s-SuStaIn: Scaling subtype and stage inference via simultaneous clustering of subjects and biomarkers

Raghav Tandon RAGHAV.TANDON@GATECH.EDU
Georgia Institute of Technology

James J. Lah JLAH@EMORY.EDU
Emory University

Cassie S. Mitchell CASSIE.MITCHELL@BME.GATECH.EDU
Georgia Insititue of Technology

Abstract

Event-based models (EBM) provide an important platform for modeling disease progression. This work successfully extends previous EBM approaches to work with larger sets of biomarkers while simultaneously modeling heterogeneity in disease progression trajectories. We develop and validate the s-SuStain method for scalable event-based modeling of disease progression subtypes using large numbers of features. s-SuStaIn is typically an order of magnitude faster than its predecessor (SuStaIn). Moreover, we perform a case study with s-SuStaIn using open access cross-sectional Alzheimer's Disease Neuroimaging (ADNI) data to stage AD patients into four subtypes based on dynamic disease progression. s-SuStaIn shows that the inferred subtypes and stages predict progression to AD among MCI subjects. The subtypes show difference in AD incidence-rates and reveal clinically meaningful progression trajectories when mapped to a brain atlas.

Data and Code Availability We use Alzheimer's Disease Neuroimaging Initiative (ADNI) data made available as a part of the TADPOLE challenge (Marinescu et al., 2019b). It was downloaded via the Laboratory Of Neuroimaging data archive at https://adni.loni.usc.edu/. The code is available at https://github.com/pathology-dynamics/s-SuStaIn.

Institutional Review Board (IRB) This work does not require IRB approval.

1. Introduction

Biomarkers and longitudinal clinical outcomes for multifactorial neurodegenerative disorders, such as Alzheimer's Disease (AD), Parkinson's Disease, Amyotrophic Lateral Sclerosis, and others are known to show variance in the diseased population. Two important sources of variance are 1) disease progression and its associated dynamics, and 2) phenotypic heterogeneity arising out of disease subtypes which may have its roots in genetic and/or environmental factors. While disease progression along a trajectory leads to temporal heterogeneity, the subtypes characterize diverse progression trajectories for the same disease. The goal of this work was to develop a method to model diverse disease progression trajectories and dynamics (e.g. disease subtypes) using large numbers of features typical of present day genomics, proteomics imaging, and other multimodal clinical datasets.

Previous event-based models (EBM) developed by Fonteijn et al. (2012); Young et al. (2014) learn a single overall or "average" disease progression trajectory from cross-sectional data. EBM hypothesizes disease progression to occur as a sequence of biomarker abnormalities and infers the characteristic sequence from cross-sectional data by using a probabilistic generative model. It is further extended to model disease subtypes via multiple progression trajectories in Young et al. (2018). However, since the disease progression trajectory is defined as a permutation over the measured biomarkers, the methods are not amenable to scaling with larger biomarker sets. The presented work scales previous EBM approaches to infer disease trajectory and its sub-

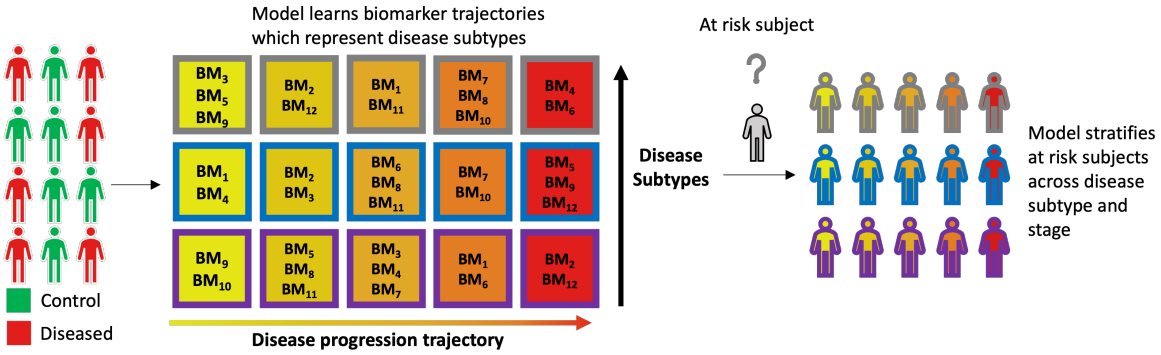


Figure 1: Overview for s-SuStaIn

s-SuStaIn uses cross-sectional data from healthy controls and diseased populations to learn a set of disease progression trajectories. The progression trajectories are defined by a sequence over biomarker clusters. In the example shown here, 14 biomarkers are assigned to 5 clusters across 3 disease trajectories, or subtypes. The disease progresses with all biomarkers in a cluster turning abnormal. These trajectories can be later applied to biomarker measurements from at-risk subjects in order to subtype and stage them for disease risk.

types in the presence of ever-increasing numbers of biomarkers. This is achieved via a shift in the EBM framework. Specifically, similar to the previously published scaled event-based model (sEBM), we hypothesize here that disease progression occurs over biomarker clusters, rather than individual biomarkers Tandon et al. (2023a). This leads to a reformulation of the model likelihood function that is easier to optimize with increasing number of biomarkers. This new method, called s-SuStaIn (scaling Subtype and Stage Inference), combines the scaling enabled by biomarker clusters with the ability to optimally subtype disease progression trajectories. A schematic of the approach and its application in stratifying risk is shown in Figure 1. Results show that s-SuStaIn is typically an order of magnitude faster than its predecessor while having a similar performance in inferring disease progression trajectories.

On real subject data from ADNI, s-SuStaIn infers 4 subtypes depicting differences in progression patterns and 6 stages (0-5) depicting disease severity, from 119 primarily neuroimaging biomarkers (greater than any previous study using EBM). The subtypes capture a difference in AD incidence-rates among MCI (mild cognitive impairment) subjects, while the subtype specific disease stages capture risk of progression to AD while adjusting for genetic risks (APOE4 status), age, gender and education.

2. Background

2.1. Event-Based Model (EBM)

EBM is a probabilistic generative model of disease progression, hypothesizing progression to be a sequence of irreversible biomarker abnormality events. This way, the disease progression trajectory is defined by a permutation of the measured biomarker set. It was first introduced by Fonteijn et al. (2012) for familial cases in Alzheimer's Disease (AD) and Huntington Disease (HD) and later extended to sporadic AD cases by Young et al. (2014). An important advantage of the event-based model is that it can use cross-sectional data for modeling disease progression.

The model takes as input N scalar biomarker measurements from J subjects. A patient j has their biomarker measurements represented as $X_j = \{x_{1,j}, x_{2,j} \dots x_{N,j}\}$. The full data can be represented as $X \in \mathbb{R}^{J \times N}$. EBM hypothesizes a characteristic sequence of biomarkers which describes disease progression. This sequence can be generally written as $S = (s(1), s(2) \dots, s(i), \dots s(N))$, where s(i) represents the biomarker taking the i^{th} position in the sequence. According to EBM, the subject is at stage k if biomarkers $s(1) \dots s(k)$ have turned abnormal while biomarkers $s(k+1) \dots s(N)$ remain normal. The model makes a key assumption - that the likelihood of measurements across biomarkers are independent, conditional on their respective event occurrence. A probabilistic expression for the data likelihood of subject j is given by Equation (1):

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

The subject stage k, i.e. number of biomarkers that have turned abnormal is a latent variable. It does not depend on the sequence S, and can be marginalized out to write the data likelihood. A priori, k is assumed to be uniformly distributed over the possible stages thereby not depending on the diagnosed clinical stage of the subject. Assuming independence of measurements from patients (X_j), the likelihood for the full data $X \in \mathbb{R}^{J\times N}$ can be written as

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

$E_{s(i)}$ denotes that the biomarker taking the i^{th} position in the sequence has turned abnormal, while $\neg E_{s(i)}$ denotes that it remains normal. Young et al. (2014) describes how $p(x_{s(i)j}|E_{s(i)})$ and $p(x_{s(i)j}|\neg E_{s(i)})$ are computed by fitting a two component Gaussian mixture model to each biomarker.

Key Modeling assumptions An important assumption in Equations (1) and (2) is the homogeneity of disease progression across all subjects. The subjects are assumed to progress along the same trajectory defined by the event sequence S. This ignores phenotypic heterogeneity seen in the disease due to different disease subtypes. An important contribution of Young et al. (2018) is to extend the EBM framework to model disease subtypes and relax the assumption of homogeneous disease progression. Other assumptions in Equations (1) and (2) include monotonic changes to biomarkers, and the biomarker measurements being conditionally independent on the event occurrence ($E_{s(i)}$ or $\neg E_{s(i)}$).

2.2. Subtype and Stage Inference (SuStaIn)

SuStaIn algorithm presented by Young et al. (2018), extends the EBM framework in two important ways.

1. It relaxes the assumption that all subjects follow the same disease progression trajectory. It does so by modeling the data as a mixture of multiple disease progression trajectories or subtypes.
2. The biomarkers are allowed to continuously accumulate with the disease progression.

In this work, we focus on the first contribution and extend it further to work with larger sets of biomarkers. The second contribution increases data dimensions beyond the number of biomarkers and will be considered in more detail in future work. However, the method presented in Section 4 for scaling to larger biomarker sets remains applicable to both.

While the event-based model (EBM) estimates only a single biomarker event sequence S for all subjects, SuStaIn estimates a mixture of T such event sequences, each representing a disease subtype $(S_1, S_2 \dots S_T)$. The overall data likelihood is expressed as a mixture of these subtypes

$$p(X|M) = \sum_{t=1}^{T} f_t \times p(X|S_t) \quad (3)$$

Here, $p(X|M)$ denotes the data likelihood for the overall model. $p(X|S_t)$ denotes the data likelihood for the event sequence S_t. f_t denotes the fraction of the subtype estimated from the data ($f_t \in [0,1]$, $\sum_{t=1}^{T} f_t = 1$). The SuStaIn algorithm in Young et al. (2018) proceeds by iteratively maximizing data-likelihood in (3) by using the expectation-maximization (E-M) algorithm to estimate $S_1, S_2 \dots S_T$ and $f_1, f_2 \dots f_T$.

3. Problem setting

The event-based model (EBM) introduced by Fonteijn et al. (2012); Young et al. (2014) can be used to study disease progression as a single sequence of biomarker abnormalities. SuStaIn (Young et al., 2018) can be used to extend EBM to infer disease subtypes to model phenotypic heterogeneity seen across subjects with the same underlying conditions. However, scaling them to work with larger number of biomarkers (data dimensions) is challenging. Three reasons for the following are given below.

1. The state space for the previous models - number of possible event orderings increases as $(N!)^T$, where N is the number of biomarkers and T is the number of disease subtypes being fitted to the model ($T = 1$ for EBM). Hence larger N and T values slows down optimization of overall data likelihood in Equation (3) by factorial and exponential increases in the search space of the optimization algorithm in Young et al. (2018).

2. Evaluating data likelihood at each point in the model's state space increases in complexity. Equation (2) shows the dependence of the data likelihood on the number of biomarkers N.

3. EBM and SuStaIn estimate uncertainty in biomarker positions along the event sequence by using posterior sampling techniques such as Markov Chain Monte Carlo (MCMC). With increasing number of biomarkers (N), MCMC sampling over biomarker sequences becomes harder due to the support of the underlying distribution increasing as $(N!)^T$. This leads to the need for a greater number of samples to cover all regions of the posterior distribution and increased over-all run times.

Scaling these models also has a clinical utility since it can enable the identification of subtle progression patterns over diverse biomarkers. This can potentially lead to uncovering of new disease signatures with diagnostic and prognostic utility. This work proposes a solution to scale SuStaIn to a larger number of biomarkers, thereby addressing all of the above challenges. It speeds up overall model optimization by reducing the state space and making likelihood computations faster. It also leads to advantages in MCMC sampling to characterize uncertainty in inferred biomarker event sequences.

4. Method

Scaling SuStaIn to higher number of biomarkers poses challenges outlined in Section 3. These challenges arise from the model viewing disease progression as a sequence of biomarkers turning abnormal, one at a time. In order to scale SuStaIn to work with increasing number of biomarkers, this work re-formulates the likelihood function and views disease progression to arise from a cluster of biomarkers turning abnormal at a time. These ideas have been previously introduced in Tandon et al. (2023a) where biomarkers are clustered along the event sequence, which allows for sets of biomarkers to turn abnormal simultaneously. However, the work by Tandon et al. (2023a) builds upon the assumption that all subjects follow the same disease trajectory which limits their clinical utility. By combining their contributions with those from Young et al. (2018), this work presents a new model which is scalable to a higher number of biomarkers, while also identifying distinct disease progression trajectories.

4.1. Clustering biomarkers along event sequence

Biomarkers characterizing the disease progression trajectory can be clustered together by relaxing the assumption that the disease advances by abnormality of a single biomarker at a time. Instead, multiple biomarkers can turn abnormal simultaneously to advance the disease stage. In this context, the set of biomarkers turning abnormal simultaneously is considered as a cluster. Biomarkers belonging to the same cluster, occupy same position in the event sequence, while clusters are ordered to characterize progressive disease stages. This work introduces some key additions unique to the previous approach in Tandon et al. (2023a). Specifically, sEBM (Tandon et al., 2023a) assumed clusters to be of fixed size, which requires user-specific choices. Instead, this work extends Tandon et al. (2023a) to include flexible cluster sizes similar to Parker et al. (2022). Another major limitation of sEBM is the inherent assumption that all subjects follow the same disease progression trajectory S. Current work relaxes this assumption by using ideas introduced in Young et al. (2018), and allows for modeling of disease subtypes which follow distinct trajectories. This ultimately results in a novel disease-progression model which can uncover disease subtypes from higher number of biomarkers. Below we introduce a mathematical framework for the above ideas.

Let the disease subtypes be defined by a sequence of biomarker clusters, i.e. $S^t = (c_1^t, c_2^t \dots c_n^t)$. c_i^t denotes the set of biomarkers which turn abnormal after the previous $i-1$ clusters in subtype t. The number of disease stages is $|S^t|$ and is set to n for all subtypes. The number of biomarkers in the i^{th} cluster of subtype t is denoted by $|c_i^t|$. Under this new formulation, Equations (1) and (2) can be re-written as

$$p(X_j|k, S^t) = \prod_{b \in \cup_{i=1}^{k} c_i^t} p(x_{b,j}|E_b) \times \prod_{b \in \cup_{i=k+1}^{n} c_i^t} p(x_{b,j}|\neg E_b) \tag{4}$$

$$p(X|S^t) = \prod_{j=1}^{J} \sum_{k=0}^{n} p(k) \Big(\prod_{b \in \cup_{i=1}^{k} c_i^t} p(x_{b,j}|E_b) \times \prod_{b \in \cup_{i=k+1}^{n} c_i^t} p(x_{b,j}|\neg E_b) \Big) \tag{5}$$

subject to the conditions - $|c_i^t| \geq C_{min}, C_{min} \in \mathbb{Z}^+, \sum_{i=1}^{n} |c_i^t| = N, c_i^t \cap c_j^t = \emptyset \quad (i \neq j), |S^t| = n, \forall \quad 1 \leq i \leq n, 1 \leq t \leq T$. $x_{b,j}$ represents measurement for biomarker b in subject j. Subsequently, the full model across all subtypes can be written similarly to Equation (3)

$$p(X|M) = \sum_{t=1}^{T} f_t \times p(X|S^t) \tag{6}$$

4.2. Assigning subjects to disease subtypes

The model maximizes the overall data likelihood in Equation (6) and results in T trajectories $(S_1, S_2 \dots S_T)$ followed by biomarkers, and their corresponding fractions in the data $(f_1, f_2 \dots f_T)$. With these estimates from the overall data, individual subject j is assigned a particular disease subtype t_j^* by maximizing the subtype specific data likelihood. The prior on k is assumed to be uniform, i.e. a priori a subject is equally likely to be in any stage.

$$t_j^* = \arg\max_t f_t \times \sum_{k=0}^{n} p(k) p(X_j|k, S^t) \tag{7}$$

4.3. Assigning subtype specific disease stage to subjects

Once subtypes have been assigned to subjects, they can be staged for the degree of disease progression within the subtype. This is done by computing the posterior for the disease stage, given the subject specific data and subtype. As in Section 4.3, the prior on k is uniform.

$$k_j^* = \arg\max_k p(k|X_j, S^{t_j^*}) \tag{8}$$

4.4. Theory : Reduction in associated permutational complexity

The original SuStaIn model has $N!^T$ possible configurations, since it uniquely orders each biomarker, and there are T subtypes. However, s-SuStaIn introduced in this work has that permutational complexity in the worst case limit. Number of possible configurations for subtype t in s-SuStaIn can be written as

$$\underbrace{\binom{N}{|c_1^t|}\binom{N-|c_1^t|}{|c_2^t|}\cdots\binom{N-\sum_{i=1}^{n-1}|c_i^t|}{|c_n^t|}}_{n \text{ clusters}} = \frac{N!}{\prod_{i=1}^{n}|c_i^t|}$$

Overall number of configurations across all subtypes is

$$\prod_{t=1}^{T} \frac{N!}{\prod_{i=1}^{n}|c_i^t|} = \frac{N!^T}{\prod_{t=1}^{T}\prod_{i=1}^{n}|c_i^t|}$$

Since $|c_i^t| \in \mathbb{Z}^+$

$$\frac{N!^T}{\prod_{t=1}^{T}\prod_{i=1}^{n}|c_i^t|} \leq N!^T$$

The equality holds in the worst case scenario only when $n = N$ and $|c_i^t| = 1, \forall i, t$.

4.5. Model fitting and inference

The model described in Equations (4), (5) and (6) is optimized by using the expectation-maximization algorithm introduced in Young et al. (2018) and made available in Aksman et al. (2021). Besides the usual parameters in the SuStaIn algorithm, two new hyperparameters are introduced. These are the number of biomarker clusters (n) and the minimum size of each cluster (C_{min}). From the fitted model, subject specific subtypes and stages are inferred using Equations (7) and (8).

5. Experiments

All experiments are performed on an Intel Xeon Gold 6136 CPU. The CPU is a multi-core processor with a clock speed of 3.00 GHz and features 48 cores. The code was adapted from Aksman et al. (2021) to reflect the changes introduced by Equations (4), (5) and (6). Where applicable, s-SuStaIn was compared to the SuStaIn algorithm. Comparison of s-SuStaIn to other non-EBM disease progression models (Lee and Van Der Schaar, 2020; Qin et al., 2023; Noroozizadeh et al., 2023) is not straightforward due to differences in data economics and model architectures which prevent a direct comparison. While SuStaIn and s-SuStaIn use cross-sectional data, models such as AC-TPC require temporal data (see Table 3).

5.1. Simulation study

The aim for the simulation study was to compare s-SuStaIn and SuStaIn on 3 different metrics, i.e. optimization times, inferring ground truth sequences associated with disease subtypes, and their respective fractions in the synthetic data (Figure 2). The data was simulated according to Young et al. (2015). The

factors of variation were number of biomarkers $N \in [50, 100, 150, 200]$, and number of disease subtypes $T \in [2, 3, 4]$. The fraction of individual subtypes was set to (0.6, 0.4) for 2 subtypes, (0.6,0.3,0.1) for 3 subtypes and (0.4,0.3,0.2,0.1) for 4 subtypes. Additionally, the s-SuStaIn model had $n = 5$ biomarker clusters, with minimum cluster size $C_{min} = 0.1 \times N$. Each experimental setting was repeated with 5 random seeds. In all cases, the number of samples was set to 200.

The distance between the ground truth sequence and inferred sequence was computed by calculating the partial Kendall-τ distance introduced in Fagin et al. (2006). For the SuStaIn results, the inferred sequences were converted to partial-rankings using cluster sizes ($|c_i^t|$) from the s-SuStaIn results. This way, the Kendall-τ is computed for each subtype and the overall metric for all subtypes is computed by taking a weighted mean, with subtype fractions serving as the weights. The inferred fractions of the subtypes are compared to their true fractions in the simulated data using cross-entropy.

5.2. Real data from ADNI study

Data used in this section has been obtained as described in the Data and Code Availability statement.

Model fitting We fit the s-SuStaIn model using cross-sectional data from 170 cognitively normal controls and 157 AD subjects. Measurements from 119 neuroimaging biomarkers are used for each subject. The maximum number of subtypes T is set to 4, and the number of biomarker clusters n is set to 5 for each subtype. C_{min} is set to 10. The fitted model is validated on 551 MCI (mild cognitive impairment) subjects which are kept as a separate held-out set.

5.2.1. MCI progression to AD as a function of disease stage

Subtype and stage for MCI subjects are inferred from s-SuStaIn trained on Control and AD subject data. Survival analysis is performed for each subtype of MCI subjects, using Cox proportional-hazards model (conversion from MCI to AD or other dementia is considered as the event of interest). The model is adjusted for covariates such as age, gender, education, and number of APOE4 allele copies which is a known genetic risk factor. Figure 3 shows the conversion risk as a function of disease stage. Table 1 compares the effect sizes of the covariates across subtypes.

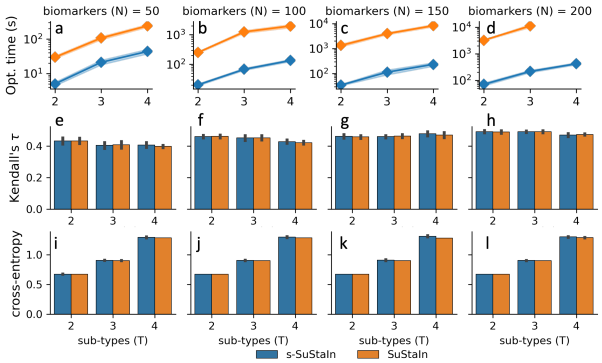
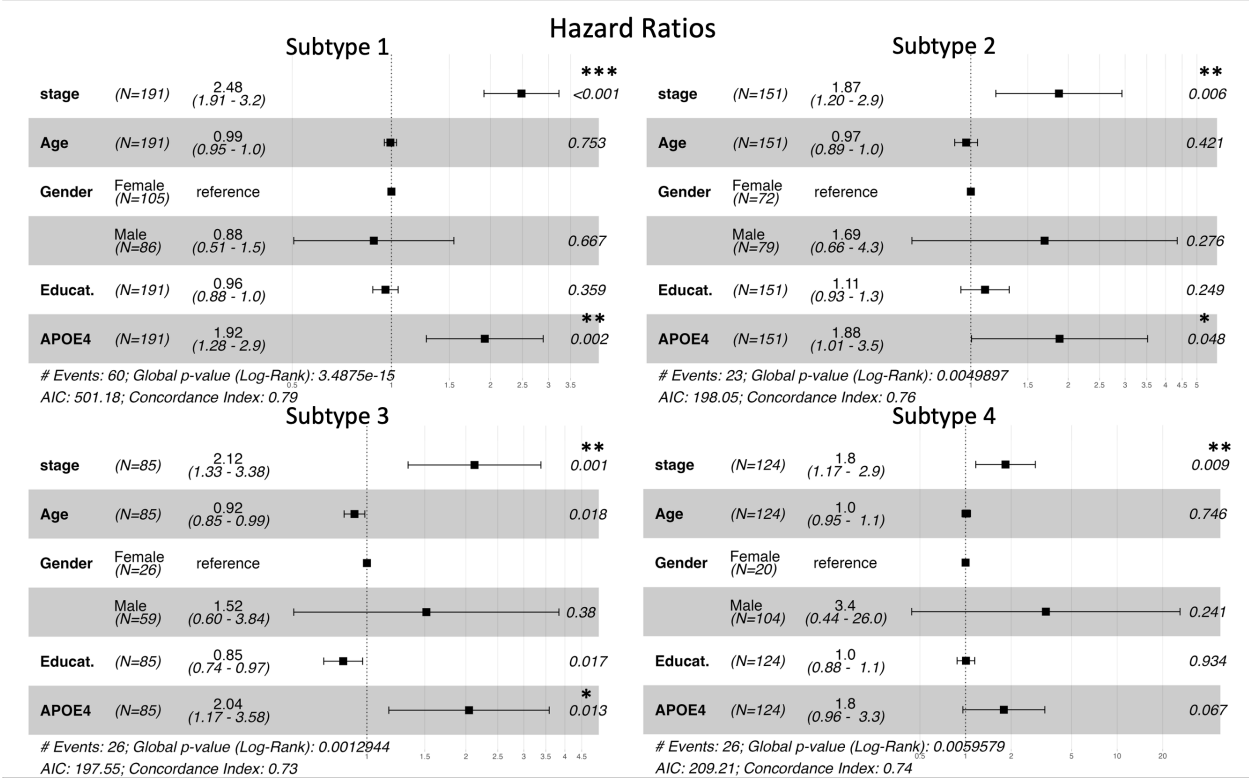


Figure 2: Simulation study
The models (s-SuStaIn and SuStaIn) are compared on simulated data with a known ground truth. The simulated data mimics cross-sectional observations of varying number of biomarkers (columns) from 200 subjects at different disease stages and following different progression trajectories. Data generation and experimental conditions are described in Section 5.1. The models are compared on 3 parameters – optimization times (Fig. a-d), recovery of ground-truth trajectories which represent subtypes (Fig. e-h), and their correct fractions in the data (Fig. i-l). Fig. a-d) The optimization step for s-SuStaIn is typically an order of magnitude faster than SuStaIn. Fig. e-h) The two models show comparable performance in recovering the ground-truth trajectories used to simulate the data. Fig. i-l) The two models also perform similarly well in inferring the subtype fractions.

5.2.2. Heterogeneity captured by subtypes

A s-SuStaIn model with 4 subtypes is compared against a similar s-SuStaIn model with only a single disease subtype to assess the advantages of modeling heterogeneous disease progression (both models trained on data described in Section 5.2). Conversion to AD among MCI subjects across subtypes and stages was compared using the Cox proportional-hazards model (as described in Section 5.2.1). The two models are compared on the Akaike information criterion (AIC), test statistic from log-rank test, and concordance index (Table 2). The s-SuStaIn model with 4 subtypes is also evaluated for 3-year incidence-rate of AD across the subtypes using Stevenson et al. (2013) (Figure 5(a)). Further, a t-SNE analysis is performed for the held-out MCI subjects to see if the lower dimensional projections capture differences in subtypes (Figure 6).

Table 1: Cox proportional-hazards modeling to predict conversion from MCI to AD.
The Cox proportional-hazards model is fitted to each of the 4 subtypes. Conversion from MCI to AD is considered as the event of interest. Each model uses subtype specific stage, age, gender, education and number of APOE4 alleles as covariates. Within each subtype, the disease stage shows a significant effect size which is comparable or greater than that for APOE4 - a known genetic risk for AD.



5.2.3. Subtype stability over time

The model fitted to Control and AD subjects is used to infer subtypes for MCI subjects in their follow-up visits. It is expected and desirable that follow-up visits will be assigned the same subtypes as the past visits for a subject (Figure 5(*b*)).

5.2.4. Visualizing progression in the brain

The progression trajectories learned by s-SuStaIn can be visualized by coloring the relevant regions on a brain atlas, in the order defined by the trajectory. We visualize the brain volumetric measurements in the trajectory over the DK brain atlas (Desikan et al., 2006) using the brainpainter software package (Marinescu et al., 2019a) in Figure 7 and Figure S2.

6. Results

6.1. s-SuStaIn is an order of magnitude faster

SuStaIn uses the expectation-maximization (EM) algorithm to learn the biomarker sequences characterizing the subtypes and their fractions in the data. s-SuStaIn uses the same optimization algorithm but shows an order of magnitude faster speeds, due to the reduced complexity in Equations (4) and (5). This is seen in Figure 2 a-d. Further, the speed-up increases with data dimensions (Figure S1). However this, does not impact s-SuStaIn's ability to learn subtype defining biomarker sequences as measured by Kendall-τ distance (Figure 2 e-h), or their relative fractions measured via cross-entropy between the ground truth and inferred values (Figure 2 i-l).

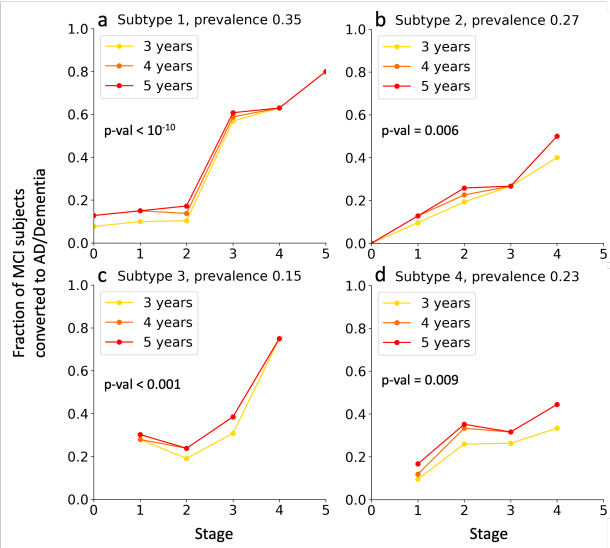


Figure 3: Conversion from MCI to AD as a function of disease stage, across disease subtypes.

Disease subtype and stage for MCI subjects are inferred from the s-SuStaIn model fitted to Control and AD subjects. The subtypes in the MCI data show varying profiles of AD progression risk as a function of disease stage. A Cox proportional-hazards model is fitted to each subtype, while adjusting for important covariates (age, gender, education and number of APOE4 alleles). In each subtype, the stage specific effect sizes for predicting AD progression was found to be significant.

6.2. Disease stages from s-SuStaIn predict progression to AD

Figure 3 shows the fraction of MCI subjects who convert to AD, as a function of subtype specific stage. For each subtype the risk of progression to AD/dementia is significantly associated with the disease stage, while adjusting for age, gender, education and the number of APOE4 allele copies (using a Cox proportional-hazards model). This shows the significance of the disease stage learned by s-SuStaIn in modeling the risk of progression among MCI subjects. Further, comparing the effects of the covariates in the Cox model shows that the effect sizes for disease-stage are comparable to those from APOE4 allele copies, a strong genetic risk factor in AD (Table 1).

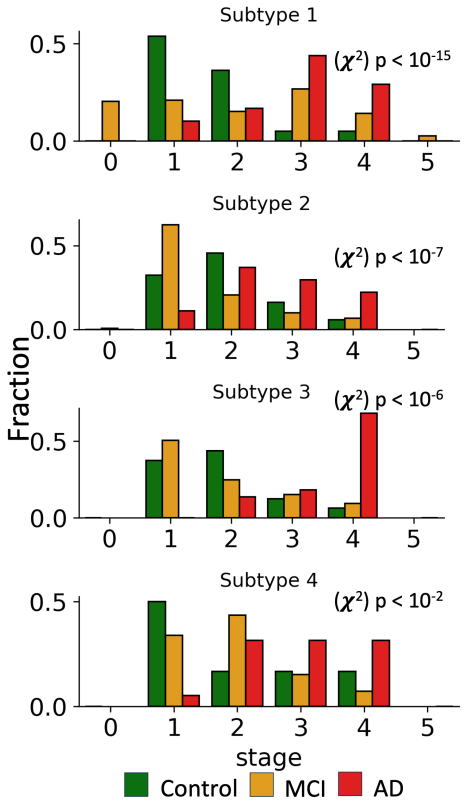


Figure 4: Disease groups show preferences for stages.

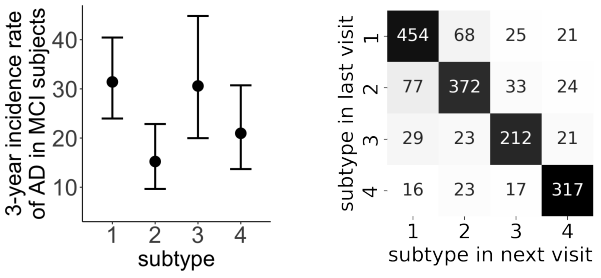
6.3. Stages correlate with known diagnosis

The s-SuStaIn model uses a uniform prior for the disease stage as described in Section 2.1, i.e. a priori disease stages are uniformly distributed. However, the estimated posteriors for disease stages show differences across diagnostic groups as seen in Figure 4 and tested using a χ^2 goodness of fit test. Controls and AD show preferences for opposite ends of staging.

6.4. Heterogeneity captured via subtypes

6.4.1. Conversion profiles across subtypes

Figure 3 show differences in AD risk profiles as a function of disease stages. Qualitatively in subtypes 1 and 3, the progression risk increases non-linearly across stages. Whereas in subtype 2, the increase in risk is rather linear and in subtype 4 it plateaus with increasing stages. Differences in stage specific effect sizes across subtypes are also observed in Table 1.



(*a*) Incidence rates (*b*) Follow-up visits

Figure 5: Differences in incidence rates across subtypes and their stability over time
a) 3-year incidence rate of AD in MCI subjects. The incidence rate is different across the subtypes (χ^2 p-val $< 5 \times 10^{-3}$). b) These subtypes show stability in their assignment over follow-up visits.

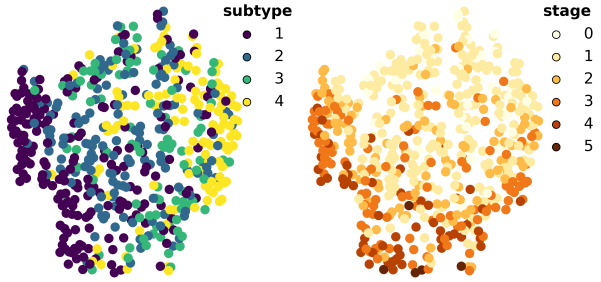


Figure 6: Separability of s-SuStain predicted subtypes and disease stages using t-SNE for visualization. The 2-D projections show differences in disease subtypes (left) and disease stages (right) in the 119-dimensional biomarker space among the held-out 551 MCI subjects.

6.4.2. AD incidence-rates across subtypes

Figure 5(*a*) shows a varying 3-year incidence-rate of AD among MCI subjects (χ^2 p-value < 0.005).

6.4.3. t-SNE analysis shows subtype and stage differences

Unsupervised analysis of the 119 biomarkers in held-out MCI subjects (n=551) using t-SNE shows differences in MCI subtypes and stages (Figure 6). The subtype and stage assignment was done using s-SuStaIn model learned from Controls and AD cases.

Table 2: Model comparison
Cox proportional-hazards model are fit to subtypes and stages inferred from two s-SuStaIn models - 1) s-SuStaIn with 4 subtypes (heterogeneous progression) and 2) s-SuStaIn with a single subtype (homogeneous progression). Other covariates used are age, gender, education, and number of APOE4 alleles. The inferred subtype and disease stage information from 1) does better in predicting conversion from MCI to AD. This is observed via 3 metrics - AIC (lower the better), p-value from log-rank test (lower the better) and Concordance index (higher the better). s-SuStaIn with 4 subtypes does better in all cases.

Metrics	**4 subtypes**	**1 subtype**
Parameters	8	5
AIC	1443.28	1469.89
p-val (log-rank)	1.25×10^{-24}	8.2×10^{-20}
Concordance index	0.773 ± 0.019	0.745 ± 0.022

6.4.4. Comparisons to a stage only model

The advantage of subtyping is further assessed by comparisons with another s-SuStaIn model with a single subtype i.e. all subjects follow the same disease progression trajectory and vary only in disease stages. A Cox proportional-hazards model is fit to inferred subtypes and stages from both s-SuStaIn models (along with covariates mentioned in Section 6.2). Table 2 shows comparisons across the two cases. The s-SuStaIn model that accounts for heterogeneity in disease progression (via subtypes) performs better than the s-SuStaIn model where all subjects follow the same progression trajectory and only vary in disease stage. This is seen through three metrics - AIC (lower the better), p-value from log rank test (lower the better) and concordance index (higher the better). In each case, the s-SuStaIn model with 4 subtypes does better than the stage only model. Further, effect size for subtype 2 shows significance at p < 0.05 (using subtype 1 as reference). This demonstrates that the model with subtypes does better at predicting progression from MCI to AD.

6.5. Stability of subtype assignment

Follow-up data from MCI subjects was assigned subtypes to study the stability in subtype assignment. Figure 5(*b*) shows subtype assignments for 1732 follow-up visits. Approximately 78% of all follow-up visits show the same subtype as the last visit.

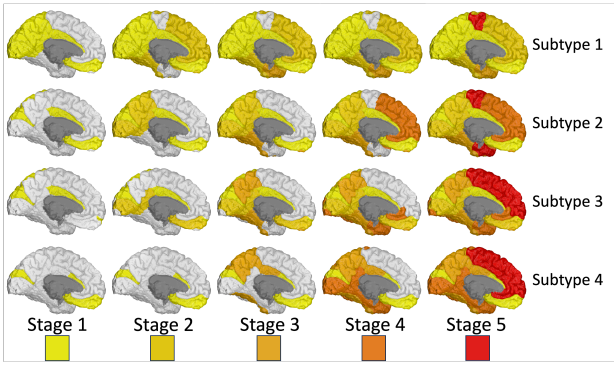


Figure 7: Visualization of disease progression across brain regions in the D-K brain atlas.

6.6. Contextual Evaluation of AD Subtypes

Even very early work suggested the presence of subtypes in AD Bondareff (1993) based on severity of neurofibrillary tangles and dementia progression Tariska and Urbanics (1995). Modern day machine learning methods, including s-SuStaIn, suggest 4 subtypes. Work by Jellinger (2020) suggested 4 major subtypes based on tau pathology and brain atrophy. Unsupervised learning of multimodal ADNI data also showed 4 clusters (Prakash et al., 2021), which varied as a function of brain volume and measured cognitive function. More research is necessary to better deduce the therapeutic implications of the 4 subtypes.

6.7. Mapping Progression to Brain Regions

The progression patterns shown in Figure 7 are described below and supported with literature findings.

- Stage 1 - All subtypes show a distinct ventral occipital lobe and medial frontal lobe involvement in Stage 1. The visual system and optic lobe plays a key role in the AD pathophysiology Cunha et al. (2016). Subtype 1 has a greater frontal lobe disease component in Stage 1. Subtypes 1, 2, 4 have more changes near the dorsal and posterior portions of the limbic lobe. Subtype 4 shows dynamic changes in only the most posterior portion of the limbic lobe.

- Stage 2 - In Stage 2, we see varying degrees of limbic and frontal lobe involvement across subtypes. The limbic system particularly the hippocampus and amygdala, are crucial for memory and emotion (Hopper and Vogel, 1976). Subtype 1 acquires changes to the entire occipital lobe. Subtypes 1 and 2 have substantive changes in the frontal lobe compared to Subtypes 3 and 4.

- Stage 3 - Stage 3 is where all subtypes are most visually differentiated. There is more frontal and temporal lobe involvement, which varies in degree across subtypes. Changes to the temporal lobe have been shown to be more pronounced in older subjects (Wilcock, 1983) and closely associated with semantic dementia (Galton et al., 2001). In Stage 3, changes to the temporal lobe spread beyond the limbic area. Additionally, subtypes 3 and 4 have a larger degree of dorsal occipital lobe sparing .

- Stage 4 - In stage 4, all subtypes show changes throughout the frontal and temporal lobes. Prior work has shown frontal lobe involvement to typically be present in all stages (Bhutani et al., 1992) with a sub-group of patients seeing earlier frontal lobe involvement (Farrow et al., 2007). The subtyping analysis presented here suggests subtypes 1, 2 have earlier, pervasive frontal lobe involvement in stages 1 and 2, whereas subtypes 3 and 4 see delays until Stages 4 and 5.

- Stage 5 - The parietal lobe is relatively spared in AD Brun and Gustafson (1976), which is also reflected in all subtypes in the present work.

7. Conclusions and future work

This work presents s-SuStaIn (scaling Subtype and Stage Inference), a data driven disease progression model which is amenable to working with larger biomarker sets. s-SuStaIn is typically an order of magnitude faster than its predecessor (SuStaIn). Using ADNI data, s-SuStaIn shows that the inferred subtypes and stages predict progression to AD among MCI subjects. In survival analysis using Cox proportional-hazards model, the adjusted effect sizes for disease stage are significant. The subtypes show difference in AD incidence-rates and meaningful progression trajectories when mapped to a brain atlas.

s-SuStaIn can be further modified to implicitly select biomarkers which are then used to determine the event sequence (Tandon et al., 2023a,b,c). Past work has shown progression in AD can be explained by a smaller The presented approach can also be extended to model biomarker evolution in disease as an accumulative process, as is done by z-score SuStaIn.

References

Leon M Aksman, Peter A Wijeratne, Neil P Oxtoby, Arman Eshaghi, Cameron Shand, Andre Altmann, Daniel C Alexander, and Alexandra L Young. pysustain: a python implementation of the subtype and stage inference algorithm. *SoftwareX*, 16: 100811, 2021.

G. E. Bhutani, D. Montaldi, David N. Brooks, and J. McCulloch. A neuropsychological investigation into frontal lobe involvement in dementia of the alzheimer type. *Neuropsychology*, 6(3):211–224, July 1992. ISSN 0894-4105. doi: 10.1037/0894-4105.6.3.211. URL http://dx.doi.org/10.1037/0894-4105.6.3.211.

William Bondareff. Evidence of subtypes of alzheimer's disease and implications for etiology. *Archives of General Psychiatry*, 50(5):350, May 1993. ISSN 0003-990X. doi: 10.1001/archpsyc.1993.01820170028004. URL http://dx.doi.org/10.1001/ARCHPSYC.1993.01820170028004.

A. Brun and L. Gustafson. Distribution of cerebral degeneration in alzheimer's disease: A clinicopathological study. *Archiv fr Psychiatrie und Nervenkrankheiten*, 223(1):15–33, 1976. ISSN 1433-8491. doi: 10.1007/bf00367450. URL http://dx.doi.org/10.1007/BF00367450.

Irene Y Chen, Rahul G Krishnan, and David Sontag. Clustering interval-censored time-series for disease phenotyping. In *Proceedings of the AAAI conference on artificial intelligence*, volume 36, pages 6211–6221, 2022.

J. P. Cunha, N. Moura-Coelho, R. P. Proença, A. Dias-Santos, J. Ferreira, C. Louro, and A. Castanheira-Dinis. Alzheimer's disease: A review of its visual system neuropathology. optical coherence tomography—a potential role as a study tool in vivo. *Graefe's Archive for Clinical and Experimental Ophthalmology*, 254(11):2079–2092, July 2016. ISSN 1435-702X. doi: 10.1007/s00417-016-3430-y. URL http://dx.doi.org/10.1007/s00417-016-3430-y.

Rahul S Desikan, Florent Ségonne, Bruce Fischl, Brian T Quinn, Bradford C Dickerson, Deborah Blacker, Randy L Buckner, Anders M Dale, R Paul Maguire, Bradley T Hyman, et al. An automated labeling system for subdividing the human cerebral cortex on mri scans into gyral based regions of interest. *Neuroimage*, 31(3):968–980, 2006.

Michael C Donohue, Hélène Jacqmin-Gadda, Mélanie Le Goff, Ronald G Thomas, Rema Raman, Anthony C Gamst, Laurel A Beckett, Clifford R Jack Jr, Michael W Weiner, Jean-François Dartigues, et al. Estimating long-term multivariate progression from short-term data. *Alzheimer's & Dementia*, 10:S400–S410, 2014.

Ronald Fagin, Ravi Kumar, Mohammad Mahdian, D Sivakumar, and Erik Vee. Comparing partial rankings. *SIAM Journal on Discrete Mathematics*, 20(3):628–648, 2006.

Tom F.D. Farrow, Subha N. Thiyagesh, Iain D. Wilkinson, Randolph W. Parks, Leanne Ingram, and Peter W.R. Woodruff. Fronto-temporal-lobe atrophy in early-stage alzheimer's disease identified using an improved detection methodology. *Psychiatry Research: Neuroimaging*, 155(1):11–19, May 2007. ISSN 0925-4927. doi: 10.1016/j.pscychresns.2006.12.013. URL http://dx.doi.org/10.1016/j.pscychresns.2006.12.013.

Nicholas C Firth, 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(7): 965–973, 2020.

Hubert M Fonteijn, Marc Modat, Matthew J Clarkson, Josephine Barnes, Manja Lehmann, Nicola Z Hobbs, Rachael I Scahill, Sarah J Tabrizi, Sebastien Ourselin, Nick C Fox, et al. An event-based model for disease progression and its application in familial alzheimer's disease and huntington's disease. *NeuroImage*, 60(3):1880–1889, 2012.

C. J. Galton, K. Patterson, K. Graham, M.A. Lambon-Ralph, G. Williams, N. Antoun, B.J. Sahakian, and J.R. Hodges. Differing patterns of temporal atrophy in alzheimer's disease and semantic dementia. *Neurology*, 57(2):216–225, July 2001. ISSN 1526-632X. doi: 10.1212/wnl.57.2.216. URL http://dx.doi.org/10.1212/WNL.57.2.216.

Sara Garbarino, Marco Lorenzi, and Alzheimer's Disease Neuroimaging Initiative. Modeling and inference of spatio-temporal protein dynamics across

brain networks. In *Information Processing in Medical Imaging: 26th International Conference, IPMI 2019, Hong Kong, China, June 2–7, 2019, Proceedings 26*, pages 57–69. Springer, 2019.

M. W. Hopper and F. S. Vogel. The limbic system in Alzheimer's disease. A neuropathologic investigation. *Am J Pathol*, 85(1):1–20, Oct 1976.

Bruno M Jedynak, Andrew Lang, Bo Liu, Elyse Katz, Yanwei Zhang, Bradley T Wyman, David Raunig, C Pierre Jedynak, Brian Caffo, Jerry L Prince, et al. A computational neurodegenerative disease progression score: method and results with the alzheimer's disease neuroimaging initiative cohort. *Neuroimage*, 63(3):1478–1486, 2012.

Kurt A. Jellinger. Pathobiological subtypes of alzheimer disease. *Dementia and Geriatric Cognitive Disorders*, 49(4):321–333, 2020. ISSN 1421-9824. doi: 10.1159/000508625. URL http://dx.doi.org/10.1159/000508625.

Changhee Lee and Mihaela Van Der Schaar. Temporal phenotyping using deep predictive clustering of disease progression. In *International conference on machine learning*, pages 5767–5777. PMLR, 2020.

Dan Li, Samuel Iddi, Wesley K Thompson, Michael C Donohue, and Alzheimer's Disease Neuroimaging Initiative. Bayesian latent time joint mixed effect models for multicohort longitudinal data. *Statistical methods in medical research*, 28(3):835–845, 2019.

Marco Lorenzi, Maurizio Filippone, Giovanni B Frisoni, Daniel C Alexander, Sébastien Ourselin, Alzheimer's Disease Neuroimaging Initiative, et al. Probabilistic disease progression modeling to characterize diagnostic uncertainty: application to staging and prediction in alzheimer's disease. *NeuroImage*, 190:56–68, 2019.

Răzvan V Marinescu, Arman Eshaghi, Daniel C Alexander, and Polina Golland. Brainpainter: A software for the visualisation of brain structures, biomarkers and associated pathological processes. In *Multimodal Brain Image Analysis and Mathematical Foundations of Computational Anatomy: 4th International Workshop, MBIA 2019, and 7th International Workshop, MFCA 2019, Held in Conjunction with MICCAI 2019, Shenzhen, China, October 17, 2019, Proceedings 4*, pages 112–120. Springer, 2019a.

Răzvan V Marinescu, Neil P Oxtoby, Alexandra L Young, Esther E Bron, Arthur W Toga, Michael W Weiner, Frederik Barkhof, Nick C Fox, Polina Golland, Stefan Klein, et al. Tadpole challenge: Accurate alzheimer's disease prediction through crowdsourced forecasting of future data. In *Predictive Intelligence in Medicine: Second International Workshop, PRIME 2019, Held in Conjunction with MICCAI 2019, Shenzhen, China, October 13, 2019, Proceedings 2*, pages 1–10. Springer, 2019b.

Shahriar Noroozizadeh, Jeremy C Weiss, and George H Chen. Temporal supervised contrastive learning for modeling patient risk progression. In *Machine Learning for Health (ML4H)*, pages 403–427. PMLR, 2023.

Neil P Oxtoby. Data-driven disease progression modeling. *Machine Learning for Brain Disorders*, pages 511–532, 2023.

CS Parker, NP Oxtoby, DC Alexander, H Zhang, and Alzheimer's Disease Neuroimaging Initiative. S-ebm: Generalising event-based modelling of disease progression for simultaneous events. *bioRxiv*, pages 2022–07, 2022.

Pierre-Emmanuel Poulet and Stanley Durrleman. Mixture modeling for identifying subtypes in disease course mapping. In *International Conference on Information Processing in Medical Imaging*, pages 571–582. Springer, 2021.

Jayant Prakash, Velda Wang, Robert E. Quinn, and Cassie S. Mitchell. Unsupervised machine learning to identify separable clinical alzheimer's disease sub-populations. *Brain Sciences*, 11(8): 977, July 2021. ISSN 2076-3425. doi: 10.3390/brainsci11080977. URL http://dx.doi.org/10.3390/brainsci11080977.

Yuchao Qin, Mihaela van der Schaar, and Changhee Lee. T-phenotype: Discovering phenotypes of predictive temporal patterns in disease progression. In *International Conference on Artificial Intelligence and Statistics*, pages 3466–3492. PMLR, 2023.

J-B Schiratti, Stéphanie Allassonniere, Alexandre Routier, Alzheimers Disease Neuroimaging Initiative, O Colliot, and S Durrleman. A mixed-effects model with time reparametrization for longitudinal univariate manifold-valued data. In *Information Processing in Medical Imaging: 24th International*

Conference, IPMI 2015, Sabhal Mor Ostaig, Isle of Skye, UK, June 28-July 3, 2015, Proceedings 24, pages 564–575. Springer, 2015a.

Jean-Baptiste Schiratti, Stéphanie Allassonniere, Olivier Colliot, and Stanley Durrleman. Learning spatiotemporal trajectories from manifold-valued longitudinal data. *Advances in neural information processing systems*, 28, 2015b.

M Stevenson, T Nunes, J Sanchez, R Thornton, J Reiczigel, J Robison-Cox, P Sebastiani, P Solymos, K Yoshida, G Jones, et al. epir: An r package for the analysis of epidemiological data. r package version 0.9-48. *Vienna, Austria: R Foundation for Statistical Computing*, 2013.

Raghav Tandon, Anna Kirkpatrick, and Cassie S Mitchell. sebm: Scaling event based models to predict disease progression via implicit biomarker selection and clustering. In *International Conference on Information Processing in Medical Imaging*, pages 208–221. Springer, 2023a.

Raghav Tandon, Allan I Levey, James J Lah, Nicholas T Seyfried, and Cassie S Mitchell. Machine learning selection of most predictive brain proteins suggests role of sugar metabolism in alzheimer's disease. *Journal of Alzheimer's Disease*, 92(2):411–424, 2023b.

Raghav Tandon, Liping Zhao, Caroline M Watson, Morgan Elmor, Craig Heilman, Katherine Sanders, Chadwick M Hales, Huiying Yang, David W Loring, Felicia C Goldstein, et al. Predictors of cognitive decline in healthy middle-aged individuals with asymptomatic alzheimer's disease. *Research Square*, 2023c.

Péter Tariska and Kinga Urbanics. Clinical subtypes of alzheimer's disease. *Archives of Gerontology and Geriatrics*, 21(1):13–20, July 1995. ISSN 0167-4943. doi: 10.1016/0167-4943(95)00641-w. URL http://dx.doi.org/10.1016/0167-4943(95)00641-W.

Vikram Venkatraghavan, Esther E Bron, Wiro J Niessen, and Stefan Klein. A discriminative event based model for alzheimer's disease progression modeling. In *Information Processing in Medical Imaging: 25th International Conference, IPMI 2017, Boone, NC, USA, June 25-30, 2017, Proceedings 25*, pages 121–133. Springer, 2017.

Peter A Wijeratne, Daniel C Alexander, and Alzheimer's Disease Neuroimaging Initiative. Learning transition times in event sequences: The temporal event-based model of disease progression. In *International Conference on Information Processing in Medical Imaging*, pages 583–595. Springer, 2021.

G.K. Wilcock. The temporal lobe in dementia of alzheimer's type. *Gerontology*, 29(5):320–324, 1983. ISSN 1423-0003. doi: 10.1159/000213133. URL http://dx.doi.org/10.1159/000213133.

Alexandra L Young, 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(9):2564–2577, 2014.

Alexandra L Young, Neil P Oxtoby, Sebastien Ourselin, Jonathan M Schott, Daniel C Alexander, Alzheimer's Disease Neuroimaging Initiative, et al. A simulation system for biomarker evolution in neurodegenerative disease. *Medical image analysis*, 26(1):47–56, 2015.

Alexandra L Young, Razvan V Marinescu, Neil P Oxtoby, Martina Bocchetta, Keir Yong, Nicholas C Firth, David M Cash, David L Thomas, Katrina M Dick, Jorge Cardoso, et al. Uncovering the heterogeneity and temporal complexity of neurodegenerative diseases with subtype and stage inference. *Nature communications*, 9(1):4273, 2018.

Alexandra L Young, Leon M Aksman, Daniel C Alexander, Peter A Wijeratne, and Alzheimer's Disease Neuroimaging Initiative. Subtype and stage inference with timescales. In *International Conference on Information Processing in Medical Imaging*, pages 15–26. Springer, 2023.

Appendix A. Speed-up factor with increasing data dimensions

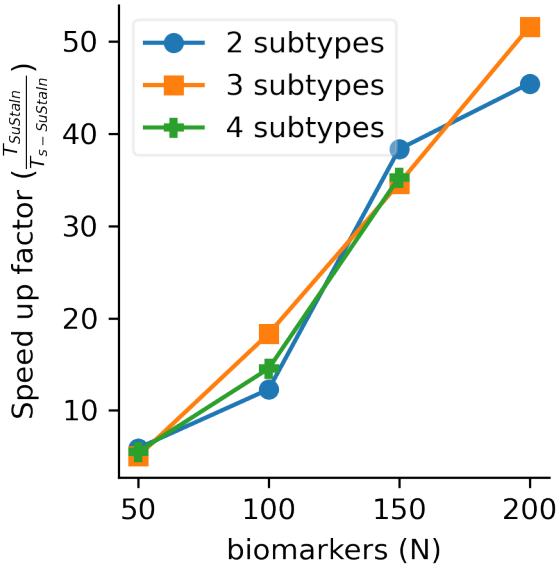


Figure S1: Speed-up factor for s-SuStaIn (compared to SuStaIn).
The speed-up factor for s-SuStaIn over SuStaIn is not constant, but increases with the number of biomarkers (data dimensionality). This is another way to see that s-SuStaIn scales favorably with increasing data dimensions, as compared to SuStaIn.

Appendix B. Spatial - (pseudo) temporal disease progression patterns observed from outer cortical, sub-cortical, top and bottom views

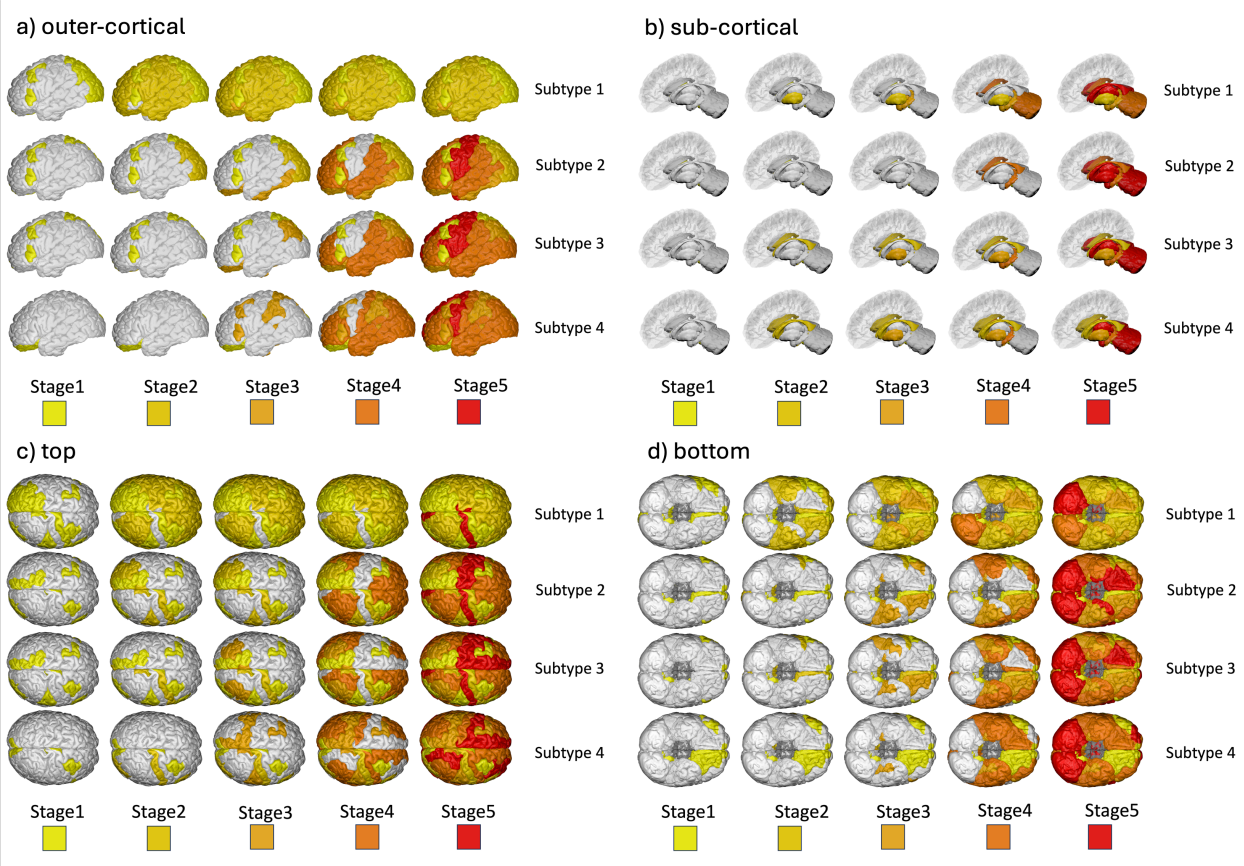


Figure S2: Visualization of disease progression across brain regions seen from different brain views. Disease progression trajectories (subtypes) are overlaid on the brain regions in Desikan-Killiany (D-K) brain atlas. These views are a) outer-cortical, b) sub-cortical, c) top, d) bottom. The progression trajectories across subtypes, and at each stage are shown. The trajectories are derived from 119 biomarkers which represent different brain regions and cannot be all seen a single view. This helps to understand the differences across the subtypes in terms of the underlying physiological changes.

Appendix C. A landscape of disease progression models

Table 3: Data driven disease progression models categorized based on data requirement and ability to model heterogeneity in disease progression (modified from Oxtoby (2023)).

Disease progression type	Cross-sectional data	Longitudinal data
Homogeneous disease progression	EBM (Fonteijn et al., 2012; Young et al., 2014) D-EBM (Venkatraghavan et al., 2017) KDE-EBM (Firth et al., 2020) s-EBM (Tandon et al., 2023a)	DPS (Jedynak et al., 2012) LTJMM (Donohue et al., 2014; Li et al., 2019) Time-warping (Schiratti et al., 2015a,b) GPPM (Lorenzi et al. (2019), Garbarino et al. (2019)) T-EBM (Wijeratne et al., 2021)
Heterogeneous disease progression	SuStaIn (Young et al., 2018)	AC-TPC (Lee and Van Der Schaar, 2020) T-phenotype (Qin et al., 2023) Temporal-SCL (Noroozizadeh et al., 2023) T-SuStaIn (Young et al., 2023) SubLign (Chen et al., 2022) Course Maps (Poulet and Durrleman, 2021)