TECHNICAL NOTE

Deep learning for clustering of multivariate clinical patient trajectories with missing values

Johann de Jong [1,*], Mohammad Asif Emon[2,3], Ping Wu[4], Reagon Karki[2,3], Meemansa Sood[2,3], Patrice Godard[5], Ashar Ahmad[3], Henri Vrooman[6,7], Martin Hofmann-Apitius[2,3] and Holger Fröhlich [1,3,*]

[1]UCB Biosciences GmbH, Alfred-Nobel-Strasse 10, 40789 Monheim, Germany; [2]Fraunhofer Institute for Algorithms and Scientific Computing, Schloss Birlinghoven, Konrad-Adenauer-Strasse, 53754 Sankt Augustin, Germany; [3]Bonn-Aachen International Center for IT, University of Bonn, Konrad-Adenauer-Strasse, 53115 Bonn, Germany; [4]UCB Pharma, Bath Road 216, Slough SL1 3WE, UK; [5]UCB Pharma, Chemin du Foriest 1, 1420 Braine-l'Alleud, Belgium; [6]Erasmus MC, University Medical Center Rotterdam, Department of Radiology, Doctor Molewaterplein 40, PO Box 2040, 3000 CA Rotterdam, Netherlands and [7]Erasmus MC, University Medical Center Rotterdam, Doctor Molewaterplein 40, Department of Medical Informatics, PO Box 2040, 3000 CA Rotterdam, Netherlands

*Correspondence address. Johann de Jong, UCB Biosciences GmbH, Alfred-Nobel-Strasse 10, 40789 Monheim, Germany. E-mail: johann.dejong@ucb.com http://orcid.org/0000-0002-2097-0915; Holger Fröhlich, Bonn-Aachen International Center for IT, University of Bonn, Konrad-Adenauer-Strasse, 53115 Bonn, Germany, E-mail: frohlich@bit.uni-bonn.de http://orcid.org/0000-0002-5328-1243

Abstract

Background: Precision medicine requires a stratification of patients by disease presentation that is sufficiently informative to allow for selecting treatments on a per-patient basis. For many diseases, such as neurological disorders, this stratification problem translates into a complex problem of clustering multivariate and relatively short time series because (i) these diseases are multifactorial and not well described by single clinical outcome variables and (ii) disease progression needs to be monitored over time. Additionally, clinical data often additionally are hindered by the presence of many missing values, further complicating any clustering attempts. **Findings:** The problem of clustering multivariate short time series with many missing values is generally not well addressed in the literature. In this work, we propose a deep learning–based method to address this issue, variational deep embedding with recurrence (VaDER). VaDER relies on a Gaussian mixture variational autoencoder framework, which is further extended to (i) model multivariate time series and (ii) directly deal with missing values. We validated VaDER by accurately recovering clusters from simulated and benchmark data with known ground truth clustering, while varying the degree of missingness. We then used VaDER to successfully stratify patients with Alzheimer disease and patients with Parkinson disease into subgroups characterized by clinically divergent disease progression profiles. Additional analyses demonstrated that these clinical differences reflected known underlying aspects of Alzheimer disease and Parkinson disease. **Conclusions:** We believe our results show that VaDER can be of great value for future efforts in patient stratification, and multivariate time-series clustering in general.

Keywords: patient stratification; deep learning; multivariate time series; multivariate longitudinal data; clustering

Findings

Background

In precision medicine, patients are stratified on the basis of their disease subtype, risk, prognosis, or treatment response by means of specialized diagnostic tests. An important question in precision medicine is how to appropriately model disease progression and accordingly decide on the right type and time point of therapy for an individual. However, the progression of many diseases, such as neurological disorders, cardiovascular diseases, diabetes, and obesity [1–5], is highly multifaceted and not well described by 1 clinical outcome measure alone. Classical univariate clustering methods are likely to miss the inherent complexity of diseases that demonstrate a highly multifaceted clinical phenotype. Accordingly, stratification of patients by disease progression translates into the challenging question of how to identify a clustering of a multivariate time series.

Clustering is a fundamental and generally well-investigated problem in machine learning and statistics. Its goal is to segment samples into groups (clusters), such that there is a higher degree of similarity between samples of the same cluster than between samples of different clusters. Following Hastie et al. [6], algorithms to solve clustering problems may be put into 3 main categories, (i) combinatorial algorithms, (ii) mixture modeling, and (iii) mode seeking. Within each of these 3 categories, a wide range of methods is available for a great diversity of clustering problems. Combinatorial algorithms do not assume any underlying probability model but work with the data directly. Examples are K-means clustering, spectral clustering [7], and hierarchical clustering [8]. Mixture models assume that the data can be described by some probabilistic model. An example is Gaussian mixture model clustering. Finally, in mode seeking one tries to directly estimate modes of the underlying multi-modal probability density. An important example here is the mean-shift algorithm [9].

For the clustering of multivariate time-series data, a few techniques have been developed [10–14]. However, these approaches generally rely on time series of far greater length than available in most longitudinal clinical datasets. Moreover, these methods are not suited for the large numbers of missing values often found in clinical data.

Missing values in clinical data can occur for different reasons: (i) patients drop out of a study, e.g., owing to worsening of symptoms; (ii) a certain diagnostic test is not taken at a particular visit (e.g., owing to lack of patient agreement), potentially resulting into missing information for entire variable groups; (iii) unclear further reasons, e.g., time constraints, data quality issues, etc. From a statistical point of view, these reasons manifest into different mechanisms of missing data [15,16]:

(1) Missing completely at random (MCAR): The probability of missing information is related neither to the specific value that is supposed to be obtained nor to other observed data. Hence, entire patient records could be skipped without introducing any bias. However, this type of missing data mechanism is probably rare in clinical studies.

(2) Missing at random (MAR): The probability of missing information depends on other observed data but is not related to the specific missing value that is expected to be obtained. An example would be patient dropout due to worsening of certain symptoms, which are at the same time recorded during the study.

(3) Missing not at random (MNAR): any reason for missing data that is neither MCAR nor MAR. MNAR is problematic because

the only way to obtain unbiased estimates is to model missing data.

Multiple-imputation methods have been proposed to deal with missing values in longitudinal patient data [16]. However, any imputation method will result in certain errors, and if imputation and clustering are performed separately, these errors will propagate through to the following clustering procedure.

To address the problem of clustering multivariate and relatively short time-series data with many missing values, in this article we propose an approach that uses techniques from deep learning. Autoencoder networks have been highly successful in learning latent representations of data (e.g., [17–20]). Specifically for clustering, autoencoders can be first used to learn a latent representation of a multivariate distribution, to then independently find clusters [21]. More recently, some authors have suggested simultaneously learning latent representations and cluster assignments. Interesting examples are deep embedded clustering [22] and variational deep embedding (VaDE) [23].

Here, we present a new method for clustering multivariate time series with potentially many missing values, variational deep embedding with recurrence (VaDER). VaDER is in part based on VaDE [23], a clustering algorithm based on variational autoencoder principles, with a latent representation forced towards a multivariate Gaussian mixture distribution. Additionally, VaDER (i) integrates 2 long short-term memory (LSTM) networks [24] into its architecture, to allow for the analysis of multivariate time series; and (ii) adopts an approach of implicit imputation and loss reweighting to account for the typically high degree of missingness in clinical data.

After a validation of VaDER via simulation and benchmark studies, we applied the method to the problem of patient stratification in Alzheimer disease (AD) and Parkinson disease (PD), using data from the Alzheimer's Disease Neuroimaging Initiative (ADNI) [25] and the Parkinson's Progression Markers Initiative (PPMI) [26], respectively. AD and PD are multifactorial and highly heterogeneous diseases, in both clinical and biological presentation, as well as in progression [27–30]. For example, PD is characterized by motor symptoms and behavioral changes (e.g., sleeping disorders), as well as cognitive impairment [31]. Cognitive impairment, one of the hallmarks of AD, is not straightforward to assess, because cognition itself is highly multifaceted, and described by, e.g., orientation, speech, and memory. Consequently, in the field of AD, a wide range of tests have been developed to assess different aspects of cognition.

This heterogeneity presents one of the major challenges in understanding these diseases and developing new treatments. As such, better clustering (stratification) of patients by disease presentation could be of great help in improving disease management and designing better clinical trials that specifically focus on treating patients whose disease is rapidly progressing.

Our analyses of the ADNI and PPMI data show that VaDER is highly effective at disentangling multivariate patient trajectories into clinically meaningful patient subgroups.

Results

Variational autoencoders for clustering

Our proposed VaDER method is in part based on VaDE [23], a variational autoencoding clustering algorithm with a multivariate Gaussian mixture prior. In variational autoencoding algorithms, the training objective is to optimize the variational lower bound

on the marginal likelihood of a data point **x** [32]:

$$\mathcal{L}(\mathbf{x}) = \mathbb{E}_{q(\mathbf{z}|\mathbf{x})}[\log(p(\mathbf{x}|\mathbf{z}))] - D_{KL}(q(\mathbf{z}|\mathbf{x})||p(\mathbf{z})). \quad (1)$$

This lower bound can be seen as composed of 2 parts. The first term corresponds to the likelihood of seeing **x** given a latent representation **z**. Its negative is often called the "reconstruction loss," and it forces the algorithm to learn good reconstructions of its input data. The negative of the second term is often called the "latent loss." It is the Kullback-Leibler divergence of the prior $p(\mathbf{z})$ to the variational posterior $q(\mathbf{z}|\mathbf{x})$, and it regularizes the latent representation **z** to lie on a manifold specified by the prior $p(\mathbf{z})$.

In VaDE, this prior is a multivariate Gaussian mixture. Accordingly including a parameter for choosing a cluster c, the variational lower bound can then be written as follows:

$$\mathcal{L}(\mathbf{x}) = \mathbb{E}_{q(\mathbf{z},c|\mathbf{x})}[\log(p(\mathbf{x}|\mathbf{z}))] - D_{KL}(q(\mathbf{z}, c|\mathbf{x})||p(\mathbf{z}, c)). \quad (2)$$

By forcing the latent representation **z** towards a multivariate Gaussian mixture distribution, VaDE has the ability to simultaneously learn latent representations and cluster assignments of its input data. For more details on variational autoencoders and VaDE, we refer the reader to Jiang et al. [23], Kingma and Welling [32], and Doersch [33].

VaDER
VaDER is an autoencoder-based method for clustering multivariate time series with potentially many missing values. For simultaneously learning latent representations and cluster assignments of its input samples, VaDER uses the VaDE latent loss as described above and in Jiang et al. [23].

To model the auto- and cross-correlations in multivariate time-series data, we integrate peephole LSTM networks [24,34] into the autoencoder architecture (Fig. 1).

To deal with missing values, we directly integrate imputation into model training. As outlined in the Background, separating imputation from clustering can potentially introduce bias. To avoid this bias, we here propose an implicit imputation scheme, which is performed within VaDER training. Our approach to imputation bears some similarity to other approaches [35,36]. However, in contrast to Lipton et al. [35], VaDER uses missingness indicators for implicit imputation as an integral part of neural network training. Additionally, in contrast to Manning et al. [36], our method of imputation is also suited for MNAR data, which are often encountered in clinical datasets.

We first define a weighted reconstruction loss on feature and sample level: imputed values are weighted to 0, non-imputed values are weighted to 1. To retain the balance with the latent loss, the resulting reconstruction loss is rescaled to match the original dimensions of the data. More formally, for a mean squared reconstruction loss, let L be the number of samples in our dataset, $\mathbf{x}^l$ a single input sample, and $\hat{\mathbf{x}}^l$ its corresponding reconstructed output ($l \in 1 \ldots L$). $\mathbf{x}^l$ and $\hat{\mathbf{x}}^l$ are matrices $\in \mathbb{R}^{N \times M}$, where N is the number of time points and M is the number of clinical outcome measures (e.g., cognitive assessments) for a particular patient. Then the unweighted mean reconstruction loss is

$$\frac{1}{L} \sum_{l=1}^{L} \sum_{i=1}^{N} \sum_{j=1}^{M} \left(x_{ij}^l - \hat{x}_{ij}^l\right)^2. \quad (3)$$

Now, let $A := \{x_{ij}^l | x_{ij}^l \text{ is missing}\}$, $\mathbf{1}_A(.)$ be the indicator function on set A, and $|A|$ be the cardinality of A. Then, the weighted mean

squared reconstruction loss is:

$$\frac{NM}{|A|} \sum_{l=1}^{L} \sum_{i=1}^{N} \sum_{j=1}^{M} \mathbf{1}_A(x_{ij}^l) \left(x_{ij}^l - \hat{x}_{ij}^l\right)^2. \quad (4)$$

In addition to the weighted reconstruction loss, we adopt an implicit imputation scheme, where imputed values are learned as an integral part of model training. More specifically, Let $\mathbf{x}^l$, N, M, x_{ij}^l, A, and $\mathbf{1}_A(.)$ be defined as above. Also assume that all x_{ij}^l for which $\mathbf{1}_A(x_{ij}^l) = 1$ are initially imputed with arbitrary finite values. Then we add 1 additional layer before the input LSTM (Fig. 1) as follows:

$$\tilde{x}_{ij}^l = x_{ij}^l \times [1 - \mathbf{1}_A(x_{ij}^l)] + b_{ij} \times \mathbf{1}_A(x_{ij}^l). \quad (5)$$

Here, x_{ij}^l is the actual observed (or missing) value of sample l at time points i and assessment j, and $\tilde{x}_{ij}^l$ serves as input to the LSTM. In other words, if x_{ij}^l is missing, then it is replaced by b_{ij} in $\tilde{\mathbf{x}}$. Parameters b_{ij} are trained as an integral part of VaDER using stochastic gradient descent and can be considered (time, assessment)-specific expected values. Note that (i) the initial arbitrary imputation does not influence the eventual clustering and (ii) the implicitly imputed values are weighted to 0 in the reconstruction loss.

VaDER achieves high accuracy on simulated data
As a first step in technically validating VaDER, we simulated data with a known ground truth clustering and assessed how well VaDER was able to recover these clusters. A natural framework to this end is the vector autoregressive (VAR) model because (i) it can express serial correlation between time points, (ii) it can express cross-correlation between variables, and (iii) given a fully parameterized VAR process, one can simulate random trajectories from that VAR process.

More specifically, to generate clusters of multivariate time series, we simulated from VAR process mixtures, for different values of a clusterability parameter λ. The clusterability parameter λ influences how easily separable the simulated clusters are (see Section Simulation experiments). Sample data are provided in the Supplementary Figure S1. We used the cluster purity measure [37] to record how well the true clustering could be recovered (for more details, see Methods).

VaDER was able to highly accurately recover the simulated clusters, achieving a cluster purity of >0.9 for $\lambda \approx 0.08$, and converging to 1.0 for larger λ (Fig. 2a). Moreover, even without extensive hyperparameter optimization, VaDER performed substantially better than hierarchical clustering using various distance measures, some of which were specifically designed for multivariate time series (multidimensional dynamic time warping [MD-DTW] [38] and Global Alignment Kernels [GAK] [39]) or short univariate time series (the STS distance [40]). Only for $\lambda < 0.04$ was VaDER outperformed by MD-DTW. This may be attributed to the fairly limited number of samples used for the simulation (n = 2,000), and omitting extensive optimization of VaDER's hyperparameters.

We used the same VAR framework to assess how varying degrees of missing values affect the performance of VaDER. Both MCAR and MNAR were simulated as described in the Methods. In the MCAR simulation, missing values were uniformly distributed across time and clinical outcome measures. In the MNAR simulation, the expected degree of missing values sigmoidally depended on time (see Methods). For varying clusterability levels λ, it can be seen that VaDER's implicit imputation
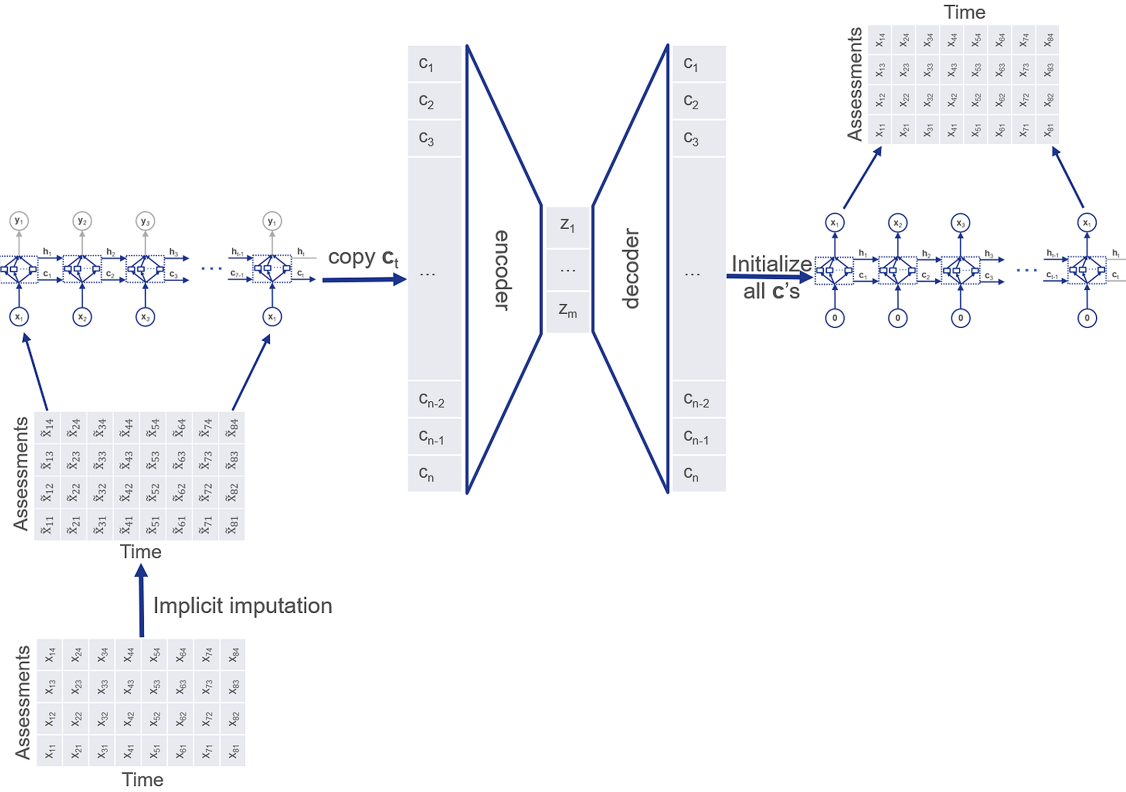
Figure 1: VaDER architecture.

Table 1. Multivariate time-series classification datasets used in this study

Name	k	n	p	n_t	n'_t	Source
ArabicDigits	10	8,800	13	4 - 93	24	UCI [41]
JapaneseVowels	9	640	12	7 - 29	15	UEA/UCR [42]
CharacterTrajectories	20	2,858	3	109 - 205	23	UCI [41]
UWave	8	4,478	3	315	25	UCI [41]

k: number of classes; n: number of samples; p: number of variables; n_t: number of time points; n'_t: number of samples after processing to equal and/or shorter length; UCI: University of California Irvine machine learning repository; UEA/UCR: University of East Anglia/University of California, Riverside time-series classification archive.

scheme is overall more robust against missing values than using VaDER with pre-imputation of missing values (Figs 2b and c).

VaDER achieves high accuracy on benchmark classification datasets
As an additional validation step towards applying VaDER to real-world clinical data, we collected a number of real-world benchmark datasets for multivariate time-series classification (Table 1). The datasets were normalized and processed to equal and/or shorter length as described in the Methods.

Comparing the ability of VaDER in recovering these a priori known classes to the other methods mentioned above reveals that VaDER consistently achieves better results (Fig. 3a). Moreover, VADER's approach of integrating imputation with model training again outperforms pre-imputation of missing values (Figs 3b and c).

***Application 1: VaDER identifies clinically diverse AD patient sub-groups***
After the technical validation using simulated and benchmark data, we applied VaDER to clinical data for identifying meaningful patient subgroups. From ADNI [25], we collected data from

689 patients who at some point received a diagnosis of dementia during the course of this study. Four different cognitive assessment scores were available at 8 different visits: ADAS13, CDRSB, MMSE, and FAQ. We pre-processed the data as described in the ADNI data preparation section. Overall, the fraction of missing values was ∼43%. We used VaDER to cluster patients by disease progression as measured using these cognitive assessments.

Hyperparameter optimization was performed by random grid search as described in the Methods. For each number of clusters $k \in \{2\ldots15\}$, the prediction strength [43] of the corresponding optimal model was compared to a null distribution (see Section Hyperparameter optimization and choice of number of clusters), which is shown in Supplementary Figure S2.

For most practical applications, determining an unambiguously correct number of clusters k is not possible, and a wide range of rules of thumb exist (see, e.g., [43–47]). For our subsequent analyses, we chose $k = 3$. This demonstrated relatively high prediction strength, significantly different from the null distribution, while still allowing VaDER to demonstrate its ability to uncover potentially interesting statistical interactions be-
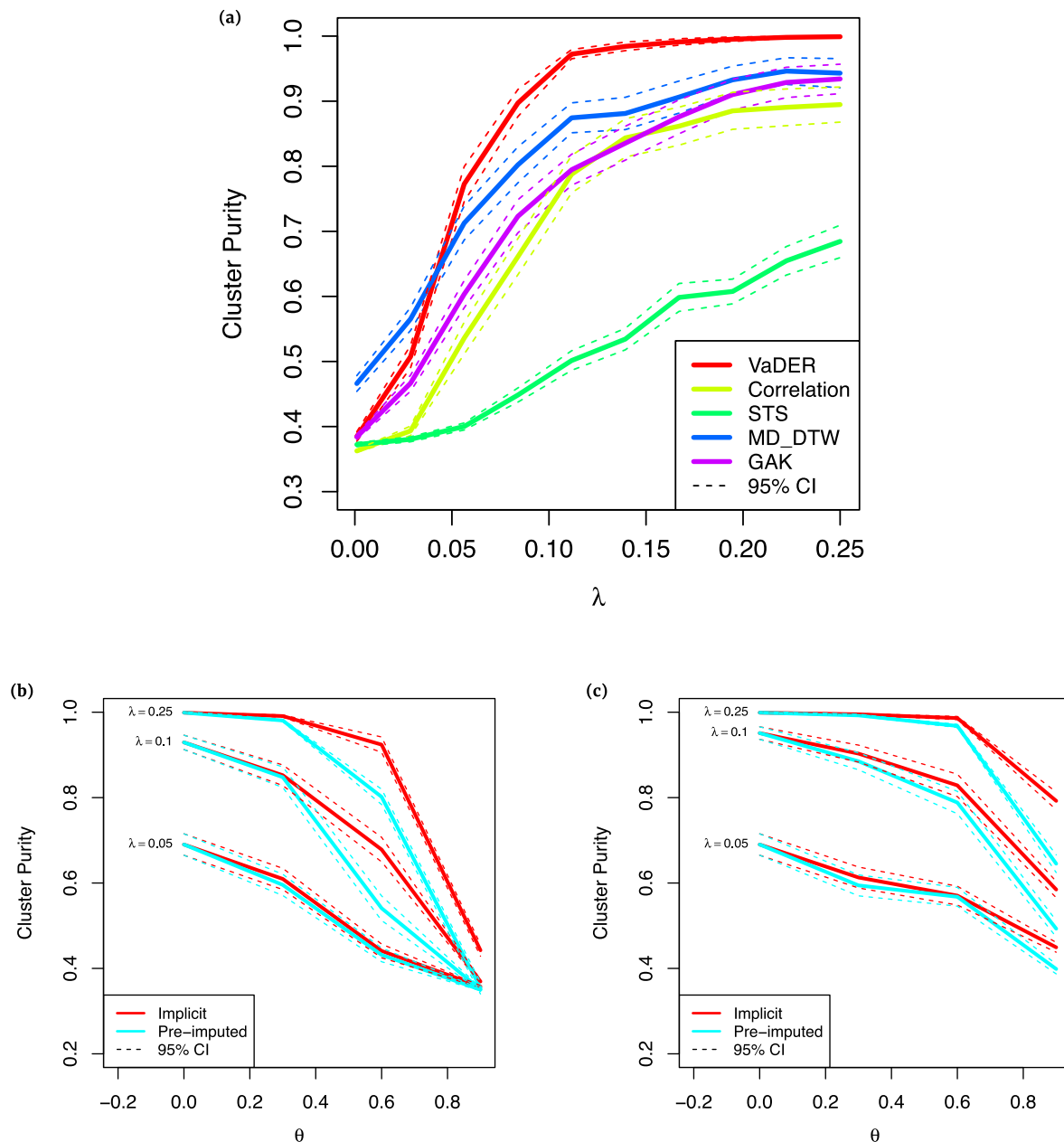
Figure 2: VaDER performance on simulated data, with varying degrees of clusterability and missingness. (a) Cluster purity [37] for clustering of simulated data as a function of the clusterability parameter λ, with higher λ implying a higher degree of similarity between profiles in the same cluster. Results are shown for VaDER as well as hierarchical clustering using 5 different distance measures, (i) Euclidean distance, (ii) Pearson correlation, (iii) the STS distance [40], (IV) multi-dimensional dynamic time warping (MD-DTW), [38] and (5) Global Alignment Kernels (GAK) [39]. (b) Cluster purity as a function of the fraction θ of values missing completely at random (MCAR), for various levels of the clusterability parameter λ, for both VaDER with implicit imputation and VaDER with pre-imputation. Confidence intervals were determined by repeating the clustering 100 times using newly generated random data and missingness patterns. (c) Cluster purity as a function of the fraction θ of values missing not at random (MNAR) (see Methods for details), for various levels of the clusterability parameter λ, for both VaDER with implicit imputation and VaDER with pre-imputation. Confidence intervals were determined by repeating the clustering 100 times using newly generated random data and missingness patterns.

tween trajectories of different cognitive assessments. A statistical interaction between different cognitive assessments could, e.g., manifest in the ability to distinguish patient subgroups based on 1 cognitive assessment, while this is not possible on another assessment. Another example would be a permuted ordering of clusters with respect to different assessment scores.

For ADNI data the resulting cluster mean trajectories are shown in Fig. 4 and demonstrate that (i) VaDER effectively clusters the data into patient subgroups showing divergent disease

progression and (ii) VaDER is able to find interactions between the different cognitive assessments, which would be principally difficult to distill from univariate analyses of the assessments. For example, the patients in Cluster 1 are the patients whose disease is the most severely progressing when assessed using ADAS13, CDRSB, and MMSE. However, the FAQ assessment (instrumental activities of daily living) does not distinguish between these patients with severely progressing disease and the patients with more moderately progressing disease in Cluster 1.
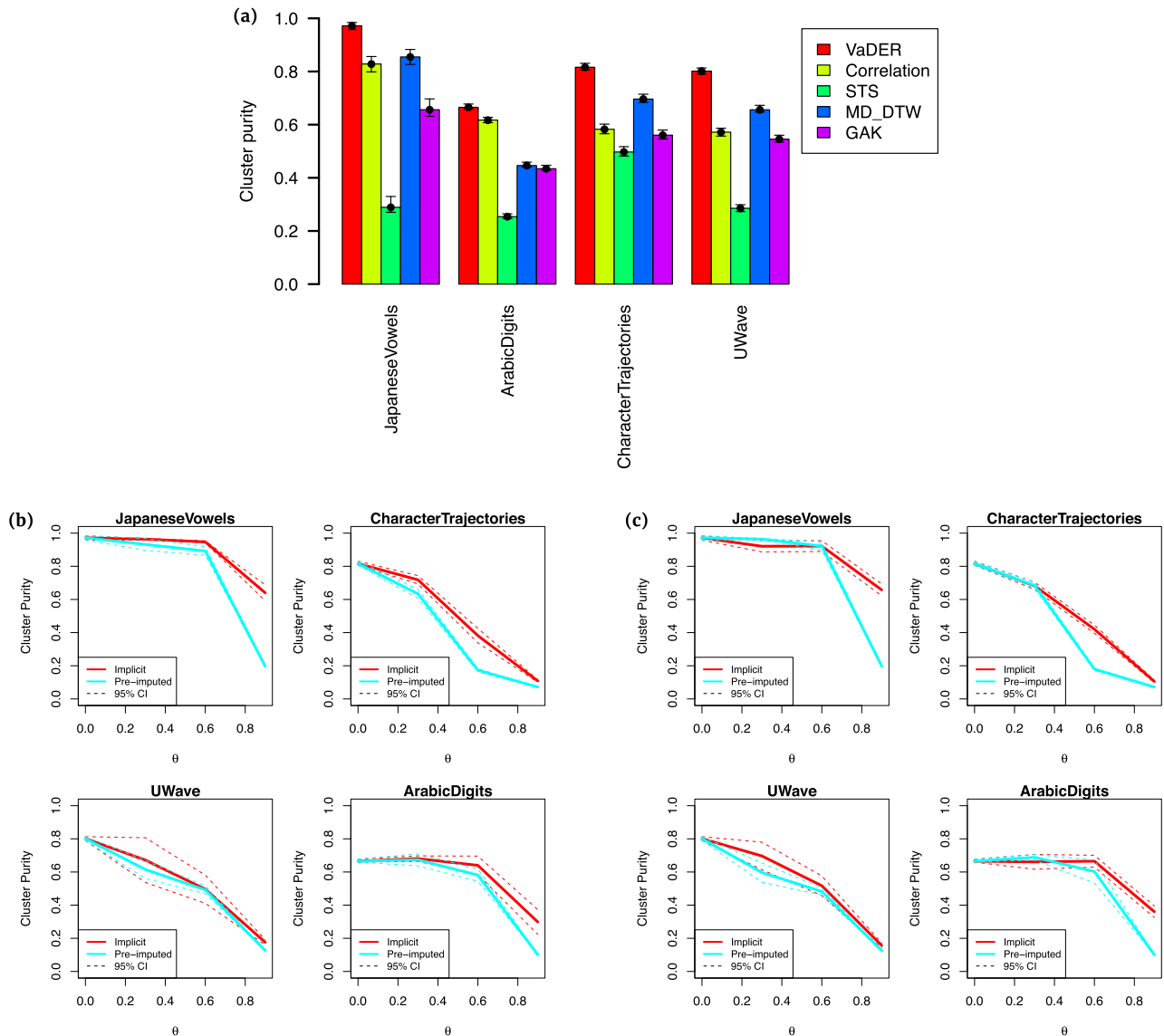
Figure 3: VaDER performance on benchmark data, for varying degrees of missingness. (a) Cluster purity [37] for clustering of benchmark data. Results are shown for VaDER as well as hierarchical clustering using 5 different distance measures, (i) Euclidean distance, (ii) Pearson correlation, (iii) the STS distance [40], (iv) multi-dimensional dynamic time warping (MD-DTW) [38], and (v) Global Alignment Kernels (GAK) [39]. For each dataset, the best performance across methods is marked by a horizontal dotted line. Confidence intervals were determined by bootstrapping the clustering 10^3 times. (b) Cluster purity as a function of the fraction θ of values missing completely at random (MCAR), for both VaDER with implicit imputation and VaDER with pre-imputation. Confidence intervals were determined by repeating the clustering 5 times using newly generated random missingness patterns. (c) Cluster purity as a function of the fraction θ of values missing not at random (MNAR), for both VaDER with implicit imputation and VaDER with pre-imputation. Confidence intervals were determined by repeating the clustering 5 times using newly generated random missingness patterns.

In addition to cognitive assessment measurements, ADNI presents a wealth of data on brain volume and various AD markers that we did not use for clustering. In these data, we identified numerous statistically significant associations with our patient subgroups. For example, the clusters strongly associated with time-to-dementia diagnosis relative to baseline, with Cluster 2 showing generally the shortest time and Cluster 0 the longest. The patients with relatively mildly progressing disease in cluster 0 also demonstrated on average a larger whole-brain volume at baseline, which moreover declined less steeply over time, compared to more patients with severely progressing disease. Especially the middle temporal gyri and fusiform gyri were larger (and shrinking more slowly over time), whereas the ventricles were smaller (and expanding more slowly over

time). Indeed, atrophy of the middle temporal gyri and fusiform gyri, as well as ventricular enlargement, have been associated with AD progression [48,49]. As another example, the patients with more severely progressing disease (Cluster 1 and especially Cluster 2) demonstrated lower cerebral glucose uptake and lower cerebrospinal Abeta42 levels, again confirming the literature [50,51] (see Methods and Supplementary Figures S3-8). These observations demonstrate that the clinical differences between our patient subgroups reflect known AD aspects.

Application 2: VaDER identifies clinically diverse PD patient subgroups

We additionally applied VaDER to clinical data from the Parkinson's Progression Markers Initiative (PPMI) [26]. From PPMI,
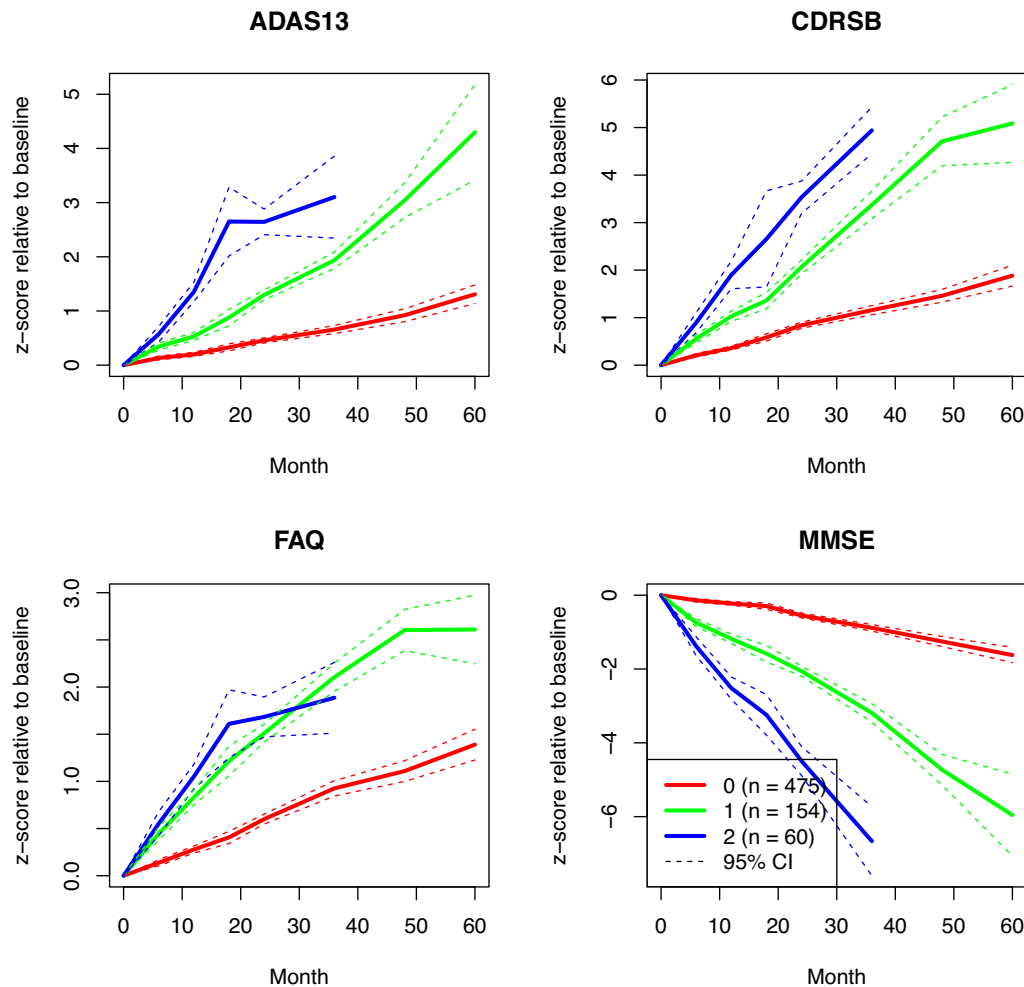
Figure 4: Normalized cluster mean trajectories relative to baseline (x-axis in months), as identified by VaDER from the ADNI cognitive assessment data.

we collected data from 362 *de novo* PD patients who had received a diagnosis within a period of 2 years before study onset and had initially not been treated. Nine variables on several motor and non-motor symptoms (UPDRS total, UPDRS1–3, tremor dominance [TD], postural instability and gait disturbance [PIGD], RBD, ESS, SCOPA-AUT) measured at either 5 or 10 time points were available. The data were pre-processed as described in the PPMI data preparation section. Overall, the fraction of missingness values was ∼17% (or ∼31%, when including time points entirely missing for some assessments). We again used VaDER to cluster patients according to disease progression as measured by these assessments.

Hyperparameter optimization and selection of the number of clusters was performed in the same way as for ADNI (see Supplementary Figure S9), and we decided on $k = 3$ patient subgroups accordingly. The resulting cluster mean trajectories are shown in Fig. 5. These again illustrate that (i) VaDER effectively clusters the data into clinically divergent patient subgroups, and (ii) VaDER is able to find interactions between the assessments that would principally be difficult to find based on univariate analyses alone. For example, Cluster 0 represents patients with a moderate progression in terms of mental impairment, behavior, and mood (UPDRS1) and autonomic dysfunction (SCOPA). However, these patients remain relatively stable, or even improve, on

many other assessments, such as TD, the self-reported ability to perform activities of daily life (UPDRS2), and motor symptoms evaluation (UPDRS3).

Similar to ADNI, PPMI presents a wealth of additional data on brain volume and various PD markers that were not used for clustering. Aligning these data with our PD patient subgroups, we found numerous statistically significant associations that confirmed existing literature, many related to quality of life and physiological changes to the brain. For example, men were over-represented in cluster 1 and showed the most severe disease progression, confirming the literature on sex differences in PD (e.g., [52]). Moreover, these patients with severely progressing disease showed an expected steeply declining ability to perform activities of daily living (modified Schwab and England score [53]), as well as rapidly developing symptoms of depression (geriatric depression scale [54]), common in patients with PD [55]. Additionally, these patients demonstrated physiological differences in the brain when compared to patients with more mildly progressing disease. Examples are the caudate nucleus and putamen brain regions, which were smaller at baseline and during follow-up examinations in the patients with more severely progressing disease in Cluster 1 and, from the literature, are known to be subject to atrophy in PD [56] (see Methods and Supplementary Figures S10-15). These observations demonstrate that the clinical differences be-
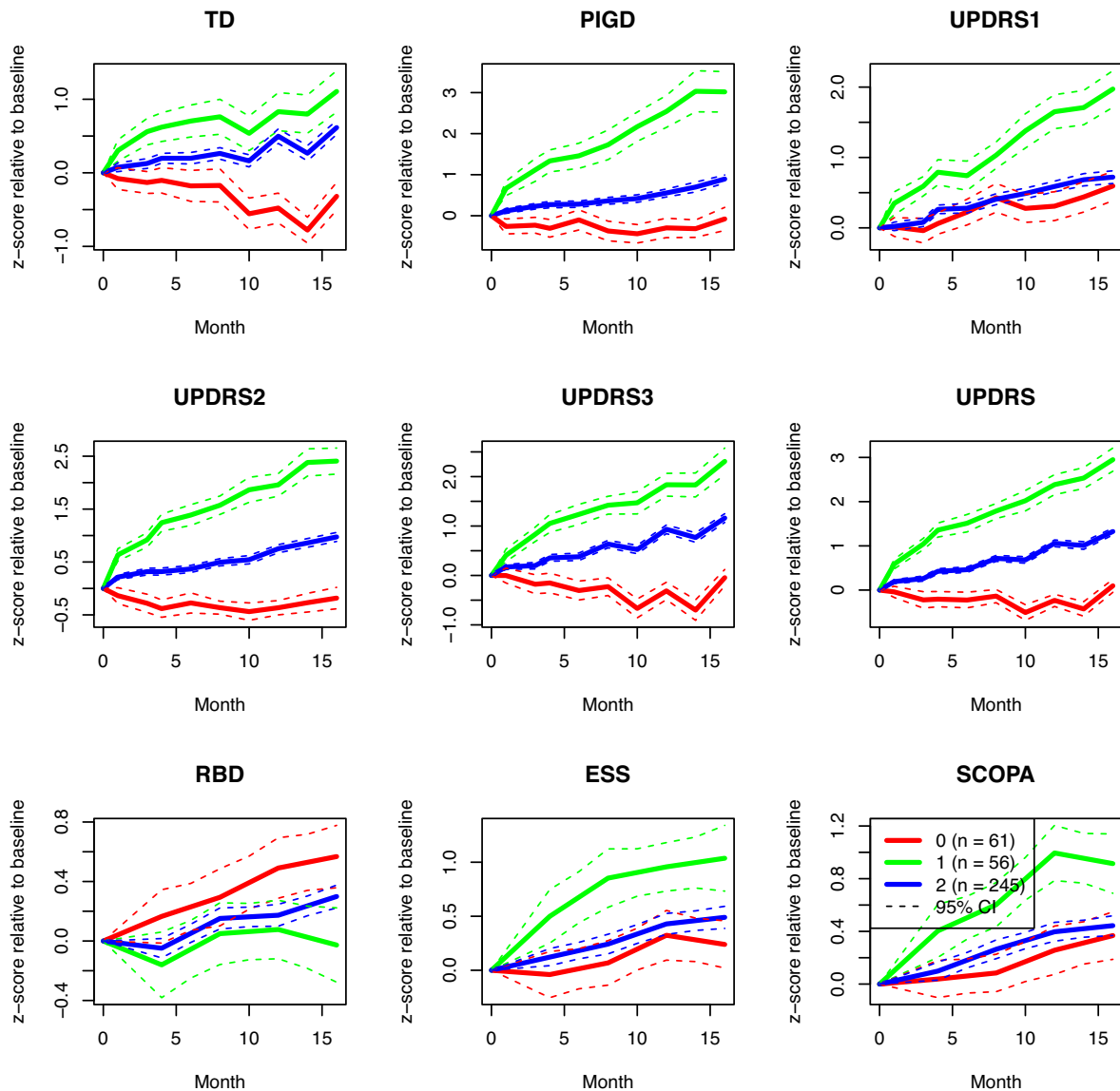
Figure 5: Normalized cluster mean trajectories relative to baseline (*x*-axis in months), as identified by VaDER from the PPMI motor/non-motor score data.

tween our patient subgroups reflect known aspects of PD disease progression.

Discussion and conclusions

Identifying subgroups of patients with similar progression patterns can help to better elucidate the heterogeneity of complex diseases. Together with predictive machine learning methods, this might improve decision making on the right time and type of treatment for an individual patient, as well as the design of clinical studies. However, one of the main challenges is the multifaceted nature of progression in many areas of disease.

In this article, we proposed VaDER, a method for clustering multivariate, potentially short, time series with many missing values, a setting that seems generally not well addressed in the literature so far but is nonetheless often encountered in clinical study data.

We validated VaDER by showing the very high accuracy on clustering simulated and real-world benchmark data with a known ground truth. We then applied VaDER to data from (i) ADNI and (ii) PPMI, resulting in subgroups characterized by clinically highly divergent disease progression profiles. A comparison with other data from ADNI and PPMI, such as brain imaging and motor and cognitive assessment data, furthermore supported the observed patient subgroups.

VaDER has 2 main distinctive features. One is that VaDER deals directly with missing values. For clinical research this is crucial because clinical datasets often show a very high degree of missing values [57, 58]. The other main distinctive feature is that, as opposed to existing methods [10–14], VaDER is specifically designed to deal with multivariate and relatively short time series that are typical for (observational) clinical studies. However, it is worthwhile to mention that the application of VaDER is not per se limited to longitudinal clinical study data or to time series of short length. Future applications (potentially requiring some adaptations) could, e.g., include data originating from electronic health records, multiple wearable sensors, video recordings, or time-series gene (co-)expression. Moreover,

VaDER could be used as a generative model: given a trained model, it is possible to generate "virtual" patient trajectories.

Altogether, we believe that our results show that VaDER has the potential to substantially enhance future patient stratification efforts and multivariate time series clustering in general.

Methods

Data preparation

ADNI data preparation
Data used in the preparation of this article were obtained from the ADNI database (adni.loni.usc.edu). The 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 (MRI), positron emission tomography, other biological markers, and clinical and neuropsychological assessment can be combined to measure the progression of mild cognitive impairment and early AD. For up-to-date information, see www.adni-info.org.

The ADNIMERGE R-package [59] contains mainly 2 categories of data, (i) longitudinal and (ii) non-longitudinal. These data represent 1,737 participants that include healthy controls and patients with a diagnosis of AD. The non-longitudinal features such as demographic characteristics and apolipoprotein E4 status were measured only once, at baseline. The longitudinal features (i.e., neuroimaging features, cerebrospinal fluid biomarkers, cognitive tests, and everyday cognition) were recorded over a span of 5 years.

Clinical data In the present study, we have focused on those participants who received a diagnosis of AD at baseline or during 1 of the follow-up visits. After this filtering step, we had a total of 689 patients. For these 689 patients, 4 cognitive assessments were selected for clustering:

- ADAS-13: Alzheimer's Disease Assessment Scale
- CDRSB: Clinical Dementia Rating Sum of Box Score
- FAQ: Functional Activities Questionnaire
- MMSE: Mini–Mental State Examination

The above assessments were taken at baseline and at 6, 12, 18, 24, 36, 48, and 60 months after baseline. For each of the 4 cognitive assessments, all time points were normalized relative to baseline by (i) subtracting the baseline mean across the 689 patients and (ii) dividing by the baseline standard deviation across the 689 patients.

Imaging data All available MRI scans (T1-weighted scans) from the ADNI database were quantified by an open-source, automated segmentation pipeline at the Erasmus University Medical Center, The Netherlands. The number of slices of the T1-weighted scans varied from 160 to 196 and the in-plane resolution was 256×256 on average, yielding an overall voxel size of $1.2 \times 1.0 \times 1.0$ mm. From the 1,715 baseline ADNI scans, the volumes of 34 bilateral cortical brain regions, 68 structures in total, were calculated using a model- and surface-based automated image segmentation procedure, incorporated in the FreeSurfer Package (v.6.0 [60]). Segmentation in Freesurfer was performed by rigid-body registration and nonlinear normalization of images to a probabilistic brain atlas. In the segmentation process, each voxel of the MRI volumes was labeled automatically as a corresponding brain region based on 2 different cortex parcellation guides [61, 62], subdividing the brain into 68 and 191 regions, respectively.

PPMI data preparation
Patients were selected if their PD diagnosis was <2 years old at baseline and if follow-up data were available for ≥48 months (5–10 time points), resulting in a total of 362 patients. For these 362 patients, a set of 9 motor and non-motor symptoms were selected for clustering:

- TD: tremor dominance
- PIGD: postural instability and gait disturbance
- UPDRS1: Unified Parkinson Disease Rating Scale, part 1: mentation, behavior, and mood
- UPDRS2: Unified Parkinson Disease Rating Scale, part 2: activities of daily living
- UPDRS3: Unified Parkinson Disease Rating Scale, part 3: motor examination
- UPDRS: Unified Parkinson Disease Rating Scale (UPDRS1 + UPDRS2 + UPDRS3)
- RBD: REM sleep behavior disorder
- ESS: Epworth Sleepiness Scale
- SCOPA-AUT: Scales for Outcomes in Parkinson Disease: Assessment of Autonomic Dysfunction

All scores were normalized relative to baseline by (i) subtracting the baseline mean across all patients and (ii) dividing by the baseline standard deviation across all patients.

For some assessments, fewer time points were available. These were treated as missing values.

Benchmark datasets for multivariate time-series classification
Because no benchmark datasets exist for multivariate time series clustering, we collected a number of benchmark datasets for multivariate time-series classification [41, 42]. Because currently, VaDER still only works with equal-length time series (see also Section Discussion and conclusions), we pre-processed all samples to equal-length time series by linear interpolation between start and end point. Following [63, 64], we chose constant lengths of $\lceil T_{\max} / \lceil \frac{T_{\max}}{25} \rceil \rceil$, where $T_{\max}$ is the maximum length of the lengths of the samples in a given dataset.

Moreover, all resulting time series were standardized to zero mean and unit variance.

VaDER

The VaDER model is extensively described in the Results section. This section describes how VaDER was trained.

Pre-training
Similar to Jiang et al. [23], we pre-train VaDER by disregarding the latent loss during the first epochs, essentially fitting a non-variational LSTM autoencoder to the data. We then fit a Gaussian mixture distribution in the latent space of this autoencoder and use its parameters to initialize the final variational training of VaDER.

Hyperparameter optimization and choice of number of clusters
We used prediction strength [43] to select suitable values for VaDER's hyperparameters. These comprise the following:

- number of layers (for both ADNI and PPMI: {1, 2})
- number of nodes per hidden layer (for ADNI: {2^0, 2^1, 2^2, 2^3, 2^4, 2^5, 2^6}; for PPMI: {2^0, 2^1, 2^2, 2^3, 2^4, 2^5, 2^6, 2^7})
- learning rate (for both ADNI and PPMI: {10^{-4}, 10^{-3}, 10^{-2}, 10^{-1}})
- mini-batch size (for both ADNI and PPMI: {2^4, 2^5, 2^6, 2^7})

Hyperparameter optimization was performed via a random grid search (i.e., by randomly sampling a predefined hyperparameter grid) with repeated cross-validation ($n = 20$), using the reconstruction loss as objective. This was done during the pre-training phase of VaDER.

After hyperparameter optimization we trained VaDER models for different numbers of clusters $k \in \{2 \dots 15\}$. For each k, prediction strength was computed by 2-fold cross-validation [43]: for a given training and test dataset:

(1) Train VaDER on the training data (the training data model).
(2) Assign clusters to the test data using the training data model.
(3) Train VaDER on the test data (the test data model).
(4) Assign clusters to the test data using the test data model.
(5) Compare the resulting 2 clusterings: for each cluster of the test data model, compute the fraction of pairs of samples in that cluster that are also assigned to the same cluster by the training data model. Prediction strength is defined as the minimum proportion across all clusters of the test data model [43].

Prediction strength was then compared to an empirical null distribution of that measure. The null distribution of the prediction strength was computed by randomly permuting the predicted cluster labels 10^3 times, then recomputing the prediction strength, and eventually taking the average of the 10^3 prediction strength values. Doing this for all 20 repeats resulted in 20 values for the eventual null distribution, which were then compared to 20 actual prediction strength values (similarly, 1 for each repeat) by a paired Wilcoxon rank-sum test.

Simulation experiments

Overview of data-generating process
To better understand the performance of VaDER we conducted an extensive simulation study: we simulated multivariate (short) time series via VAR processes [65] because (i) they can model the auto-correlation between time points, (ii) they can model the cross-correlation between variables, and (iii) given a VAR, one can generate random trajectories from that VAR.

We used mixtures of VAR processes to simulate multivariate time-series data of the same dimensions as the ADNI data: 4 variables measured over 8 time points. Given a clusterability factor λ, we generated trajectories as follows:

(1) Sample coefficient matrices for 3 VAR(8) processes, by randomly sampling the individual entries of each 4×4 matrix from the uniform distribution $\mathcal{U}(-0.1, 0.1)$. Multiply each of the matrix entries by λ.
(2) Randomly sample 3 additional 4×4 matrices from $\mathcal{U}(-0.1, 0.1)$ and multiply each by its own transpose. Let each of the results correspond to the variance-covariance matrix of 1 of the 3 VAR(8) processes.
(3) Repeat 10^3 times:
   (1) Randomly select 1 of the 3 VAR(8) processes (with equal probability).
   (2) Generate a random trajectory from the selected VAR(8) process.

The above generates 1 set of random data. Given a value of λ, the entire sampling process was repeated 100 times, and each of the 100 datasets was clustered using both VaDER and hierarchical clustering.

For computational reasons, hyper-parameters for VaDER were fixed and not further optimized during our simulation (10^2 epochs of both pre-training and training, learning rate: 10^{-4}, 2 hidden layers: [36, 4], batch size: 64).

Comparison against hierarchical clustering
We compared VaDER against a conventional hierarchical clustering (complete linkage), in which we flatten the $N \times M$ data matrices of each patient into vectors. We considered 3 distance measures for these vectors:

- Pearson correlation
- Euclidean distance
- STS distance [40], a distance measure specifically developed for univariate short time series. The STS distance relies on the difference between adjacent time points. Here we first computed the STS distance for each of the different clinical outcome measures and then summed these up to arrive at an aggregated STS distance across the M clinical measures.

Additionally, we compared VaDER against hierarchical clustering using 2 distance measures specifically designed for multivariate time series:

- MD-DTW [38]
- Fast GAK [39]

Given that VaDER is non-deterministic, we ran 100 replicates for each (simulated/benchmark) dataset and determined the consensus clustering by hierarchically clustering a consensus matrix listing, for each pair of samples, how often these 2 samples were clustered together across the 100 replicates.

Simulating missing data
To test the ability of VaDER to deal with missing data we performed a separate set of simulations: Let L be the number of patients in our dataset and $\mathbf{x}^l \in \mathbb{R}^{N \times M}$ a single patient trajectory ($l \in 1 \dots L$), where N is the number of time points and M is the number of measured features. MCAR were simulated by an individual entry $\mathbf{x}^l_{ij}$ to missing with probability θ.

MNAR was simulated by letting the probability of a missing value for entry $\mathbf{x}^l_{ij}$ depend on time. More specifically, each individual entry $\mathbf{x}^l_{ij}$ was set to missing with probability $1/(1 + e^{i_0 - i/k})$, where $i_0 = (1 + N)/2$ and k was varied to result in different overall missingness fractions θ.

To compare VADER's implicit imputation with pre-imputation, missing values generated using the above approach were additionally imputed using mean substitution: each missing value was substituted with the average conditioned on the relevant time point and variable.

Given that VaDER is non-deterministic, we ran 20 replicates for each (simulated/benchmark) dataset and determined the consensus clustering by hierarchically clustering a consensus matrix listing, for each pair of samples, how often these 2 samples were clustered together across the 20 replicates. Confidence intervals were determined by repeating the aforementioned procedure 100 times (simulation experiments) or 5 times (benchmark experiments) with newly generated missingness patterns (simulation/benchmark experiments) and/or data (simulation experiments).

Estimating clustering performance
For the simulation and benchmark datasets, the number of clusters is a priori known. Hence, an intuitive measure of comparing the performance between the different algorithms is cluster purity [37]. Cluster purity can be interpreted as the fraction of correctly clustered samples and is calculated as follows:

(1) For each cluster, count the number of samples from the majority class in that cluster.
(2) Sum the above counts.
(3) Divide by the total number of samples.

For the ADNI and PPMI data, the number of clusters is not a priori known. Hence, performance was recorded using the adjusted Rand index [66, 67] for different values of λ in the interval [0.001, 0.25]. For $\lambda \gtrsim 0.25$, generating coefficient matrices that lead to stable VARs becomes very difficult.

Post hoc analysis of patient clusters

We collected a wide range of additional variables from ADNI and PPMI and assessed the association of the identified patient subgroups with a given variable by multinomial logistic regression. For any baseline variable x, we first fitted the following full model:

$$\text{subgroup} \sim x + \text{confounders.} \tag{6}$$

Each of these full models was then compared to a null model:

$$\text{subgroup} \sim \text{confounders} \tag{7}$$

by means of a likelihood ratio test.

For any longitudinal variable x measured at time points t, we first fitted the following multinomial logistic regression model:

$$\text{subgroup} \sim x + t + x * t + \text{confounders.} \tag{8}$$

We tested this model against the null model:

$$\text{subgroup} \sim \text{confounders} \tag{9}$$

by performing a likelihood ratio test and applying a false discovery rate correction for multiple testing. If the above test was found to be significant ($q < 0.05$), we tested the effects of the individual terms $x * t$, x, and t against the same null model above.

Confounders considered were age, education, and sex but were only included when univariate results were significantly associated with subgroup. For ADNI, this was only age ($P = 0.0029$, ANOVA F-test). For PPMI, this was only sex ($P = 0.0017$, χ^2 test).

In the post hoc analysis, only complete cases were included; i.e., patients with missing values were ignored.

Availability of Supporting Source Code and Requirements

A complete implementation of VaDER in Python/Tensorflow: https://github.com/johanndejong/VaDER.
An R-package for streamlining the processing of PPMI data: https://github.com/patzaw/PPMI-R-package-generator.
Other code used for generating results presented in this article: https://github.com/johanndejong/VaDER_supporting_code.
Snapshots of all the above code and other supporting data are also available in the GigaScience database, GigaDB [68].

Additional Files

Supplementary information: Supplementary Methods and Results are available via the additional file associated with this article.
Supplementary Figure S1 Multivariate short time series data simulated using vector autoregressive processes, for 4 variables, 8 time points and 3 clusters, and different levels of the similarity parameter λ.
Supplementary Figure S2 ADNI: prediction strength of VaDER for each k (blue) and the corresponding permutation-based null distribution.
Supplementary Figure S3 ADNI: associations of the VaDER clustering with a wide range of other baseline data available from ADNI.
Supplementary Figure S4 ADNI: associations of the VaDER clustering with a wide range of other baseline data available from ADNI.
Supplementary Figure S5 ADNI: associations of the VaDER clustering with a wide range of other baseline data available from ADNI.
Supplementary Figure S6 ADNI: associations of the VaDER clustering with a wide range of other baseline data available from ADNI.
Supplementary Figure S7 ADNI: associations of the VaDER clustering with a wide range of other baseline data available from ADNI.
Supplementary Figure S8 ADNI: associations of the VaDER clustering with a wide range of other longitudinal data available from ADNI.
Supplementary Figure S9 PPMI: prediction strength of VaDER for each k (blue) and the corresponding permutation-based null distribution.
Supplementary Figure S10 PPMI: associations of the VaDER clustering with a wide range of other baseline data available from PPMI.
Supplementary Figure S11 PPMI: associations of the VaDER clustering with a wide range of other baseline data available from PPMI.
Supplementary Figure S12 PPMI: associations of the VaDER clustering with a wide range of other baseline data available from PPMI.
Supplementary Figure S13 PPMI: associations of the VaDER clustering with a wide range of other longitudinal data available from PPMI.
Supplementary Figure S14 PPMI: associations of the VaDER clustering with a wide range of other longitudinal data available from PPMI.
Supplementary Figure S15 PPMI: associations of the VaDER clustering with a wide range of other longitudinal data available from PPMI.

Abbreviations

AD: Alzheimer disease; ADAS-13: Alzheimer Disease Assessment Scale; ADNI: Alzheimer's Disease Neuroimaging Initiative; ANOVA: analysis of variance; CDRSB: Clinical Dementia Rating Sum of Box Score; ESS: Epworth Sleepiness Scale; FAQ: Functional Activities Questionnaire; GAK: Global Alignment Kernels; LSTM: long short-term memory; MAR: missing at random; MCAR: missing completely at random; MD-DTW: multidimensional dynamic time warping; MMSE: Mini–Mental State Examination; MNAR: missing not at random; MRI: magnetic

resonance imaging; PD: Parkinson disease; PIGD: postural instability and gait disturbance; PPMI: Parkinson's Progression Markers Initiative; RBD: REM sleep behavior disorder; SCOPA: Scales for Outcomes in Parkinson's Disease; STS distance: short-time-series distance; TD: tremor dominance; UPDRS: Unified Parkinson's Disease Rating Scale; UPDRS1: Unified Parkinson's Disease Rating Scale, Part 1; UPDRS2: Unified Parkinson's Disease Rating Scale, Part 2; UPDRS3: Unified Parkinson's Disease Rating Scale, Part 3; UCI: University of California Irvine Machine Learning Repository; UEA/UCR: University of East Anglia/University of California, Riverside Time-Series Classification Archive; VaDE: variational deep embedding; VaDER: variational deep embedding with recurrence; VAR: vector autoregression; VCF: variant call format.

Competing Interests

Funding

Authors' Contributions

Method development: J.d.J., H.F.; implementation and testing: J.d.J.; Data preparation: M.A.E., P.W., R.K., M.S., A.A., P.G.; image analysis: H.V.; supervision: H.F., M.H.A.; definition of research project: H.F.

Acknowledgments

References

1. Hruby A, Hu FB. The epidemiology of obesity: a big picture. Pharmacoeconomics 2015;**33**(7):673–89.
2. van Tilburg J, van Haeften TW, Pearson P, et al. Defining the genetic contribution of type 2 diabetes mellitus. J Med Genet 2001;**38**(9):569–78.
3. Cordell HJ, Todd JA. Multifactorial inheritance in type 1 diabetes. Trends Genet 1995;**11**(12):499–504.
4. Ruppert V, Maisch B. Genetics of human hypertension. Herz 2003;**28**(8):655–62.
5. Poulter N. Coronary heart disease is a multifactorial disease. Am J Hypertens 1999;**12**(10):92S–5S.
6. Hastie T, Tibshirani R, Friedman JH. The Elements of Statistical Learning: Data Mining, Inference, and Prediction. 2nd Ed. Springer Series in Statistics. New York, NY: Springer; 2009.
7. Kannan R, Vempala S. On clusterings - good, bad and spectral. In: Proceedings 41st Annual Symposium on Foundations of Computer Science, Redondo Beach, CA. IEEE; 2000:367–77.
8. Jain A, Dubes R. Algorithms for Clustering Data. Englewood Cliffs, NJ: Prentice-Hall; 1988.
9. Fukunaga K, Hostetler LD. The estimation of the gradient of a density function, with applications in pattern recognition. IEEE Trans Inf Theory 1975;**21**:32–9.
10. Aghabozorgi S, Seyed Shirkhorshidi A, Ying Wah T. Time-series clustering - a decade review. Inf Syst 2015;**53**(C):16–38.
11. Rani S, Sikka G. Recent techniques of clustering of time series data: a survey. Int J Comput Appl 2012;**52**(15):1–9.
12. Liao TW. Clustering of time series data: a survey. Pattern Recognit 2005;**38**(11):1857–74.
13. Ghassempour S, Girosi F, Maeder A. Clustering multivariate time series using hidden Markov models. Int J Environ Res Public Health 2014;**11**(3):2741–63.
14. Sun J. Clustering multivariate time series based on Riemannian manifold. Electron Lett 2016;**52**(2):1607–9.
15. Rubin DB. Inference and missing data. Biometrika 1976;**63**(3):581–92.
16. Kang H. The prevention and handling of the missing data. Korean J Anesthes 2013;**64**(5):402–6.
17. Mikolov T, Sutskever I, Chen K, et al. Distributed representations of words and phrases and their compositionality. In: Proceedings of the 26th International Conference on Neural Information Processing Systems - Volume 2 NIPS'13, Lake Tahoe, NV. Curran Associates Inc.; 2013:3111–9.
18. Hinton GE, Salakhutdinov RR. Reducing the dimensionality of data with neural networks. Science 2006;**313**(5786): 504–7.
19. Frome A, Corrado GS, Shlens J, et al. DeViSE: A Deep Visual-Semantic Embedding Model. In: Burges CJC, Bottou L, Welling M , et al., eds. Advances in Neural Information Processing Systems. Curran Associates; 2013:2121–9.
20. Asgari E, Mofrad MRK. Continuous distributed representation of biological sequences for deep proteomics and genomics. PLoS One 2015; **10**(11):1–15.

21. Trigeorgis G, Bousmalis K, Zafeiriou S, et al. A deep semi-NMF model for learning hidden representations. In: Xing EP, Jebara T , eds. Proceedings of the 31st International Conference on Machine Learning, Beijing, China. PMLR; 2014:1692–700.

22. Xie J, Girshick R, Farhadi A. Unsupervised deep embedding for clustering analysis. In: Proceedings of the 33rd International Conference on International Conference on Machine Learning - Volume 48 ICML'16. JMLR; 2016:478–487.

23. Jiang Z, Zheng Y, Tan H, et al. Variational deep embedding: an unsupervised and generative approach to clustering. In: Proceedings of the Twenty-Sixth International Joint Conference on Artificial Intelligence (IJCAI-17). 2017:1965–72.

24. Hochreiter S, Schmidhuber J. Long short-term memory. Neural Comput 1997;**9**(8):1735–80.

25. Petersen RC, Aisen PS, Beckett LA, et al. Alzheimer's Disease Neuroimaging Initiative (ADNI). Neurology 2010;**74**(3): 201–9.

26. Marek K, Jennings D, Lasch S, et al. The Parkinson Progression Marker Initiative (PPMI). Prog Neurobiol 2011;**95**(4):629–35.

27. Komarova NL, Thalhauser CJ. High degree of heterogeneity in Alzheimer's disease progression patterns. PLoS Comput Biol 2011;**7**(11), doi:10.1371/journal.pcbi.1002251.

28. Lam B, Masellis M, Freedman M, et al. Clinical, imaging, and pathological heterogeneity of the Alzheimer's disease syndrome. Alzheimers Res Ther 2013;**5**(1):1, doi:10.1186/alzrt155.

29. Lewis SJG, Foltynie T, Blackwell AD, et al. Heterogeneity of Parkinson's disease in the early clinical stages using a data driven approach. J Neurol Neurosurg Psychiatry 2005;**76**(3):343–8.

30. von Coelln R, Shulman LM. Clinical subtypes and genetic heterogeneity: of lumping and splitting in Parkinson disease. Curr Opin Neurol 2016;**29**(6):727–34.

31. Parkinson's Disease Information Page. https://www.ninds.nih.gov/Disorders/All-Disorders/Parkinsons-Disease-Information-Page.

32. Kingma DP, Welling M. Auto-encoding variational Bayes. arXiv 2013:1312.6114.

33. Doersch C. Tutorial on variational autoencoders. arXiv 2016:1606.05908.

34. Gers FA, Schraudolph NN, Schmidhuber J. Learning precise timing with LSTM recurrent networks. J Mach Learn Res 2003;**3**:115–43.

35. Lipton ZC, Kale DC, Wetzel RC. Directly modeling missing data in sequences with RNNs: improved classification of clinical time series. In: Proceedings of the 1st Machine Learning for Healthcare Conference, PMLR 56. JMLR; 2016: 253–270.

36. Nazábal A, Olmos PM, Ghahramani Z, et al. Handling incomplete heterogeneous data using VAEs. arXiv 2018:1807.03653.

37. Manning CD, Raghavan P, Schütze H. Introduction to Information Retrieval. New York, NY: Cambridge University Press; 2008.

38. Tormene P, Giorgino T, Quaglini S, et al. Matching incomplete time series with dynamic time warping: an algorithm and an application to post-stroke rehabilitation. Artif Intell Med 2008;**45**(1):11–34.

39. Cuturi M. Fast global alignment kernels. In: Getoor L, Scheffer T , eds. ICML Omnipress; 2011:929–36.

40. Möller-Levet CS, Klawonn F, Cho K, et al. Fuzzy clustering of short time-series and unevenly distributed sampling points. In: Berthold MR, Lenz H, Bradley E , et al., eds. Advances in Intelligent Data Analysis V, 5th International Symposium on Intelligent Data Analysis, IDA 2003, Berlin, Germany. Springer; 2003:330–40.

41. Dua D, Graff C. UCI Machine Learning Repository. 2017. http://archive.ics.uci.edu/ml.

42. Bagnall A, Lines J, Vickers W. The UEA, UCR Time Series Classification Repository. http://www.timeseriesclassification.com. Accessed 15 August 2019.

43. Tibshirani R, Walther G. Cluster validation by prediction strength. J Comput Graph Stat 2005;**14**(3):511–28.

44. Tibshirani R, Walther G, Hastie T. Estimating the number of clusters in a data set via the gap statistic. J R Stat Soc Series B Stat Methodol 2001;**63**(2):411–23.

45. Sugar CA, James GM. Finding the number of clusters in a dataset: an information-theoretic approach. J Am Stat Assoc 2003;**98**(463):750–63.

46. Thorndike RL. Who belongs in the family. Psychometrika 1953;**18**(4):267–76.

47. Rousseeuw PJ. Silhouettes: a graphical aid to the interpretation and validation of cluster analysis. J Comput Appl Math 1987;**20**:53–65.

48. Convit A, de Asis J, de Leon MJ, et al. Atrophy of the medial occipitotemporal, inferior, and middle temporal gyri in non-demented elderly predict decline to Alzheimer's disease. Neurobiol Aging 2000;**21**(1):19–26.

49. Nestor SM, Rupsingh R, Borrie M, et al. Ventricular enlargement as a possible measure of Alzheimer's disease progression validated using the Alzheimer's Disease Neuroimaging Initiative database. Brain 2008;**131**(9):2443–54.

50. Butterfield DA, Halliwell B. Oxidative stress, dysfunctional glucose metabolism and Alzheimer disease. Nat Rev Neurosci 2019;**20**(3):148–60.

51. Tapiola T, Alafuzoff I, Herukka SK, et al. Cerebrospinal fluid beta-amyloid 42 and tau proteins as biomarkers of Alzheimer-type pathologic changes in the brain. JAMA Neurol 2009;**66**(3):382–9.

52. Moisan F, Kab S, Mohamed F, et al. Parkinson disease male-to-female ratios increase with age: French nationwide study and meta-analysis. J Neurol Neurosurg Psychiatry 2016;**87**(9):952–7.

53. Schrag A, Jahanshahi M, Quinn N. What contributes to quality of life in patients with Parkinson's disease? J Neurol Neurosurg Psychiatry 2000;**69**(3):308–12.

54. Sheikh JI, Yesavage JA. Geriatric Depression Scale (GDS): Recent evidence and development of a shorter version. Clinical Gerontol 1986;**5**(1-2):165–73.

55. Marsh L. Depression and Parkinson's disease: current knowledge. Curr Neurol Neurosci Rep 2013;**13**(12):409.

56. Pitcher TL, Melzer TR, MacAskill MR, et al. Reduced striatal volumes in Parkinson's disease: a magnetic resonance imaging study. Transl Neurodegener 2012;**1**(1):17.

57. Pedersen AB, Mikkelsen EM, Cronin-Fenton D, et al. Missing data and multiple imputation in clinical epidemiological research. Clin Epidemiol 2017;**9**:157–66.

58. Marston L, Carpenter JR, Walters KR, et al. Issues in multiple imputation of missing data for large general practice clinical databases. Pharmacoepidemiol Drug Saf 2010;**19**(6):618–26.

59. ADNI Team. ADNIMERGE: Alzheimer's Disease Neuroimaging Initiative. 2018. R package version 0.0.1.

60. FreeSurfer. http://surfer.nmr.mgh.harvard.edu/.

61. Desikan R, Ségonne F, Fischl B, et al. An automated labeling system for subdividing the human cerebral cortex on MRI scans into gyral based regions of interest. NeuroImage 2006;**31**(3):968–80.

62. Destrieux C, Fischl B, Dale AM, et al. Automatic parcellation of human cortical gyri and sulci using standard anatomical nomenclature. NeuroImage 2010;**53**(1):1–15.

63. Wang L, Wang Z, Liu S. An effective multivariate time series classification approach using echo state network and adaptive differential evolution algorithm. Expert Syst Appl 2016;**43**(C):237–49.

64. Øyvind Mikalsen K, Bianchi FM, Soguero-Ruiz C, et al. Time series cluster kernel for learning similarities between multivariate time series with missing data. Pattern Recognit 2018;**76**:569–81.

65. Sims C. Macroeconomics and reality. Econometrica 1980;**48**(1):1–48.

66. Rand WM. Objective criteria for the evaluation of clustering methods. J Am Stat Assoc 1971;**66**(336): 846–50.

67. Hubert L, Arabie P. Comparing partitions. J Classif 1985;**2**(1):193–218.

68. de Jong J, Emon MA, Wu P, et al. Supporting data for "Deep learning for clustering of multivariate clinical patient trajectories with missing values." GigaScience Database 2019. http://dx.doi.org/10.5524/100662.