Nature Communications*, 13, 2022.
12. P.E. Poulet and S. Durrleman. Mixture modeling for identifyingsubtypes in disease course mapping. *IPMI LNCS*, 12729:571–582, 2021.
13. I.Y. Chen, R.G. Krishnan, and D. Sontag. Clustering interval-censored time-series for disease phenotyping. *arXiv*, 2021.
14. L.R. Rabiner. A tutorial on hidden markov models and selected applications in speech recognition. *IEEE*, 77:257–286, 1989.
15. S. G. Mueller, M. W. Weiner, L. J. Thal, et al. The alzheimer's disease neuroimaging initiative. *Neuroimaging Clin N Am*, 15:869–877, 2005.
16. D. Ferreira, A. Nordberg, and E. Westman. Biological subtypes of alzheimer disease: A systematic review and meta-analysis. *Neurology*, 94, 2020.
17. P. Metzner, I. Horenko, and C. Schütte. Generator estimation of markov jump processes based on incomplete observations non-equidistant in time. *Phys Rev E Stat Nonlin Soft Matter Phys.*, 76, 2007.
18. L. M. Aksman, P. A. Wijeratne, N. P. Oxtoby, et al. pysustain: A python implementation of the subtype and stage inference algorithm. *SoftwareX*, 16:100811, 2021.
19. A. L. Young, J. W. Vogel, L. M. Aksman, et al. Ordinal sustain: Subtype and stage inference for clinical scores, visual ratings, and other ordinal data. *Front Artif Intell.*, 4:613261, 2021.