Research Article

Constructing longitudinal disease progression curves using sparse, short-term individual data with an application to Alzheimer's disease

C.A. Budgeon,[a,b*†]   K. Murray,[c]   B.A. Turlach,[a]   S. Baker,[d] V.L. Villemagne,[e,f] S.C. Burnham[b]   and for the Alzheimer's Disease Neuroimaging Initiative[‡]

In epidemiology, cohort studies utilised to monitor and assess disease status and progression often result in short-term and sparse follow-up data. Thus, gaining an understanding of the full-term disease pathogenesis can be difficult, requiring shorter-term data from many individuals to be collated. We investigate and evaluate methods to construct and quantify the underlying long-term longitudinal trajectories for disease markers using short-term follow-up data, specifically applied to Alzheimer's disease. We generate individuals' follow-up data to investigate approaches to this problem adopting a four-step modelling approach that (i) determines individual slopes and anchor points for their short-term trajectory, (ii) fits polynomials to these slopes and anchor points, (iii) integrates the reciprocated polynomials and (iv) inverts the resulting curve providing an estimate of the underlying longitudinal trajectory. To alleviate the potential problem of roots of polynomials falling into the region over which we integrate, we propose the use of non-negative polynomials in Step 2. We demonstrate that our approach can construct underlying sigmoidal trajectories from individuals' sparse, short-term follow-up data. Furthermore, to determine an optimal methodology, we consider variations to our modelling approach including contrasting linear mixed effects regression to linear regression in Step 1 and investigating different orders of polynomials in Step 2. Cubic order polynomials provided more accurate results, and there were negligible differences between regression methodologies. We use bootstrap confidence intervals to quantify the variability in our estimates of the underlying longitudinal trajectory and apply these methods to data from the Alzheimer's Disease Neuroimaging Initiative to demonstrate their practical use. Copyright © 2017 John Wiley & Sons, Ltd.

Keywords: Alzheimer's disease; sigmoidal curves; longitudinal trajectories

1. Introduction

Alzheimer's disease (AD), the leading cause of dementia, represents a worldwide epidemic and currently costs 1.09% of global gross domestic product [1]. AD is a neurodegenerative disease resulting in loss of cognitive function due to structural changes in the neocortex. These changes are mainly encompassed by

[a]*Centre for Applied Statistics, University of Western Australia, Crawley, Western Australia, Australia*
[b]*eHealth, Health and Biosecurity, Commonwealth Scientific and Industrial Research Organisation (CSIRO), Floreat, Western Australia, Australia*
[c]*School of Population and Global Health, University of Western Australia, Crawley, Western Australia, Australia*
[d]*Janssen Research and Development, Titusville, NJ, USA*
[e]*Department of Nuclear Medicine and Centre for PET, Austin Health, Heidelberg, Victoria, Australia*
[f]*The Florey Institute for Neuroscience and Mental Health, The University of Melbourne, Victoria, Australia*
*Correspondence to: C.A. Budgeon, Centre for Applied Statistics (M019), University of Western Australia, 35 Stirling Highway, Crawley, WA 6009, Australia.
†*E-mail: charley.budgeon@uwa.edu.au*

the formation of extracellular Aβ-amyloid (Aβ) plaques and intracellular neurofibrillary tangles [2, 3]. Neurofibrillary tangles consist of intra-neuronal paired helical filaments of phosphorylated Tau protein. They are not specific to AD and can be found in other neurodegenerative diseases such as frontotemporal dementia. Aβ plaques are composed of extracellular aggregates of amyloid β-peptide, which is thought to be toxic to neurons, and current therapeutic effort for AD is concentrated on reducing neocortical Aβ burden and/or toxicity.

Current thinking suggests that once clinical symptoms are present and diagnoses can be made, irreversible structural damage has already occurred [4]. Therefore, therapeutic strategies are predicted to have most efficacy if delivered at the prodromal and even preclinical stages of disease [5]. To be able to appropriately time such intervention and therapeutic strategies, an understanding about the disease pathogenesis is crucial. Longitudinal monitoring of individuals with AD is becoming more common through a number of cohort studies around the world. Two examples include the Alzheimer's Disease Neuroimaging Initiative (ADNI) [6] and the Australian Imaging, Biomarker and Lifestyle Study (AIBL) [7] both of which record data including neocortical Aβ burden as measured by positron emission tomography (PET). However, these studies generally have sparse information with recordings taken every 1–2 years and with less than 8 years of total follow-up on any individual in AIBL and generally less than 10 years of follow-up in ADNI, although not all measures span this time. Consequently, the full disease course that is suspected to evolve over an approximate 30-year period [8] is not fully captured by these studies.

Given therapeutic targets are aimed at modifying neocortical Aβ burden and it is desirable to implement these at the prodromal or preclinical phase of disease, it is paramount to understand the longitudinal deposition of Aβ plaques in the neocortex. In this paper, we specifically refer to AD, whose long-term trajectory, with respect to certain biomarkers, for example, neocortical Aβ burden, has been shown to be sigmoidal in nature, characterised by an initial phase of steady, slow increase, then a stage of rapid growth, followed finally by a plateau. This is described in [9, 10] with the latter of these providing a thorough discussion and justification of the use of the sigmoidal shape in AD progression.

To date, several studies have conducted longitudinal modelling in various areas of AD research (for example, Huijbers *et al.* [11], Yau *et al.* [12], Bateman *et al.* [13] and Samtani *et al.* [14] and references therein). Bateman *et al.* [13] considered participants who were at risk of carrying a mutation for autosomal dominant AD and collected information on their parents in relation to estimated years from expected symptom onset to determine the relative order and magnitude of pathophysiological changes. Capuano *et al.* [15] provide a method for longitudinal modelling for sigmoidal trajectories to incorporate correlated data to analyse retrospective cognition data for deceased individuals. However, in this instance, the individuals had a known end-point, and longitudinal trajectories were estimated up until death, which requires having sufficient data on all individuals. Sabuncu *et al.* [16] investigated the use of a cumulative diffusion model to predict longitudinal atrophy also proposing the use of a sigmoidal shape without directly modelling the progression time. Yang *et al.* [17] modelled disease progression time assuming an exponential model for ADAS-Cog 13 scores, using late mild cognitively impaired (MCI) and AD individuals. Jack *et al.* [18] hypothesised that biomarker trajectories have a sigmoidal shape but did not directly test this because of limitations in the available data, which was sparse in the areas it was most needed in order to detect areas of acceleration and deceleration. Instead, they aimed to characterise the shape of AD progression as a function of the mini-mental state exam (MMSE), where using cognition measures has been seen as a common solution to modelling AD progression without the knowledge of specific disease progression times (for example, Caroli *et al.* [19]). Furthermore, they state that creating reliable models that accurately describe the full trajectory of disease will require significant amounts of longitudinal data spanning from healthy status through to the end stage of disease. Aisen *et al.* [20] also indicate that to test the proposed trajectory of disease progression, it is essential to acquire long-term longitudinal follow-up. ADNI are currently in the process of collecting such data.

Most of these works either do not consider the problem created by having short-term data on the study time scale in order to provide long-term predictions on a disease progression scale or deal with the sigmoidal shape covering the full time span of disease progression. To our knowledge, only Villemagne *et al.* [8] have made efforts to address both of these issues making initial progress in attempting to explore longitudinal trajectories using short-term data. However, in attempting to apply this methodology, it became apparent that data-specific decisions needed to be made in order to use their approach and that expansion and robustification of the various steps of this method was necessary in order to make it readily replicable and effective. As such, we aim to develop (and evaluate) methodologies that will take sparse short-term follow-up data from a number of individuals and construct the underlying full-term trajectory.

The rest of this paper is organised as follows. Section 2 introduces the motivation behind this research. In Section 3, we describe the methods developed and a detailed description of our four-step approach, including a description of bootstrapping techniques used. In Section 4, we outline the numerical experiment and describe the results and model comparisons in Section 5. We illustrate our four-step approach using the publicly available ADNI data in Section 6 and conclude in Section 7 with a short discussion. Throughout, we specifically demonstrate our methods in order to quantify the disease pathogenesis associated with the longitudinal neocortical Aβ burden.

2. Motivation and background

Typically, in AD and many other diseases, data on individuals are available over short periods of time only. Currently, in ADNI, or other databases that focus on AD, any individual can have either a single baseline measurement or as few as two longitudinal data points. Having such short-term individual follow-up data poses obvious challenges in approximating long-term trajectories for disease progression. Whilst the methodology that we are proposing in this paper is suitable for the construction of sigmoidal, longitudinal disease progression curves, our specific focus will be on AD, whose long-term trajectory for certain biomarkers has been shown to have such a shape. Assuming each individual's disease progression has an underlying sigmoidal shape, we aim to estimate the population trajectory through a novel four-step modelling approach that is outlined in the succeeding discussion and graphically illustrated in Figure 1. Our methodology uses short-term, longitudinal data on individuals measured in follow-up time (t) since entering the study (a subset of such data is shown in the left panel of Figure 1) to estimate the longitudinal disease progression curve in terms of disease progression time (τ) (right panel of Figure 1).

Initially, we consider responses y_{ij}, which are measures for a surrogate marker of AD on the i-th individual, at the j-th follow-up time point t_{ij}. From these, we calculate estimates of the mean $\hat{\mu}$ and slope $\hat{\mu}'$ for each individual's disease progression by fitting a linear regression model to each individual's data (subset shown in Figure 1, Step 1). This fitted model is evaluated at the midpoint $\bar{t}_{i\bullet} = \frac{1}{n_i} \sum_{j=1}^{n_i} t_{ij}$ of an individual's follow-up time points to give estimates:

$$(\hat{\mu}(\bar{t}_{i\bullet}), \hat{\mu}'(\bar{t}_{i\bullet})) \tag{1}$$

The second step involves fitting a (non-negative) polynomial to the estimated means and slopes from (1) (Figure 1, Step 2). If we denote the fitted values from this polynomial as $\hat{\mu}'(t)$, of which we take the reciprocal, and continue to denote the scale of the input variable for the polynomial as $\hat{\mu}(t)$, we obtain

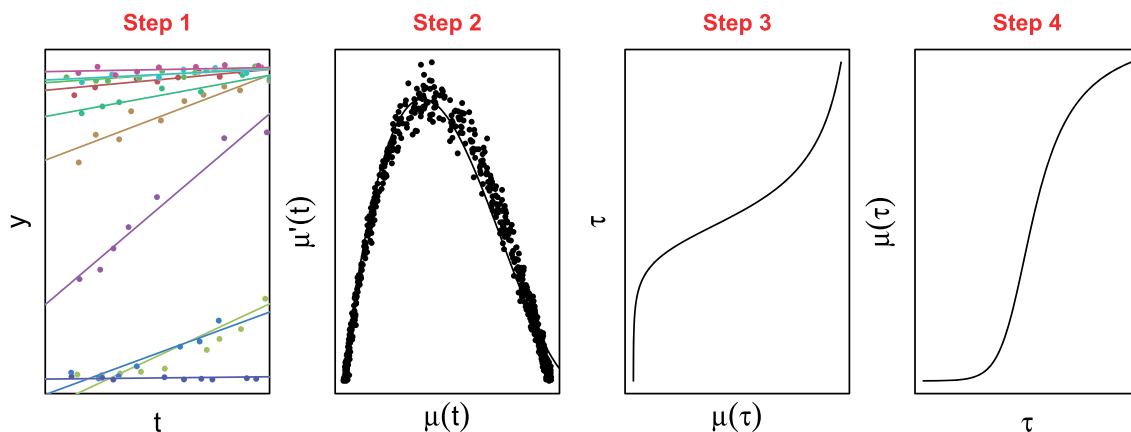$$\left(\hat{\mu}(t), \frac{1}{\hat{\mu}'(t)} \right) \tag{2}$$

Figure 1. Step 1 – individual simulated longitudinal data for a surrogate marker of Alzheimer's disease (AD) against study time for a subset of individuals. Step 2 – estimated means against their respective slopes based on the fitted regression lines, with a polynomial fit. Step 3 – integrating the reciprocal of the fitted polynomial. Step 4 – inverting the function to provide a surrogate marker of AD against disease progression time.

In the third step, we integrate either analytically or numerically the function from (2) (Figure 1, Step 3), that is, we integrate $\frac{1}{\hat{\mu}'(t)}$ with respect to $\hat{\mu}(t)$ and the result can be written as

$$(\hat{\mu}(\tau), \tau) \tag{3}$$

where τ is now disease progression time rather than follow-up time t. We note that the origin of the disease progression time is such that for $\tau = 0$, $\hat{\mu}(\tau)$ equals the smallest observed mean among the $\hat{\mu}(\bar{t}_{i\bullet})$, making it data dependent. However, this is not problematic, because a typical question of interest is how long it takes to progress from one specific disease progression state to another, rendering the origin inconsequential. The final step involves an inversion of (3) to provide a long-term trajectory of disease progression (Figure 1, Step 4):

$$(\tau, \hat{\mu}(\tau)) \tag{4}$$

These final two steps allow us to progress from the curve fitted to the means and slopes, that is, the short-term individual changes in a surrogate marker of AD, to a final estimated curve that can be interpreted as a long-term trajectory for disease progression enabling us to infer about the long-term trajectory of an individual.

It may seem natural in (2) to determine the functional form of $\frac{1}{\mu'(t)}$ from the assumed sigmoidal shape of the disease progression curve and use nonlinear least squares to fit the resulting function to the data calculated in (1). However, if each individual has its own sigmoidal curve arising from a nonlinear mixed model based on a specific sigmoidal function, as, for example, in our simulation study in which we use the five-parameter logistic (5PL) function, the population curve (i.e. the mean of all the individual's curves) is itself not necessarily of the same algebraic form as the individual's curves, and it is highly unlikely this population curve will have a closed form expression. Thus, basing the form of the relationship between the means and slopes in this step on such backtracking from an assumed specific sigmoidal population curve may be unreasonable. In light of this, and the fact that the means and slopes generated from (1) typically produce an inverted u-shaped point cloud, a natural and relatively flexible candidate set of models to consider here is the set of low-order polynomials.

As mentioned previously, it is generally of interest to predict the length of time it takes for an individual to progress from one point on the curve to another. Specifically, we wish to know how many years an individual will take to move from the healthy control (HC) state to the beginning stages of AD, and what the error is associated with this prediction. Quantifying this time and understanding the relationship between surrogate markers of AD and disease progression will enable us to incorporate the design and timing of therapeutic information. It will also assist in the planning for events such as residential care and predicting decline. Subsequently, this progression time could assist in personalising the care of an individual.

3. Methods

In order to fit regression models to short-term data from all of the individuals, we consider two approaches. We can fit a simple linear regression model (LM) for each subject i, each with n_i time points t_{ij}, of the following form:

$$y_{ij} = \beta_{0i} + \beta_{1i}t_{ij} + \epsilon_{ij} \tag{5}$$

where β_{0i} and β_{1i} are the intercept and slope, respectively, of the regression line fitted to individual i and $\epsilon_{ij} \sim N(0, \sigma^2)$. Alternatively, we can fit a linear mixed model (LMM) of the following form:

$$y_{ij} = (\beta_0 + b_{0i}) + (\beta_1 + b_{1i})t_{ij} + \epsilon_{ij} \tag{6}$$

where β_0 and β_1 are the intercept and slope, respectively, of the (population) regression line, b_{0i} and b_{1i} are the deviations from the population parameters for individual i modelled as being bivariate normal with mean 0 and general variance–covariance matrix and $\epsilon_{ij} \sim N(0, \sigma^2)$. We use either (5) or (6) to extract estimated slopes, $\hat{\mu}'_i$s, and fitted values, $\hat{\mu}_i$s, evaluated at $\bar{t}_{i\bullet}$, and these estimates are then modelled against each other as described in Section 2.

Whilst LMMs are preferred over LMs (see among others Verbeke and Molenberghs [21] or Fitzmaurice *et al.* [22]) if the models in (5) or (6) are of primary interest, it is not clear whether these advantages carry

through in our application. Hence, we not only investigate further the LM approach in the first step, as implemented in Villemagne *et al.* [8], but also extend the methodology to explore the use of LMMs.

Previous research has suggested that the relationship between the means and slopes is concave in nature [8]; thus, for simplicity, a low-order polynomial of the following form should provide a reasonable model for this relationship:

$$p(x) = p(x; \boldsymbol{\gamma}) = \gamma_0 + \gamma_1 x + \gamma_2 x^2 + \cdots + \gamma_q x^q \tag{7}$$

where q is the degree of the polynomial and $\boldsymbol{\gamma} = (\gamma_0, \gamma_1, \dots, \gamma_q)^T$. We restrict the order to cubics ($q = 3$) and quintics ($q = 5$) as they provide, in our experience, reasonable fits and are relatively simple to use in practice, whereas the use of higher degree polynomials becomes more complex in subsequent steps of our approach and are likely to result in non-concave fits.

However, when fitting polynomials to the estimated means and slopes in (1), the roots of the fitted polynomial may fall inside the range of the data, creating obvious problems in the subsequent integration step after reciprocating the fitted polynomial. In this situation, we need to impose constraints to ensure that the roots of the fitted polynomial do not fall within the range of the estimated means. Specifically, we ensure that the fitted polynomial is positive over this range by enforcing non-negativity over a slightly larger range. In order to enforce non-negativity, we adopt a sum of squared polynomials parametisation (see, for example, Brickman [23] for details and, more recently, Murray *et al.* [24]). Typically, in the problems we consider, we are using cubics and quintics (i.e. $q = 2K + 1$ for some integer K), which are now constrained to be non-negative over an interval $[a, b]$; hence, (7) becomes

$$p(x) = (x - a)p_1(x)^2 + (b - x)p_2(x)^2, \quad \forall x \in \mathbb{R} \tag{8}$$

where $p_1(x)$ and $p_2(x)$ are polynomials being at most of degree K.

3.1. Bootstrap variance estimation

In order to quantify the uncertainty in the disease progression time at a certain level of a biomarker or surrogate marker, we can estimate confidence intervals (CI) for τ given a value of $\mu(\tau)$. An analytical solution is not straightforward, so we use a bootstrap methodology to estimate point-wise CIs. There are numerous bootstrap approaches that could be utilised, for example, the parametric bootstrap, the residual bootstrap and the cases bootstrap all of which are detailed in De Leeuw *et al.* [25]. Field and Welsh [26] and Field *et al.* [27] also suggest three main approaches for bootstrapping LMMs including the random-effects bootstrap, the modified bootstrap and the generalised cluster bootstrap. Sherman and le Cessie [28] call the cases bootstrap the 'all-block' bootstrap and advocate its use. Chambers and Chandra [29] indicate that it is hard to justify the use of the parametric bootstrap when the model is an approximation, in particular if the distributional assumptions on the random effects are more for convenience and not presumed to be part of the data-generating process. The same idea is expressed by Thai *et al.* [30] who state that a non-parametric bootstrap approach is more suitable than the standard parametric bootstrap if the distributional assumptions of the random effects are not tenable. In light of this discussion and our extensive numerical experiences, we advocate and describe the use of the cases bootstrap for our methodology. This consists of resampling with replacement the entire subjects (case), that is, all longitudinal measurements for a resampled individual are included. Specifically, our approach can be described as follows:

(1) Let $(\mathbf{t}_i, \mathbf{y}_i)$ be the predictor and response vectors for individual i, that is, the vectors for n individuals is $(\mathbf{t}_1, \mathbf{y}_1), (\mathbf{t}_2, \mathbf{y}_2), \dots, (\mathbf{t}_n, \mathbf{y}_n)$.
(2) For $b = 1, \dots, B$ bootstrap iterations, do the following:

    (a) Resample r times, where $r = n$, with replacement from $\{1, \dots, n\}$. If i is the jth resampled value, set $(\mathbf{t}_j^*, \mathbf{y}_j^*)$ to be $(\mathbf{t}_i, \mathbf{y}_i)$.
    (b) With each bootstrap sample, follow the procedure described in Section 2, where we initially calculate estimated slopes and means for each case using either LMs or LMMs, fit a non-negative polynomial and integrate the reciprocal of this polynomial. By doing so, we obtain at any given value $\mu(\tau)$ bootstrap replicates $\hat{\tau}^{*,b}$ of $\hat{\tau}$ possibly by interpolation.

(3) To obtain CIs for τ, for a given value of $\mu(\tau)$, we use the percentile method on our bootstrap replicates $\hat{\tau}^{*,b}$ [31]. For each bootstrap iteration, the time taken to progress from one point in the progression curve to another is calculated, and percentiles are taken of these times to obtain our CIs.

4. Numerical experiments

Based on the methodology described in Section 3, we conducted an extensive simulation study for modelling short-term follow-up data in order to construct a long-term disease progression curve for AD. For this investigation, we generated data from a 5PL function equivalent to that described in Gottschalk and Dunn [32]:

$$R(\tau) = c + \frac{f - c}{[1 + e^{-a(\tau - b^*)}]^g} \qquad (9)$$

where τ is disease progression time, $a > 0$ is a slope parameter, c is the lower asymptote, f is the upper asymptote, g is an asymmetry parameter, b^* is $b + \log(2^{1/g} - 1)/a$ and b is the mid-point. This function reflects the sigmoidal nature proposed by previous researchers for disease progression [9, 10].

In our simulations, all parameters in (9) are sampled using the distributions described as follows:

$$b \sim U(10, 16)$$
$$a \sim N_-^+(0.40, 0, \infty, 0.02^2)$$
$$c \sim N_-^+(0.01, 0, \infty, 0.005^2)$$
$$f \sim N_-^+(0.98, -\infty, 1, 0.005^2)$$
$$g \sim N(20, 1)$$

where $N_-^+(\mu, \mu^-, \mu^+, \sigma^2)$ denotes the two-sided truncated normal distribution, that is, the normal distribution, $N(\mu, \sigma^2)$, truncated to the interval $[\mu^-, \mu^+]$. The values for each distribution are chosen such that they provide a reasonable sigmoidal curve reflecting a real-life scenario. An example of a 5PL curve is shown in Figure 2 and shows disease progression spanning approximately 30 years for some surrogate marker of AD.

For each simulation, progression response data are generated, over a grid spanning 30 years, for all individuals from individual specific 5PL functions, whose parameters are generated in each simulation according to the distributions of the parameters described previously. To reflect a real-life scenario, uniformly distributed noise, whose range depends on the part of the progression curve from which the data originates, is added to each response value. In AD (and other diseases that also follow a sigmoidal trajectory), there is generally more variation in measurements around the years of most rapid change in progression compared with the beginning and final stages of disease. In Figure 2, this rapid change in
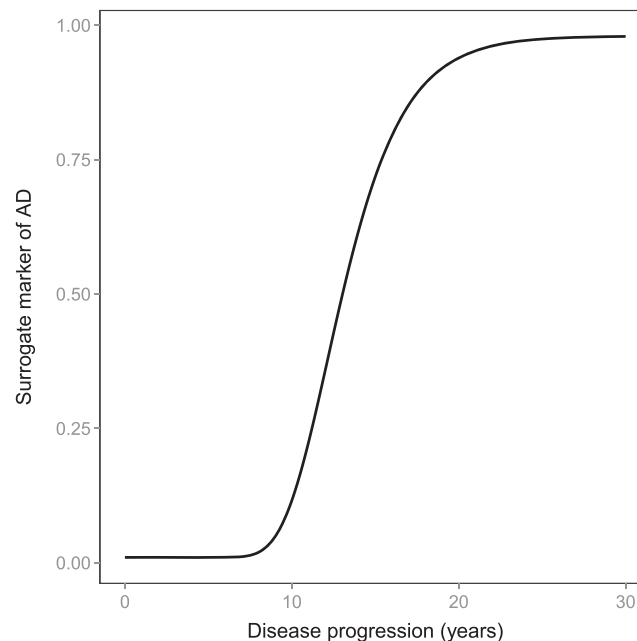

Figure 2. An example of a five-parameter logistic curve modelled over 30 years for a surrogate marker of Alzheimer's disease (AD). The following parameter values are utilised for this figure $b = 13$, $a = 0.4$, $c = 0.01$, $f = 0.98$ and $g = 20$.

progression is seen between 10 and 20 years, and the distribution of the uniformly distributed noise reflects this. This provides us with *n* individual sets of data, each spanning 30 years.

As we are interested in the construction of the population trajectory curve from short-term follow-up data, we use for each individual's set of data only the data over *t* years, where *t* takes various values in our simulation settings but is typically less than 5 years. The selection of the start point for the subset of data used for each individual is randomly generated from a mixture of two normal distributions, reflecting the nature of real-world data where there are generally fewer individuals studied around the area of rapid change of progression. Once this starting time is randomly generated for each individual, we extract from that individual's data *m* consecutive observations from the starting point such that these are *l* time units apart and within *t* years. For all subsequent steps of the proposed methodology and our simulations, our knowledge is restricted to the study time point for an individual, and we assume that every individual's first observation is their first study time point. This reflects the real world scenario where it is unknown what stage of the disease the individual is at, or what part of the curve their data came from.

The first panel of Figure 1 displays a subset of individual data for one simulation, with a regression line fitted for each individual. These regression lines are then used to obtain estimates of the means and slopes for a surrogate marker of AD each individual using methods such as LMs or LMMs. The second panel displays these estimated slopes against the estimated means, overlaid by a fitted (non-negative) polynomial. Finally, the integral of the reciprocated fitted non-negative polynomial was evaluated (third panel), and the resulting function was inverted to produce an estimated disease progression curve (fourth panel).

5. Simulation results

In this section, we perform a variety of simulations to further investigate the properties of our proposed methodology for constructing the underlying longitudinal trajectory of AD progression. In Section 5.2, we look at comparing methodology for estimating means and slopes and varying the degree of polynomial used. In Section 5.3, we investigate the impact of changing the short-term follow up length, number of consecutive observations per individual and consequently their time between consecutive observations.

5.1. A simple example

Initially, we describe results for a specific scenario over a 5-year collection period, with $n = 1000$ individuals, $l = 0.5$ years between observations resulting in each individual having 11 consecutive time points. Data from this scenario were generated 1000 times yielding 1000 respective estimated prediction curves. For comparison purposes, we generated a population average curve using Monte Carlo simulations based on 1,000,000 individual simulated curves and the sampled parameters described in Section 4 and compared each estimated prediction curve to this. The left panel of Figure 3 shows the fitted regression lines for a subset of individuals from one simulated dataset, the estimated prediction curve based on this dataset and the population average curve. The accuracy of the four-step approach is evident from the small differences seen between the population average and estimated prediction curve. The middle panel of Figure 3 displays the realigned short-term trajectories of the simulated individual's data where every individual's first observation is now their starting point. We also provide 95% CIs based on the bootstrapping techniques described in Section 3.1, which can be seen in the right panel of Figure 3. Note that these were constructed based on 100 individuals, as the CIs become very narrow with a larger number of individuals in our simulations.

5.2. Model comparisons

Using the same distributions for the model parameters as described previously, we compare four different methods obtained by crossing the method of estimating the means and slopes, by either LMs or LMMs, with the degree of the non-negative polynomial used, either cubic (3) or quintic (5). For the rest of this paper, we refer to these methods as LM3, LM5, LMM3 and LMM5.

To examine the difference between the population average and estimated prediction curve, the area between these curves is calculated over a fine grid using the trapezoidal rule. Figure 4 provides boxplots for the area between the population average curve and the 1000 estimated predicted curves for the four methods described previously. We observe more variability in the area between the population average curve and the estimated predicted curves for the quintic models in comparison with the cubic models, with the interquartile range spanning a wider range for these quintic combinations. There is also evidence
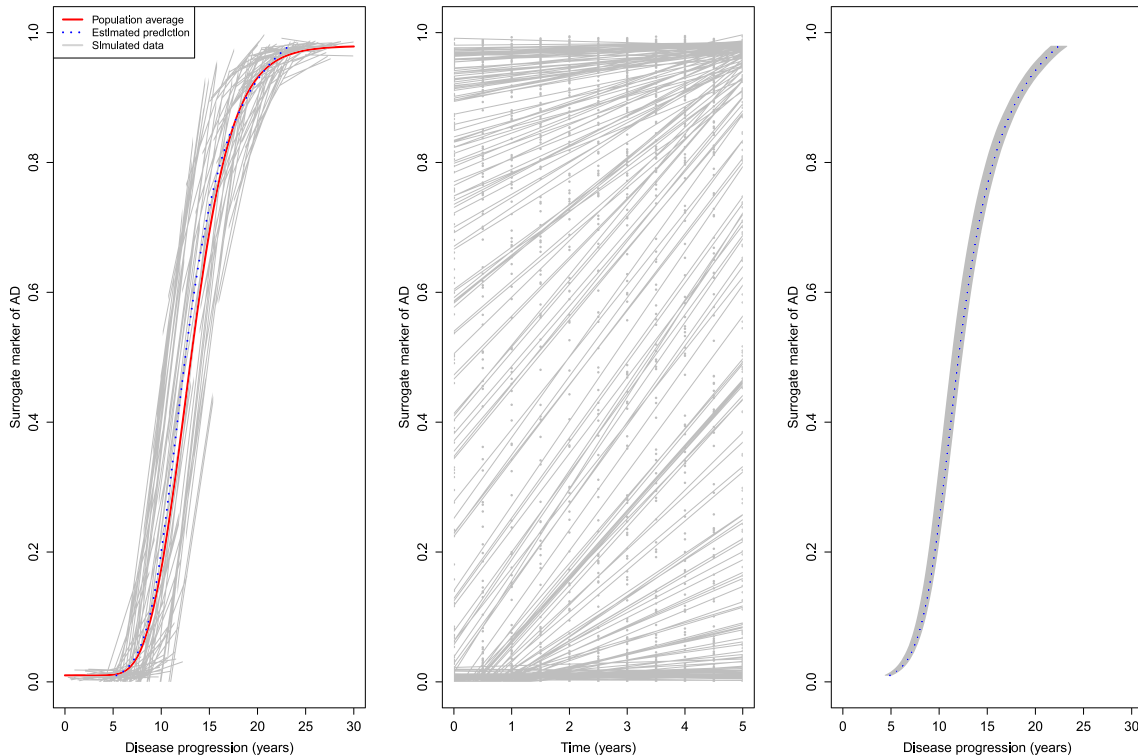
Figure 3. Left panel: the results of one simulation following our proposed four-step approach showing individual short-term data (grey), the estimated predicted curve (blue) and the population average curve (red). Middle panel: individual short-term data rescaled where every individuals first observation is their starting point. Right panel: the estimated predicted curve and 95% CI after conducting 1000 bootstraps for each model based on 100 individuals. AD, Alzheimer's disease.
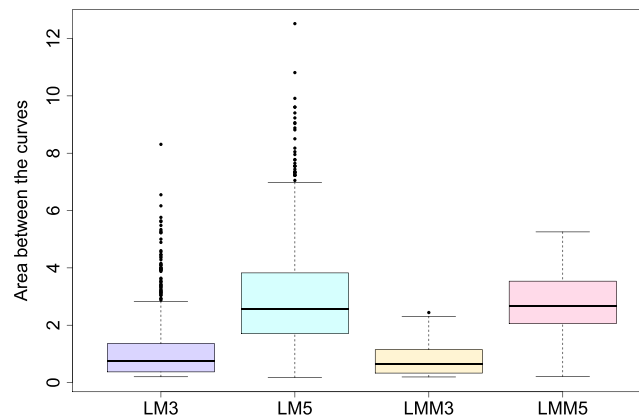

Figure 4. Boxplots displaying the range of values for the area between the population average and estimated prediction curves for the four models. [Colour figure can be viewed at wileyonlinelibrary.com]

of extreme values and a long right tail in the distribution of the areas between the curves with the LM combinations, which is not apparent in the LMM combinations. These extreme values demonstrate a tendency for an inaccurate construction of the underlying disease progression curve for the LM combinations.

Overall, the LMM with a fitted cubic polynomial provides the smallest area between the curves, with minimal difference evident between the LMs and LMMs within cubics and within quintics.

5.3. Parameter variations

In addition to comparing the different model combinations, variations in parameters are also investigated. These include varying the time interval between consecutive observations, l, the number of

individual consecutive observations, m, and consequently the total follow up time, t, which is defined by the previous two parameters. For each variation, 1000 simulations are conducted. In this section, we use $l = 0.25, 0.5, 0.75, \dots, 3$ years and $m = 3, 4, 5, \dots, 16$ consecutive observations. Only combinations of these parameters that had a maximum follow-up time of 14 years were investigated. These variations are explored in order to provide advice on the optimal sampling methodology for constructing sigmoidal progression curves based on our proposed four-step approach.

Figure 5 displays line graphs for the area between the population average and the estimated predicted curves comparing the four different models, LM3, LM5, LMM3 and LMM5, separately for $0.5, 1, 1.5, 2, 2.5$ and 3 years between consecutive observations. In the top left panel, where the time between consecutive observations is half a year, we observe similar results for the cubic model combinations with smaller areas between the curves compared with the quintic model combinations. When the time between consecutive observations is increased, the results become less consistent, but generally, the cubic variations perform better than the quintic models with the latter displaying a great deal of variation in the area between the curves over time. Arguably, the best overall combination is the cubic LMM.

The heat maps in Figure 6 show an alternative way of comparing the different combinations of the parameter variations for LM3, LM5, LMM3 and LMM5. Each plot displays the area between the curves for the number of consecutive observations against the time between consecutive observations. The areas of darker colour represent the better performing models, with smaller areas between the population average and the estimated prediction curves.

In general, we can see that the cubic models seem to perform better overall, with increased areas of darker shading compared with the quintic variations. However, within each of the combinations, we can see some obvious patterns. In the bottom left corner of each plot, we see an increased discrepancy between the curves with evidence of lighter shading, representing shorter times between consecutive observations, and fewer number of observations, subsequently resulting in a shorter follow-up period and consequently
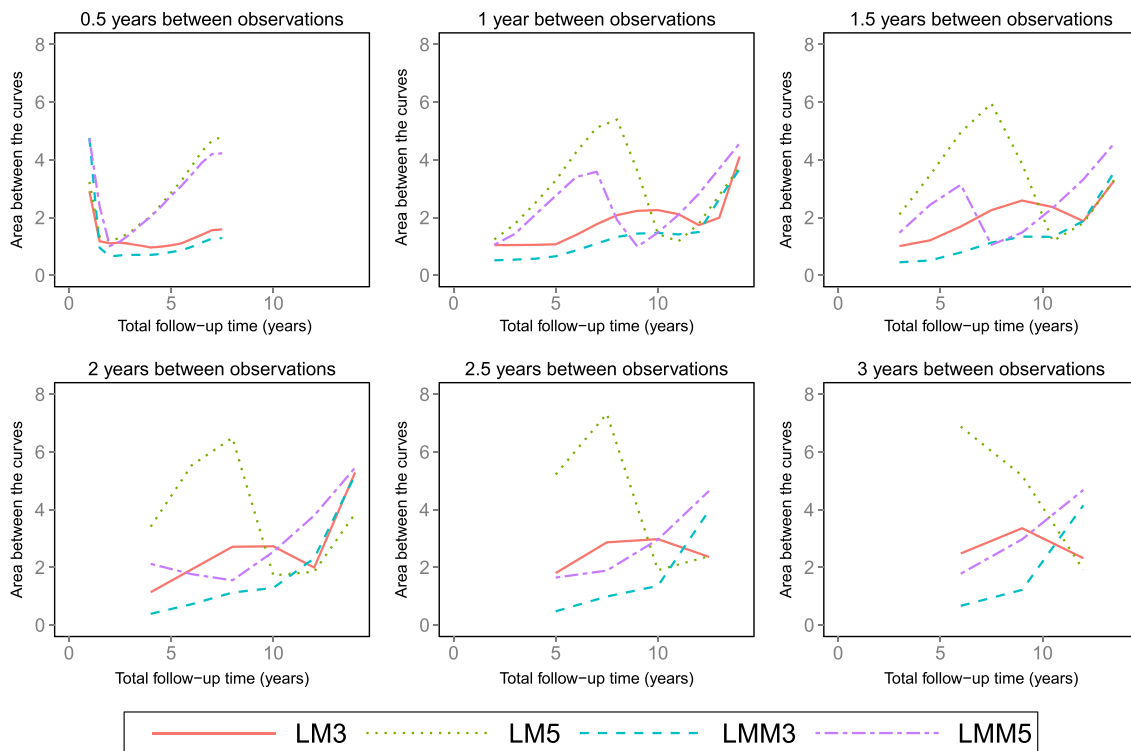

Figure 5. A comparison in the area between the population average curve and estimated prediction curves. Each plot displays a different time interval between consecutive observations against total follow-up time. Top left panel uses $l = 0.5$ and $m = 3, 4, 5 \dots, 16$, with t between 1 and 7.5 years. Top middle panel uses $l = 1$ and $m = 3, 4, 5 \dots, 15$, with t between 2 and 14 years. Top right panel uses $l = 1.5$ and $m = 3, 4, 5 \dots, 10$, with t between 3 and 13.5 years. Bottom left panel uses $l = 2$ and $m = 3, 4, 5 \dots, 8$, with t between 4 and 14 years. Bottom middle panel uses $l = 2.5$ and $m = 3, 4, 5, 6$, with t between 5 and 12.5 years. Bottom right panel uses $l = 3$ and $m = 3, 4, 5$, with t between 6 and 12 years.
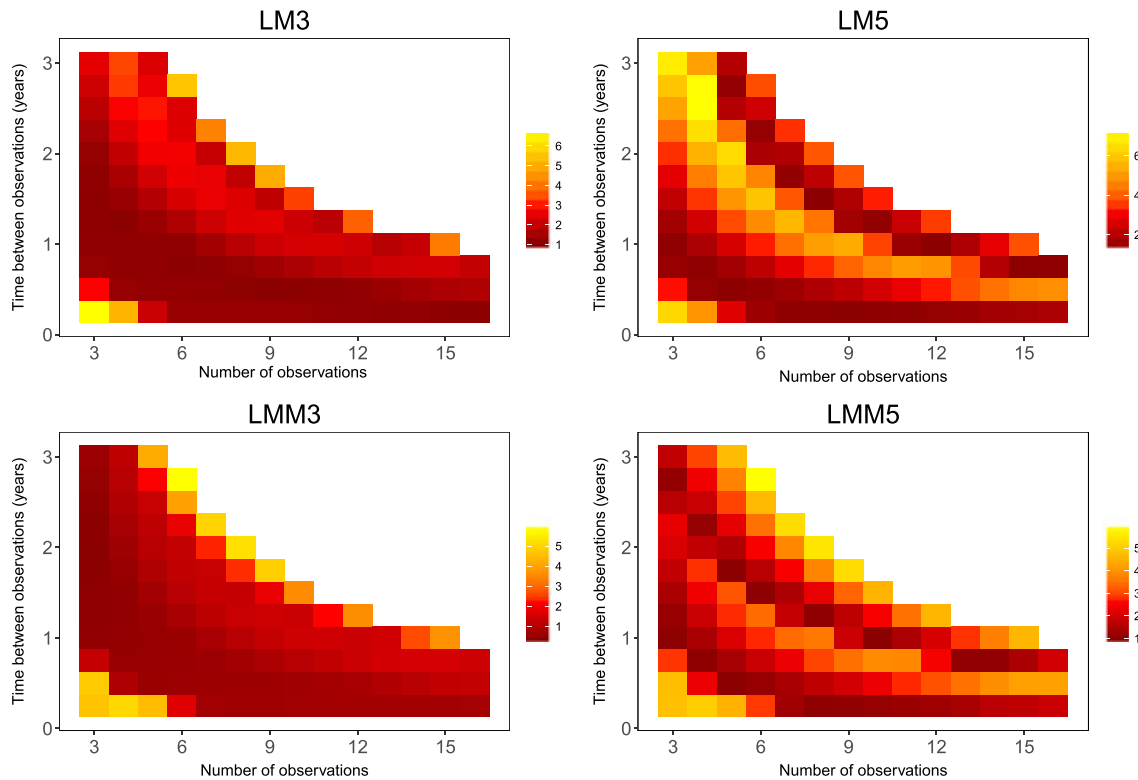
Figure 6. Heat maps of the area between the population average curve and estimated predicted curve when varying the number of consecutive observations and time between these observations for each model. Darker areas indicate better performing variations.

poorer outcomes. For the two cubic models, a gradient along the diagonal can be seen such that there is an initial improvement in the performance for all settings orthogonal to this gradient and later deterioration. For the quintic models, this behaviour is less pronounced. Specifically, with respect to follow-up time and the cubic models, it appears that most combinations that lead to a total follow up time of about 3–6 years appears to perform well. Overall, the combination that provides the smallest area between the curves was the LMM3 variation, using an interval of 2 years between consecutive observations and three observations in total, therefore a total follow-up time of 4 years. The combination leading to the most difference between the population average and estimated prediction curve is using an interval of 2.5 years and four observations (7.5 years of follow up) for the LM5 combination.

6. Real-world application

To illustrate our methodology, we apply our four-step approach to subjects that were recruited through ADNI, an ongoing, longitudinal, multi-centre study launched in 2003 to develop clinical, imaging, genetic and biochemical biomarkers for the early detection and tracking of AD [6]. Data used in the preparation of this article were obtained from ADNI's database (adni.loni.usc.edu). ADNI was launched in 2003 as a public–private partnership, led by Principal Investigator Michael W. Weiner, MD. The primary goal of ADNI has been to test whether serial magnetic resonance imaging, PET, other biological markers, and clinical and neuropsychological assessment can be combined to measure the progression of MCI and early AD. For up-to-date information, see www.adni-info.org

For the purpose of demonstrating our approach, we used neocortical $A\beta$ burden as measured by the standard uptake value ratio (SUVR) from ^{18}F-florbetapir PET imaging normalised to the whole cerebellum. Our proposed method requires longitudinal data and subjects having at least three observations; thus, individuals with only one or two observations were not considered. The population included 156 ADNI subjects all having three longitudinal measurements. To apply our methodology to ADNI data, an individual's first $A\beta$ measurement was considered a baseline measurement, and subsequent longitudinal times were calculated from this.

Table I. Demographics (mean (SD), n(%)) of the ADNI data for each diagnosis group.

	HC ($N = 48$)	MCI ($N = 102$)	AD ($N = 6$)
Baseline age (years)	75.1 ± 5.4	70.0 ± 7.8	71.1 ± 9.7
Sex (male)	27 (56%)	55 (54%)	2 (33%)
Baseline neocortical Aβ burden	0.78 ± 0.11	0.85 ± 0.12	0.95 ± 0.18
E4 carrier (yes)	9 (19%)	40 (39%)	3 (50%)
Education (years)	16.4 ± 2.8	16.0 ± 2.8	15.3 ± 3.7
Baseline MMSE	29.2 ± 1.3	28.1 ± 1.8	21.2 ± 4.4

ADNI, Alzheimer's Disease Neuroimaging Initiative; HC, healthy control; MCI, mild cognitively impaired; AD, Alzheimer's disease; MMSE, mini-mental state exam.
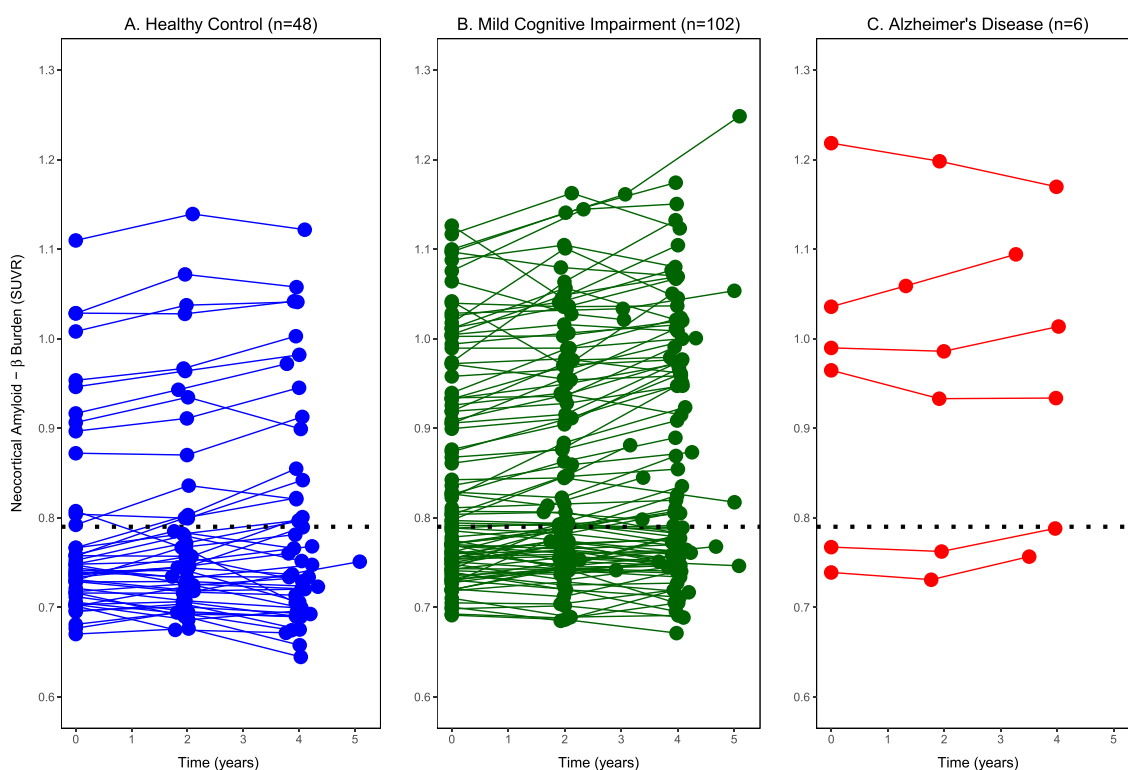
Figure 7. Individual longitudinal neocortical Aβ burden (standard uptake value ratio (SUVR)) by diagnosis group: left panel displays the healthy controls; middle panel displays the mild cognitively impaired; and right panel displays individuals with Alzheimer's disease. The horizontal dotted line at 0.79 represents the cut-off for individuals classified as amyloid β positive. [Colour figure can be viewed at wileyonlinelibrary.com]

We first provide a summary of the ADNI cohort used in this study but note that categorisation of subjects into diagnostic groups (HC, MCI and AD), or use of subject-specific demographics is not necessary to implement our four-step modelling approach. In our cohort, at the defined baseline, of the 156 subjects, 48 were HC, 102 MCI and 6 were diagnosed as having AD. Demographic details split by diagnosis group are provided in Table I. At baseline, subjects diagnosed with AD had a significantly higher Aβ than both the MCI and the HC subjects (estimated mean difference 0.101 (95% CI 0.001–0.202) and 0.167 (95% CI 0.064–0.271), respectively). The MCI subjects had a significantly higher Aβ than HC subjects (estimated mean difference 0.066 (95% CI 0.024–0.108)). At the first, follow-up visit after baseline 100 (64%) of subjects showed positive changes in neocortical Aβ burden.

Figure 7 displays the trajectories of Aβ for all study individuals in our cohort, and Figure 8 displays the disease progression curve for AD after applying the four-step approach described in Section 3 to these data. For the latter figure, linear mixed modelling was used to estimate the means and slopes of all individuals and a cubic polynomial was fitted to these data; see Figure A1 in Appendix A for the progression curve when linear models were used. Figure 7 also displays in each panel a horizontal dotted
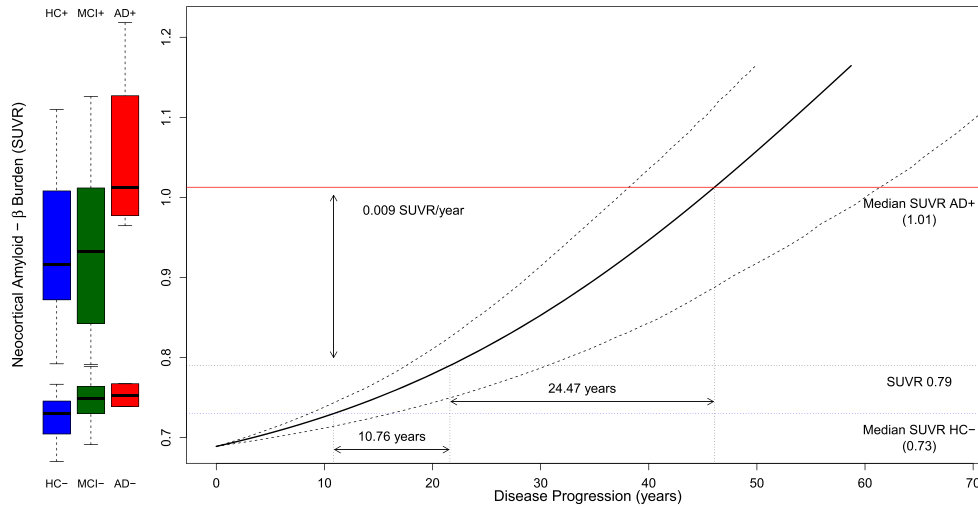
Figure 8. Application of our approach using linear mixed models to Alzheimer's Disease Neuroimaging Initiative data with bootstrapped 95% CI including boxplots of neocortical Aβ burden (standard uptake value ratio (SUVR)) for each diagnosis group, separately for amyloid–β positive and negative individuals. It takes 24.47 years to progress from an SUVR of 0.79 to 1.01. This is equivalent to a rate of 0.009 increase in SUVR per year. Similarly, it takes 10.76 years to progress from an SUVR of 0.73 to 0.79. See the text for further details. HC, healthy control; MCI, mild cognitively impaired; AD, Alzheimer's disease. [Colour figure can be viewed at wileyonlinelibrary.com]

line at the commonly used clinical threshold of 0.79, which is based on the estimated 95% quantile for the distribution of Aβ in a group of young HCs for use in longitudinal data [33, 34]. Individuals with an SUVR above or equal to this value are defined as Aβ positive, and below are Aβ negative.

In Figure 8, it is of interest to estimate the time it takes to move from HC- (the median SUVR for HCs that are Aβ negative in our cohort) to this clinical threshold and from this threshold to AD+ (the median SUVR for ADs that are Aβ positive in our cohort). The results indicate that it takes 24.47 years (95% CI 20.17–32.15) to progress from the SUVR clinical threshold of 0.79 to a value of 1.01 (the estimated median SUVR of the AD+ individuals in our dataset). This is equivalent to a rate of 0.009 increase in SUVR per year (95% CI 0.007–0.011). Similarly, it takes 10.76 years (95% CI 8.73–14.70) to progress from 0.73 (the estimated median SUVR of the HC- individuals in our dataset) to the SUVR clinical threshold of 0.79.

7. Discussion

Alzheimer's disease is a complex disease, and it is currently rare to have data that encompasses the full course of the disease, with only short-term follow-up data typically being available. In this paper, we have demonstrated that it is possible to construct the full underlying longitudinal trajectories using such short-term data from multiple individuals. We used simulated data generated from a 5PL function, possessing a shape similar to that hypothesised for AD progression [9], to explore variations in our approach.

Our investigations suggest that the use of cubic polynomials outperforms quintic polynomials when constructing the full progression curve and the fitting of non-negative polynomials is essential. The use of a more flexible quintic polynomial could potentially increase the variability observed because of over-fitting of the data. There were minimal differences between the linear regression models and linear mixed regression models when constructing the full progression curve, with a marginal preference for LMMs. Note that in order to fit LMMs, an adequate number of follow-up observations for the individuals are necessary. Finally, for a hypothesised 30-year progression curve, we demonstrated that the best range of follow-up times for our methodology is between 3 and 6 years for a fixed number of visits, with follow-up times outside this range providing less accurate construction of the longitudinal trajectory.

We applied our approach to PET data measuring neocortical Aβ burden (SUVR) to ADNI, where individuals have between 3 and 5 years of follow-up for this specific measure. Our method was able to quantify the time it takes to progress through the disease course, with results similar to those reported in Villemagne *et al*. [8] and Bateman *et al*. [13]. This quantification could promote understanding of the

relationship between AD biomarkers and disease progression and is crucial for the introduction of early therapeutic or intervention strategies to assist in delaying the onset of clinical symptoms and slowing cognitive decline.

Using bootstrap methodology, we were also able to provide CIs for our predictions. We have described and implemented the cases bootstrap, which has been advocated in many previous works and is known to be robust against violations of the distributional assumptions of the random effects. However, we recognise different bootstrap methodologies for longitudinal and mixed effects models, and their effectiveness is still being studied.

Using our methodology, it is less likely to experience problems associated with missing data, mainly because we utilise only short-term follow-up data. Individuals' data are used to estimate the means and slopes based on data available from the time they entered the study and as such an individual missing a measurement should not have an undue impact on the ability to provide such estimates. Furthermore, those individuals who may have a smaller number of measurements due to death could still make a useful contribution as it is expected these individuals would contribute higher means and lower slopes and provide information on the upper asymptote of the sigmoidal disease progression curve. As such, drop out due to death during the duration of the data collection period should have minimal impact on the ability to construct the disease progression curve. However, whilst shorter follow-ups do reduce the chance of missing data, there is still the chance of some selection bias. For example, fast progressors may possibly die earlier, which potentially impacts the estimation of the population curve later in disease. With the use of measures such as neocortical Aβ burden, we suspect this to have little impact, but it should be considered when investigating other measures, for example, MMSE.

In future work, we expect to explore the effectiveness of our methodology for different markers of AD such as, MMSE, Clinical Dementia Rating Scale or hippocampal volume and will also be able to provide separate progression paths for covariates of interest and interactions between these covariates and time, which will consequently assist in understanding the staging of different events.

The implementation of all methods described in this paper was carried out using R: A Language and Environment for Statistical Computing [35], and code will be made available upon request.
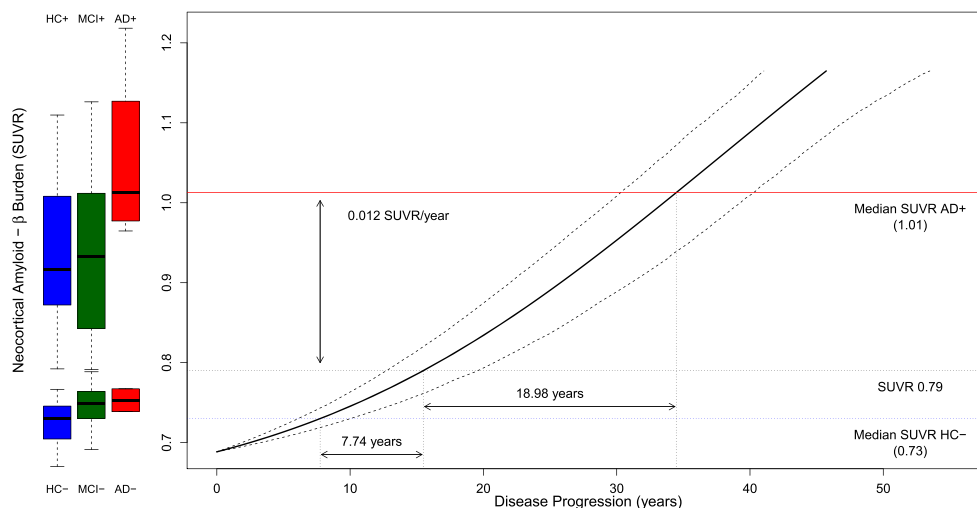
Appendix A

Figure A1. Application of our approach using linear models to Alzheimer's Disease Neuroimaging Initiative data with bootstrapped 95% CI including boxplots of neocortical Aβ burden (standard uptake value ratio (SUVR)) for each diagnosis group, separately for amyloid–β positive and negative individuals. It takes 18.98 (95% CI 16.55–22.19) years to progress from an SUVR of 0.79 to 1.01. This is equivalent to a rate of 0.012 (95% CI 0.010–0.0135) increase in SUVR per year. Similarly, it takes 7.74 (95% CI 6.53–9.58) years to progress from an SUVR of 0.73 to 0.79. HC, healthy control; MCI, mild cognitively impaired; AD, Alzheimer's disease. [Colour figure can be viewed at wileyonlinelibrary.com]

Acknowledgements

References

1. Prince MJ. *World Alzheimer Report 2015: The Global Impact of Dementia: An Analysis of Prevalence, Incidence, Cost and Trends*, 2015. Available from: https://www.alz.co.uk/research/world-report-2015 [accessed 14th November 2016].
2. Selkoe DJ. Alzheimer's disease: genes, proteins, and therapy. *Physiological Reviews* 2001; **81**(2):741–766.
3. Masters CL, Beyreuther K. The neuropathology of Alzheimer's disease in the year 2005. In *Neurodegenerative Diseases: Neurobiology, Pathogenesis and Therapeutics*. Cambridge University Press, 2005; 433–440, Cambridge Books Online.
4. Masters CL, Cappai R, Barnham KJ, Villemagne VL. Molecular mechanisms for Alzheimer's disease: implications for neuroimaging and therapeutics. *Journal of Neurochemistry* 2006; **97**(6):1700–1725.
5. Dubois B, Feldman HH, Jacova C, Dekosky ST, Barberger-Gateau P, Cummings J, Delacourte A, Galasko D, Gauthier S, Jicha G, Meguro K, O'brien J, Pasquier F, Robert P, Rossor M, Salloway S, Stern Y, Visser PJ, Scheltens P. Research criteria for the diagnosis of Alzheimer's disease: revising the NINCDS–ADRDA criteria. *The Lancet Neurology* 2007; **6**(8):734–746.
6. Weiner MW, Veitch DP, Aisen PS, Beckett LA, Cairns NJ, Green RC, Harvey D, Jack CR, Jagust W, Liu E, Morris JC, Petersen RC, Saykin AJ, Schmidt ME, Shaw L, Shen L, Siuciak JA, Soares H, Toga AW, Trojanowski JQ, Alzheimer's Disease Neuroimaging Initiative. The Alzheimer's Disease Neuroimaging Initiative: a review of papers published since its inception. *Alzheimer's & Dementia* 2013; **9**(5):e111–e194.
7. Ellis KA, Bush AI, Darby D, De Fazio D, Foster J, Hudson P, Lautenschlager NT, Lenzo N, Martins RN, Maruff P, Masters C, Milner A, Pike K, Rowe C, Savage G, Szoeke C, Taddei K, Villemagne V, Woodward M, Ames D, AIBL Research Group. The Australian Imaging, Biomarkers and Lifestyle (AIBL) study of aging: methodology and baseline characteristics of 1112 individuals recruited for a longitudinal study of Alzheimer's disease. *International Psychogeriatrics* 2009; **21**(04):672–687.
8. Villemagne VL, Burnham S, Bourgeat P, Brown B, Ellis KA, Salvado O, Szoeke C, Macaulay SL, Martins R, Maruff P, Ames D, Rowe CC, Masters CL, Australian Imaging Biomarkers and Lifestyle (AIBL) Research Group. Amyloid β deposition, neurodegeneration, and cognitive decline in sporadic Alzheimer's disease: a prospective cohort study. *The Lancet Neurology* 2013; **12**(4):357–367.
9. Jack CR, Knopman DS, Jagust WJ, Shaw LM, Aisen PS, Weiner MW, Petersen RC, Trojanowski JQ. Hypothetical model of dynamic biomarkers of the Alzheimer's pathological cascade. *The Lancet Neurology* 2010; **9**(1):119–128.
10. Jack CR Jr, Knopman DS, Jagust WJ, Petersen RC, Weiner MW, Aisen PS, Shaw LM, Vemuri P, Wiste HJ, Weigand SD, Lesnick TG, Pankratz VS, Donohue MC, Trojanowski JQ. Tracking pathophysiological processes in Alzheimer's disease: an updated hypothetical model of dynamic biomarkers. *The Lancet Neurology* 2013; **12**(2):207–216.
11. Huijbers W, Mormino EC, Schultz AP, Wigman S, Ward AM, Larvie M, Amariglio RE, Marshall GA, Rentz DM, Johnson KA, Sperling RA. Amyloid-β deposition in mild cognitive impairment is associated with increased hippocampal activity, atrophy and clinical progression. *Brain* 2015; **138**(4):1023–1035. awv007.
12. Yau WY, Tudorascu DL, McDade EM, Ikonomovic S, James JA, Minhas D, Mowrey W, Sheu LK, Snitz BE, Weissfeld L, Gianaros PJ, Aizenstein HJ, Price JC, Mathis CA, Lopez OL, Klunk WE. Longitudinal assessment of neuroimaging and clinical markers in autosomal dominant Alzheimer's disease: a prospective cohort study. *The Lancet Neurology* 2015; **14**(8):804–813.
13. Bateman RJ, Xiong C, Benzinger TLS, Fagan AM, Goate A, Fox NC, Marcus DS, Cairns NJ, Xie X, Blazey TM. Clinical and biomarker changes in dominantly inherited Alzheimer's disease. *New England Journal of Medicine* 2012; **367**(9): 795–804.
14. Samtani MN, Raghavan N, Novak G, Nandy P, Narayan VA. Disease progression model for clinical dementia rating–sum

of boxes in mild cognitive impairment and Alzheimer's subjects from the Alzheimer's Disease Neuroimaging Initiative. *Neuropsychiatric Disease and Treatment* 2014; **10**:929–952.

15. Capuano AW, Wilson RS, Leurgans SE, Dawson JD, Bennett DA, Hedeker D. Sigmoidal mixed models for longitudinal data. *Statistical Methods in Medical Research* 2016. In press.

16. Sabuncu MR, Desikan RS, Sepulcre J, Yeo BT, Liu H, Schmansky NJ, Reuter M, Weiner MW, Buckner RL, Sperling RA, Fischl B, Alzheimer's Disease Neuroimaging Initiative. The dynamics of cortical and hippocampal atrophy in Alzheimer disease. *Archives of Neurology* 2011; **68**(8):1040–1048.

17. Yang E, Farnum M, Lobanov V, Schultz T, Raghavan N, Samtani M.N, Novak G, Narayan V, DiBernardo A. Quantifying the pathophysiological timeline of Alzheimer's disease. *Journal of Alzheimer's Disease* 2011; **26**(4):745–753.

18. Jack CR Jr, Vemuri P, Wiste HJ, Weigand SD, Lesnick TG, Lowe V, Kantarci K, Bernstein MA, Senjem ML, Gunter JL, Boeve BF, Trojanowski JQ, Shaw LM, Aisen PS, Weiner MW, Petersen RC, Knopman DS, for the Alzheimers Disease Neuroimaging Initiative. Shapes of the trajectories of 5 major biomarkers of Alzheimer disease. *Archives of Neurology* 2012; **69**(7):856–867.

19. Caroli A, Frisoni GB, Initiative ADN. The dynamics of Alzheimer's disease biomarkers in the Alzheimer's disease neuroimaging initiative cohort. *Neurobiology of Aging* 2010; **31**(8):1263–1274.

20. Aisen PS, Petersen RC, Donohue MC, Gamst A, Raman R, Thomas RG, Walter S, Trojanowski JQ, Shaw LM, Beckett LA, Jack CR Jr, Jagust W, Toga AW, Saykin AJ, Morris JC, Green RC, Weiner MW, Alzheimer's Disease Neuroimaging Initiative. Clinical core of the Alzheimer's disease neuroimaging initiative: progress and plans. *Alzheimer's & Dementia* 2010; **6**(3):239–246.

21. Verbeke G, Molenberghs G. *Linear Mixed Models for Longitudinal Data*, Springer Series in Statistics. Springer-Verlag: New York, 2009.

22. Fitzmaurice GM, Laird NM, Ware JH. *Applied Longitudinal Analysis*, Wiley Series in Probability and Statistics. John Wiley & Sons: Hoboken, 2011.

23. Brickman L. On nonnegative polynomials. *The American Mathematical Monthly* 1962; **69**(3):218–221.

24. Murray K, Müller S, Turlach BA. Fast and flexible methods for monotone polynomial fitting. *Journal of Statistical Computation and Simulation* 2016; **86**(15):2946–2966.

25. De Leeuw J, Meijer E, Goldstein H. *Handbook of Multilevel Analysis*. Springer-Verlag: New York, 2008.

26. Field CA, Welsh AH. Bootstrapping clustered data. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)* 2007; **69**(3):369–390.

27. Field CA, Pang Z, Welsh AH. Bootstrapping robust estimates for clustered data. *Journal of the American Statistical Association* 2010; **105**(492):1606–1616.

28. Sherman M, le Cessie S. A comparison between bootstrap methods and generalized estimating equations for correlated outcomes in generalized linear models. *Communications in Statistics - Simulation and Computation* 1997; **26**(3):901–925.

29. Chambers R, Chandra H. A random effect block bootstrap for clustered data. *Journal of Computational and Graphical Statistics* 2013; **22**(2):452–470.

30. Thai H-T, Mentré F, Holford NHG, Veyrat-Follet C, Comets E. A comparison of bootstrap approaches for estimating uncertainty of parameters in linear mixed-effects models. *Pharmaceutical statistics* 2013; **12**(3):129–140.

31. Efron B, Tibshirani RJ. *An Introduction to the Bootstrap*, Chapman & Hall/CRC Monographs on Statistics & Applied Probability. Taylor & Francis: New York, 1994.

32. Gottschalk PG, Dunn JR. The five-parameter logistic: a characterization and comparison with the four-parameter logistic. *Analytical Biochemistry* 2005; **343**(1):54–65.

33. Joshi AD, Pontecorvo MJ, Clark CM, Carpenter AP, Jennings DL, Sadowsky CH, Adler LP, Kovnat KD, Seibyl JP, Arora A, Saha K, Burns JD, Lowrey MJ, Mintun MA, Skovronsky DM, Florbetapir F 18 Study Investigators. Performance characteristics of amyloid pet with florbetapir F 18 in patients with Alzheimer's disease and cognitively normal subjects. *Journal of Nuclear Medicine* 2012; **53**(3):378–384.

34. Landau S, Jagust W. *Florbetapir Processing Methods*, 2011. http://adni.bitbucket.org/docs/UCBERKELEYAV45/ UCBERKELEY_AV45_Methods_12.03.15.pdf [accessed 14th November 2016].

35. R Core Team. *R: A language and environment for statistical computing*. R Foundation for Statistical Computing: Vienna, Austria, 2016. https://www.R-project.org/.