Artificial Intelligence in Medicine (2006) **36**, 245–255

ELSEVIER

ARTIFICIAL INTELLIGENCE IN MEDICINE

http://www.intl.elsevierhealth.com/journals/aiim

Neural networks for longitudinal studies in Alzheimer's disease

Reeti Tandon [a,*], **Sudeshna Adak** [a], **Jeffrey A. Kaye** [b]

[a] *Computational Biology and Biostatistics Lab, GE India Technology Center, EPIP Phase II, Hoodi Village, Whitefield Road, Bangalore, Karnataka 560066, India*
[b] *Layton Aging and Alzheimer's Disease Center, Oregon Health & Science University, 3181 SW Sam Jackson Park Road, Portland, OR 97201, USA*

Received 23 November 2004; received in revised form 20 October 2005; accepted 20 October 2005

KEYWORDS
Neurodegenerative diseases;
Longitudinal;
Random effects;
Mixed effects;
Misclassification;
Disease course;
Prognosis

Summary

Objective: Alzheimer's disease affects a growing population of elderly people today. The predictions about the course of the disease is a key component of health care decision making for patients with Alzheimer's. The physician's prognosis and predicted trajectory of cognitive decline often form the basis of treatment and health care decisions taken by patients and their families. These predictions are difficult to make because of the high variability and non-linearity exhibited by individual patterns of cognitive decline. This paper presents a new method of predicting the course of a disease using longitudinal data collected through multiple clinic visits. Longitudinal databases are similar to temporal databases, with some important differences—data is collected at irregular time intervals that are patient specific and also a varying number of observations are made for each patient, depending upon the number of times the patient visited the clinic. We propose a new type of neural network called the mixed effects neural network (MENN) model that can incorporate this type of longitudinal information.
Material and methods: We have used longitudinal data on 704 subjects enrolled at the Layton aging and research center (LAARC) at Oregon Health and Science University. A back-propagation algorithm, modified for longitudinal data is used to obtain the weight parameters of the MENN. The modified back-propagation algorithm is further embedded in an iterative procedure that estimates the noise variance and the parameters that capture the longitudinal (temporal) correlation structure.

* Corresponding author. Tel.: +91 80 5032411; fax: +91 80 8413189.
E-mail address: reeti.tandon@geind.ge.com (R. Tandon).

0933-3657/$ – see front matter
doi:10.1016/j.artmed.2005.10.007

Results: We have compared the performance of the MENN with linear mixed effects models and standard neural networks (NN). MENN show better performance (misclassification rate = 0.13 and relative MSE = 0.35) as compared to standard NN (misclassification rate = 0.34 and relative MSE = 2.74) and linear mixed effects models (misclassification rate = 0.14 and relative MSE = 0.4).
Conclusion: The results show that this method can be a useful tool for predicting non-linear disease trajectories and uncovering significant prognostic factors in longitudinal databases.

1. Introduction

The role of artificial neural networks in medical diagnosis, prognosis, treatment and clinical decision support has been well established since the earliest days of computing [1]. Neural networks have been applied across a wide variety of medical disciplines including oncology, cardiology, intensive care, radiology, neurology and many others. Fields of application have ranged from modelling the learning process of the brain [2] and signal processing to classification, prediction and survival analysis. Examples include the use of a neural network in review of pap smear slides using PAPNET testing [3], predicting outcomes of intensive care patients using the APACHE system [4], modelling cancer survival [5], classifying Alzheimer's disease and vascular dementia using analysis of brain SPECT image data [6], etc. For a more in-depth look at the uses and benefits of neural networks in medicine, we refer the reader to comprehensive reviews by various authors [7—10].

Neural networks that use temporal information to capture variations in non-stationary processes like radar, signals from the engine of an automobile, etc. have also been well studied. For these kinds of applications, temporal inputs have been built into a neural network system by various methods including traditional ones such as [11]:

- use of time delays to incorporate short-term memory in the system;
- recurrent networks by adding local feedback at the level of single neuron in the network or global feedback encompassing the whole network.

However, most of the existing neural network architectures do not model longitudinal data arising from multiple clinic visits by the same patient. Longitudinal data is a special class of temporal data where multiple measurements are made over time for each patient as and when the patient undergoes evaluation. The main difference between typical temporal databases like ECG [12] and longitudinal databases is that measurements in longitudinal data are made at unequal time intervals and each subject may not have an equal number of observations.

The modelling and analysis of longitudinal data has been a major challenge in clinical and epidemiological research and has resulted in many developments in statistical methodologies [13]. Current statistical models are geared towards generalizing regression methods to handle longitudinal data. They explicitly model individual trajectories and more fully exploit the information contained in repeated measurements on a patient collected from multiple clinic visits, and therefore, account for both intra- and inter-patient variability.

Learning from information collected over the disease course by following patients over time is the key to building accurate prognostic and predictive models. Predicting stages of many diseases accurately over time is still a challenge today especially given the non-linear progression of diseases. This article provides a new method for modelling the non-linear disease course. A simple artificial neural network model is not capable of handling longitudinal data which motivates us to propose a new method of "mixed effects neural networks" (MENN) which combines (1) temporal information arising from multiple clinic visits; (2) allowing for varying number of visits occurring at varying time intervals as and when patients visit the clinic; (3) neural networks that can flexibly model the response as a function of input variables. While there has been prior work in more parametric non-linear models for longitudinal data [13,14], this work proposes to integrate the more non-parametric flexible approach of neural networks with the capability of mixed effects models to handle longitudinal data. We describe an estimation procedure for obtaining the weights in the MENN and demonstrate its use for longitudinal clinical studies.

2. Longitudinal data

2.1. Alzheimer's disease study

Alzheimer's disease (AD) is a neurodegenrative disorder characterized by progressive decline in cognitive function. AD has come to be recognized as the

most common cause of functional disability among the elderly, affecting about 12% of individuals over the age of 65 years and as many as 40% of the individuals over the age of 85 years in the United States [15]. With an aging population, the number of AD cases are expected to quadruple in the next 50 years and can potentially become a formidable public health issue [16].

AD is an irreversible disease that slowly destroys memory and thinking skills. As the disease progresses, patients become increasingly dependent on their family caregivers who are all too often unprepared for their role [17]. Therefore, it is very important for physicians to be able to predict the time-course of the disease, i.e. predicting the cognitive and functional status as well as disease stage in the future.

Our study uses data from individuals enrolled in ongoing longitudinal studies at the Layton Aging and Alzheimer's Research Center (LAARC) at Oregon Health and Science University. There were a total of 4423 records from 704 subjects who were followed longitudinally from 1988 to 2002. Mean follow-up duration was 3.3 years (range 0.3—13.1 years).

Subjects entered into the LAARC database underwent standard neurological assessment, including a medical interview, and behavioral and cognitive status examinations at each visit. These examinations included the Mini-Mental State Examination (MMSE) [18] and the Clinical Dementia Rating Scale (CDR) [19]. The Mini-Mental State Examination is one of the most commonly used rating scales in assessing cognitive impairment. It is an 11-question measure that tests five areas of cognitive function: orientation, registration, attention and calculation, recall, and language. The different stages of dementia as defined by the MMSE score are shown in Table 1. The patients have been divided into four stages: normal (NOR), mild impairment (MILD), moderate impairment (MOD) and severe impairment (SEV) based on the MMSE score.

CDR is a measure of disease stage that is scored on a 5-point scale: 0 (normal), 0.5 (questionable dementia), 1 (mild dementia), 2 (moderate dementia) and 3 (severe dementia), and serves as an alternate measure for staging disease progression and incorporates both cognitive and functional status.

Table 1 MMSE score and dementia stages

MMSE score	Stage
24—30	Normal (NOR)
18—23	Mild impairment (MILD)
10—17	Moderate impairment (MOD)
<10	Severe impairment (SEV)

The factors associated with the changes in cognitive function over time such as age [20], gender, education [30], family history and base rate of progression [22] were collected at the time of enrollment. The base rate of progression is a measure that combines the information about the cognitive impairment at initial visit and the duration of symptoms and has been shown [22] to be predictor of cognitive decline. It is computed as: (expected MMSE score —baseline MMSE score at entry)/duration of symptoms in years, where expected MMSE score was obtained from age and education dependent population-based norms. This was used to stratify individuals as "non-progressors" (those with no symptoms at baseline), "slow progressors" (base rate of progression is 0—2 MMSE points per year) and "medium/fast progressors" (base rate of progression >2 MMSE points per year). A summary of these clinical and socio-demographic factors is shown in Table 2.

2.2. Linear mixed effects model for longitudinal data

Currently, most models for Alzheimer's disease are based on assuming that the MMSE score declines

Table 2 The factors used as inputs in the neural network models

Factor	Subtypes	N	%N
Gender	Female	392	56
	Male	312	44
Education	Less than 12 years	299	43
	Greater than 12 years	405	57
Live with status	Alone	200	28
	Not alone	504	72
Diagnosis at enrollment	Alzheimer's disease	336	48
	No dementia	308	44
	Questionable dementia	60	8
Base rate of progression	Non	250	22
	Slow	153	35
	Medium/fast	301	43
CDR	0	308	44
	0.5	150	21
	1	185	26
	2	53	8
	3	8	1
Age at enrollment		Mean = 78, S.D. = 10	

linearly [15,20,23,24] and that the intercept and slope follow a multivariate gaussian distribution. Thus, a standard model for MMSE is:

$$y_{ij} = a_i + s_i t_{ij} + \varepsilon_{ij}, \tag{1}$$

where

- y_{ij} is the response (MMSE score in our case study) observed for the i th patient at the j th clinic visit, $i = 1, 2, \ldots, M$; $j = 1, 2, \ldots, n_i$;
- t_{ij} is the time of the j th visit for the i th patient;
- the random intercept and slope $(a_i, s_i) \sim$ multivariate gaussian, where $E(a_i), E(s_i)$ are defined as functions of covariates of interest such as clinical and socio-demographic variables listed in Table 2.

These models are a special case of a general framework known as mixed effects models and are primarily used to describe relationship between a response variable and the input variables in longitudinal data. The linear mixed effects models can be denoted as [25]:

$$\begin{aligned} &\mathbf{y}_i = \mathbf{X}_i\boldsymbol{\beta} + \mathbf{Z}_i\boldsymbol{b}_i + \varepsilon_i, \quad i = 1, 2, \ldots, M \\ &\boldsymbol{b}_i \sim N_q(\mathbf{0}, \sigma^2 D), \quad \varepsilon_i \sim N_{n_i}(\mathbf{0}, \sigma^2 \mathcal{I}) \\ &\boldsymbol{b}_1, \boldsymbol{b}_2, \ldots, \boldsymbol{b}_M \quad \varepsilon_1, \varepsilon_2, \ldots, \varepsilon_M \text{ are independent} \end{aligned} \tag{2}$$

where

- $\mathbf{y}_i = (y_{i1}, y_{i2}, \ldots, y_{in_i})$is the MMSE scores for patient i collected in n_i visits;
- $\mathbf{X}_i$ is the $n_i \times p$ matrix corresponding to the input variables and $\boldsymbol{\beta}$ is a $p \times 1$ vector of "fixed effects" or population effects of the input variables;
- $\mathbf{Z}_i$ is the $n_i \times q$ design matrix corresponding to the "random effects" or subject-specific effects $\boldsymbol{b}_i$.

Assuming a linear trajectory with random intercept and slope, as in (1), would imply that the random effects matrix $\mathbf{Z}_i$ has two columns, the first being a column of 1's and the second being the times $(t_{ij};\ j = 1, 2, \ldots, n_i)$at which the visits occurred. The corresponding random effects coefficients $\boldsymbol{b}_i$ are the random subject-specific effect on the intercept and slope of cognitive decline. The input variables would be the covariates of interest that effect the slope and the intercept as well as their interactions (product) with time.

Note that (2) implies that

$$\mathbf{y}_i \sim N(\mathbf{X}_i\boldsymbol{\beta}, \sigma^2(I + \mathbf{Z}_i D \mathbf{Z}_i')) \tag{3}$$

which allows a general correlation structure between all the observations for a particular patient.

3. Methods

3.1. Non-linear disease course

Many longitudinal studies have examined the temporal course of AD using cognitive scores like MMSE and current models in these studies have usually assumed a linear rate of progression in predicting the time-course of AD [15,20,24]. The actual course of most of the diseases including AD is, however, non-linear with a long latent period of symptom development, followed by a period of obvious deterioration followed by a late plateau phase in the end stages of the disease. In AD, the middle phase is the period of rapid decline in cognitive, behaviorial and social functions and late plateau period where behavior and nursing are the primary concerns [26] as shown in Fig. 1.

Several models have been applied to understand the dynamics of the change associated with disease progression. For example, a trilinear model has been used to estimate at which point decline begins and ends [27]. However, ceiling and floor effects of the measures used to estimate severity limit the usefulness of most scales in the earliest and latest phases of the illness and may impact on such estimates of decline [28,29]. Other methods including growth curve approaches (S-shaped curves) have also been useful for representing the change that occurs in cognitive test performance over time [21]. Further, another analytic approach using non-parametric smoothing has shown that scores from several measures can be combined and related to a "time-index" estimation of dementia severity based on disease time-course or duration [31]. While the linear models are clearly inadequate in capturing the non-linear course of the disease, non-linear models have been traditionally of parametric form

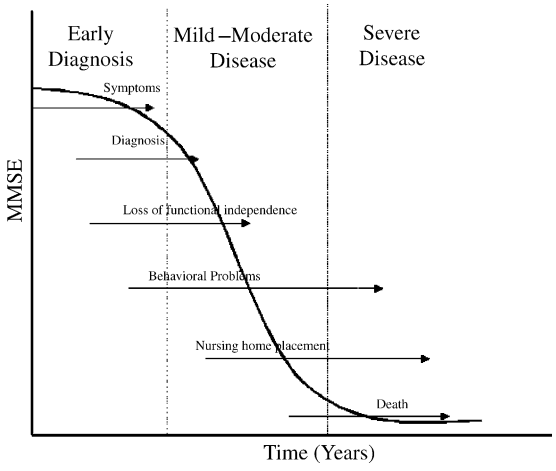

Figure 1 Non-linear trajectory of Alzheimer's disease.

(trilinear, S-shape, etc.). In the subsequent sections, we describe a more flexible method based on neural networks creating complex non-linear models. The new method described in the next section simultaneously adapts traditional neural networks to incorporate the generalized correlation structure of longitudinal data as in (3).

3.2. MENN

The MENN generalize the linear mixed effects models (2) by incorporating a general non-linear function of the input variables. The MENN model can be generally denoted as:

$$\mathbf{y}_i = f(\mathbf{X}_i, \boldsymbol{\beta}) + \mathbf{Z}_i \boldsymbol{b}_i + \varepsilon_i, \quad i = 1, 2, \ldots, M \tag{4}$$

where $\mathbf{X}_i, \boldsymbol{\beta}, \mathbf{Z}_i, \boldsymbol{b}_i, \varepsilon_i$ are as in (2) and f represents the neural network with inputs $\mathbf{X}_i$.

A special case of the MENN is defined by assuming a non-linear trajectory of the form:

$$y_{ij} = f(a_i, s_i t_{ij}; \boldsymbol{\beta}) + Z_{ij} \boldsymbol{b}_i + \varepsilon_{ij} \tag{5}$$

where a_i and s_i are the linear functions of covariates and $Z_{ij} = [1 \;\; t_{ij}]$. For example, a simple non-linear mixed effects neural network model with no neuron in the hidden layer and one in the output layer with sigmoid transfer function is:

$$y_{ij} = \frac{1}{1 + e^{a_i + s_i t_{ij}}} + Z_{ij} \boldsymbol{b}_i + \varepsilon_{ij}$$

This is the type of S-shaped model that has been proposed for modelling AD [21].

3.3. Estimating the parameters of MENN

In the standard neural networks (NN), the weight parameters $\boldsymbol{\beta}$ are determined by minimizing a squared-error loss function. When the noise is assumed to be gaussian, this can also be interpreted as maximizing the likelihood to estimate the weight parameters $\boldsymbol{\beta}$. In the MENN, we generalize this likelihood principle to estimate the weight parameters $\boldsymbol{\beta}$. This is achieved by maximizing $L(\boldsymbol{\beta}, D, \sigma^2 | \mathbf{y})$ which is the joint likelihood defined by incorporating the random effects $\boldsymbol{b}_i$ as follows:

$$\begin{aligned} &L(\boldsymbol{\beta}, D, \sigma^2 | \mathbf{y}_1, \mathbf{y}_2, \ldots, \mathbf{y}_M) \\ &\quad = \prod_{i=1}^{M} p(\mathbf{y}_i | \boldsymbol{\beta}, D, \sigma^2) \\ &\quad = \prod_{i=1}^{M} \int p(\mathbf{y}_i | \boldsymbol{b}_i, \boldsymbol{\beta}, \sigma^2)\, p(\boldsymbol{b}_i | D, \sigma^2)\, \mathrm{d}\boldsymbol{b}_i \end{aligned}$$

Proposition 1.

$$\begin{aligned} &L(\boldsymbol{\beta}, D, \sigma^2 | \mathbf{y}_1, \mathbf{y}_2, \ldots, \mathbf{y}_M) \\ &= \prod_{i=1}^{M} \frac{1}{\sqrt{(2\pi\sigma^2)^{n_i} |D|}} \\ &\times \int \frac{\exp(-1/2\sigma^2 (\|\mathbf{y}_i - f(\mathbf{X}_i, \boldsymbol{\beta}) - \mathbf{Z}_i \boldsymbol{b}_i\|^2 + \boldsymbol{b}_i' D^{-1} \boldsymbol{b}_i))}{2\pi\sigma^2} \mathrm{d}\boldsymbol{b}_i \end{aligned}$$

The proof of proposition 1 follows from the assumptions that $\boldsymbol{b}_i$ and $\mathbf{y}_i | \boldsymbol{b}_i$ have gaussian distributions.

Proposition 2.

$$\begin{aligned} &L(\boldsymbol{\beta}, D, \sigma^2 | \mathbf{y}_1, \mathbf{y}_2, \ldots, \mathbf{y}_M) \\ &\quad = (2\pi\sigma^2)^{-N/2} \times \prod_{i=1}^{M} \mathrm{abs}\left(\frac{|\Delta|}{|R_{11(i)}|} \right) \\ &\qquad \times \exp\left[\frac{-1}{2\sigma^2} \sum_{i=1}^{M} \| Q_{12(i)}' (\mathbf{y}_i - f(\mathbf{X}_i, \boldsymbol{\beta})) \|^2 \right] \end{aligned} \tag{6}$$

where Δ is a matrix such that $\Delta'\Delta = D^{-1}$, $N = \sum n_i$, and $Q_{12(i)}$ and $R_{11(i)}$ are matrices derived from a QR-decomposition involving $\mathbf{Z}_i$ and D as described in Appendix A.

Proof. Given in Appendix A. □

Note that in the special case described in (5), $\mathbf{Z}_i$ is a function of visit times t_{ij}. So, $Q_{12(i)}$ and $R_{11(i)}$ incorporate longitudinal information through

1. the times at which the i th patient visited the clinic;
2. D that captures the correlation structure over time.

Proposition 2 has a very important consequence in the estimation procedure. For a given value of D and σ^2, maximizing the likelihood (6) to obtain the estimates of $\boldsymbol{\beta}$ reduces to

$$\hat{\boldsymbol{\beta}} | D, \sigma^2 = \arg \min_{\boldsymbol{\beta}} \| Q_{12(i)}' (\mathbf{y}_i - f(\mathbf{X}_i, \boldsymbol{\beta})) \|^2, \tag{7}$$

everything else being constant. So, the weight parameters in the MENN, $\boldsymbol{\beta}$ can be estimated by minimizing the modified "mean squared error" criterion in the above expression (7). Note that this criterion is a modification of the standard mean squared error criterion used in a neural network and the modification involves the incorporation of longitudinal information through the patient-specific $Q_{12(i)}$ matrix.

Once $\hat{\boldsymbol{\beta}}$ is obtained, these estimates are used in deriving optimal estimates of D and σ^2, and this is repeated recursively till convergence. For the current estimate of $\boldsymbol{\beta}$ in the recursive algorithm, we can

derive the explicit expression for the optimal value of σ^2, conditional on D, as:

$$\hat{\sigma}^2|D,\hat{\boldsymbol{\beta}} = \sum_{i=1}^{M} \frac{\|Q'_{12(i)}(\mathbf{y}_i - f(\mathbf{X}_i,\hat{\boldsymbol{\beta}}))\|^2}{N} \tag{8}$$

It remains to estimate D given a current estimate of the parameters $\hat{\boldsymbol{\beta}}$ and $\hat{\sigma}^2$. Once we get the estimates of $\boldsymbol{\beta}$ and σ^2, (6) reduces to,

$$(2\pi\hat{\sigma}^2)^{-N/2} \times \prod_{i=1}^{M} \text{abs}\left(\frac{|\Delta|}{|R_{11(i)}|}\right) \times \exp\left[\frac{-1}{2\hat{\sigma}^2}\sum_{i=1}^{M}\|r_i\|^2\right] \tag{9}$$

r_i being the residuals obtained from the MENN model and defined as $Q'_{12(i)}(\mathbf{y}_i - f(X_i,\hat{\boldsymbol{\beta}}))$. The expression (9) looks exactly the same as the equation for linear models, except that the residuals of the linear model being replaced by the residuals from mixed effect model. Therefore, we have adapted the expectation maximization-based algorithms [13,32] for estimating D as in a linear mixed effects models.

The algorithm implementing this recursive maximization of the likelihood to obtain the weight parameters $\boldsymbol{\beta}$ and the noise parameter σ^2 as well as the correlation parameters D is described below.

Iterative estimation algorithm. The minimization over $\boldsymbol{\beta}$ and $\boldsymbol{b}_i$ is done in following steps:

1. Prepare an initial Δ matrix.
2. Repeat the following steps until norm $\left(\frac{\Delta_{\text{new}}-\Delta_{\text{old}}}{\Delta_{\text{old}}}\right) < 0.001$:
   - prepare $[\mathbf{Z}_i' \ \Delta']'$ matrix as described in Appendix A and perform its QR decomposition;
   - get the estimates for $\boldsymbol{\beta}$ by minimizing $\|Q'_{12(i)}(\mathbf{y}_i - f(\mathbf{X}_i,\boldsymbol{\beta}))\|^2$;
   - calculate new σ^2 using Eq. (7);
   - calculate the new Δmatrix as described in Section 3.3.1;
   - update the Q and R matrix.
3. Use this new Δ matrix to calculate the random effects.

3.3.1. The algorithm for estimation of D

Instead of directly estimating D, we estimate $\boldsymbol{\theta}$ which is a parametrization of Δ that depends upon the correlation structure assumed [13]. In our case, we have assumed that the random effects are uncorrelated, and thus

$$\Delta = \begin{bmatrix} \theta_1 & 0 \\ 0 & \theta_2 \end{bmatrix}$$

The estimates of $D$are obtained by maximizing the following condition:

$$\max_{\theta}\left(M\log|D^{-1}| - \sum_{i=1}^{M}\frac{1}{\sigma^2}E(\boldsymbol{b}_i'D^{-1}\boldsymbol{b}_i)\right) \tag{10}$$

3.3.2. Starting values of θ parameters

We generate simple starting values of $\theta^{(0)}$ as a diagonal matrix where each diagonal element is some fraction λ of root-mean-square length of corresponding column of $\mathbf{Z}_i$ matrices. That is, letting $\mathbf{Z}_i(k)$ denote the k th column of $\mathbf{Z}_i$, the initial value for k th diagonal element of Δ is $\lambda \times (\sum_{i=1}^{M}\|\mathbf{Z}_i(k)\|^2/M)^{1/2}$. The value of f used is 0.375.

4. Predictions

In a mixed effects model, predictions are made at two levels: the population level and at the subject-specific level. Population level predictions in a linear mixed effects model are given by $\mathbf{y}_i = \mathbf{X}_i\hat{\boldsymbol{\beta}}$ and predicted values for the subject, commonly known as best linear unbiased predictors (BLUPs) are defined as $\mathbf{y}_i = \mathbf{X}_i\hat{\boldsymbol{\beta}} + \mathbf{Z}_i\hat{\boldsymbol{b}}_i$ where $\hat{\boldsymbol{\beta}}$ and $\hat{\boldsymbol{b}}_i$ are the estimated values of $\boldsymbol{\beta}$ and $\boldsymbol{b}_i$.

In the MENN model, we similarly define population level and subject-specific level predictions. The population level predictions are based on the estimated weight parameters $\hat{\boldsymbol{\beta}}$ defined as in a standard NN and

$$\hat{\boldsymbol{y}}_i = f(\mathbf{X}_i,\hat{\boldsymbol{\beta}})$$

and the subject-specific level predictions are defined as

$$\hat{\boldsymbol{y}}_i = f(\mathbf{X}_i,\hat{\boldsymbol{\beta}}) + \mathbf{Z}_i\hat{\boldsymbol{b}}_i, \text{ where}$$
$$\hat{\boldsymbol{b}}_i = (\mathbf{Z}_i'\mathbf{Z}_i + \hat{D}^{-1})^{-1}\mathbf{Z}_i'(\mathbf{y}_i - f(\mathbf{X}_i,\hat{\boldsymbol{\beta}}))$$

5. Results and discussion

The neural networks system computes the Mini-Mental State Examination score for a particular visit of a patient. The inputs to the MENN are the covariates (fixed effects) along with the visit times for the patient as listed in Table 2.

5.1. Selection of the number of neurons

A neural network architecture was selected by choosing an optimal number of neurons in the hidden layer. The common parameters of the networks employed during training are listed in Table 3.

Table 3 Neural network architecture parameters

Network topology	Multilayer perceptron
Learning algorithm	Levenberg-Marquardt
Learning rule	Generalized delta rule
Output data	MMSE score
Learning constant	0.15
Momentum constant	0.4
Number of epochs	1000

The number of elements in the hidden layer were increased from 0 to 4 (adding one element at a time) until there was no further improvement in the network performance. The network with no hidden layer gave the best performance, so it was selected for the rest of the analysis. The transfer function used for architectures with 1 or more hidden neurons is "logsig" (sigmoid with output between 0 and 1) in the hidden layer while that in the output layer is "purelin" (linear). In the architecture with no hidden neuron "logsig" was used in the outer layer. All the inputs were scaled in the range of −1to 1. We used a standard training set (90% of the dataset) to train the network and test set, and test set (10% of the dataset) used to validate its predictive performance.

Using the selected architecture, the standard NN and the MENN were compared to a linear mixed effects model. Then, we have used a 10-fold "cases" cross validation to measure the performance of MENN. This is the same as 10-fold cross validation except that the partitions are made based on the patients and then data from all the visits are placed in whichever set (training or test) that the patient is placed into. The final results are obtained by averaging over the 10 validation sets for linear mixed effects models, standard NN and MENN to compare the predictive performance of the models.

5.2. Comparison of standard NN, MENN and linear mixed effects model

We have used following metrics to evaluate the networks predictive performance.

- Relative mean squared error (relative MSE)—is the mean square error from the model in comparison to what we would get from the naive model which uses the MMSE from the previous visit. So, we take the ratio of the MSE in predictions from our model to the MSE from using the last visit as our prediction. So, if the relative MSE is much less than 1, our model is good. This provides a good metric of comparison of all the models to the common "last-visit" model.
- Misclassification error—is the misclassification of predicted disease stage in comparison to actual disease stage. The disease stages are defined by the MMSE stage. Thus, the misclassification error estimates the percentage of people who belonged to one disease stage but were predicted to be at another disease stage due to our models.

R^2-value and MSE (which are more standard in the neural networks literature) were also computed. However, since the main aim of the model is MMSE prediction and our objective is to compare the accuracy of the predictions from MENN with those of the naive model predictions, we use relative MSE and misclassification error as metrics for assessing the performance of the various models.

Fig. 2 compares the predictions of standard NN, linear mixed effects models and MENN for a patient. This selected case is an example of a situation where the MENN are able to capture the non-linearity in decline of MMSE much better than all the other models.

Table 4 compares the above performance metrics of the three models—linear mixed models, standard NN and MENN.

The above results show that MENN performed better than the standard NN and linear mixed effects models in terms of relative mean squared error and misclassification rate.

MENN predicts disease stage better than linear mixed effects model but when we look across the 10-cross validation sets there is a high degree of variability. Fig. 3 shows the misclassification rate as observed across the 10-cross validation sets.

The misclassification rate varies from as low as 5.5% to as high as 18.5% (range = 13%) across the 10-cross validation sets. We explored the potential causes of this high variability namely fluctuations,

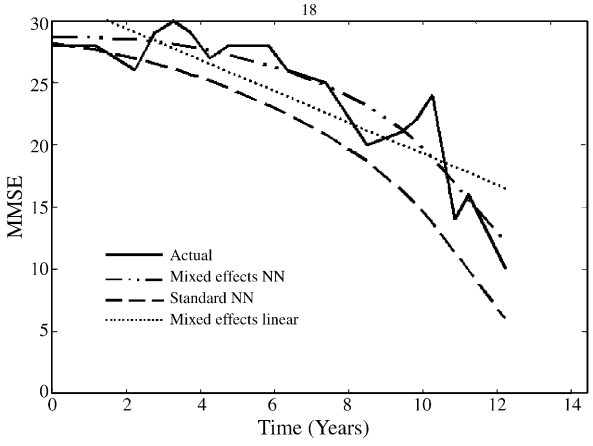

Figure 2 MMSE decline for a patient-comparison of linear mixed effects models, standard NN and MENN.

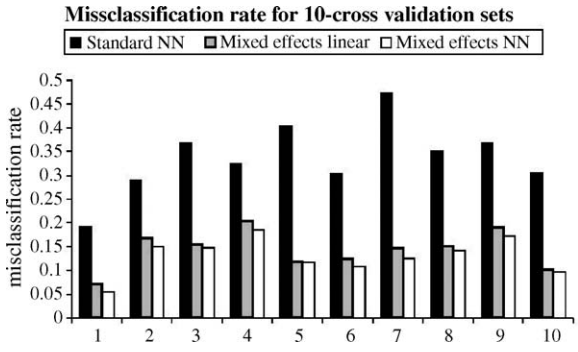

Figure 3 Misclassification rate for 10-cross validation sets.

linearity of the profile, effect of gender or age of enrollment. The main reason for this high variability accounted for by the fact that the actual MMSE scores in the data have a lot of fluctuations. R^2-value was computed for the straight line fit to an individuals MMSE profile. High fluctuations were defined as deviation from this straight line fit. All the patients having R^2 less than 60% were defined to have fluctuating MMSE score. By this definition 22% of the cases had fluctuating MMSE scores.

We performed three experiments to test the effect of fluctuations:

- Experiment 1: All the cases with fluctuating MMSE scores in training set. The remaining cases in the training set were randomly chosen from the remaining non-fluctuating cases.
- Experiment 2: We randomly selected 10% of the fluctuating cases as the test set. All remaining cases were placed in the training set.
- Experiment 3: 50% of fluctuating cases in training set and 50% in validation set.

We found that the standard deviation of the misclassification rate across the 10-cross validation sets can be reduced to 0.01 (12% of the mean) if the patients are grouped such that all the patients who have fluctuating MMSE score go in the validation set, while the standard deviation becomes as high as 0.04 (29% of the mean) when the fluctuating cases are distributed randomly across the training and test set. The misclassification rate across 10-cross validation sets varied from 6.2 to 10%, the range being 3.8% which is quite low as compared to 13% when the fluctuating cases are randomly distributed. Note that the average misclassification rate for the MENN was reduced from 12.9% (as reported in Table 4) to 7.8%. This is still an improvement over the linear mixed effects model which had an average misclassification rate of 8.5% in this controlled experiment. Note that in all the three experiments the MENN performed better than linear mixed effects models with regards to relative MSE and misclassification rate as metrics 5.

6. Conclusions and future work

In this paper, we have introduced a new methodology that uses neural networks applied to longitudinal data for modelling the time-course of a disease. Longitudinal databases are the most common type of databases seen in chronic illnesses like Alzheimer's, where the data is collected from multiple clinic visits by a cohort of patients. The MENN models present a new paradigm for AI modelling and a new approach for more flexible models in comparison to current statistical techniques. There were two main benefits using the MENN models on the Alzheimer's disease longitudinal studies:

- Alzheimer's disease is one of the most devastating illnesses that affects the elderly. The trajectory of cognitive decline is clearly non-linear, highly variable and dependent on a variety of factors that are not clearly understood today. The flexibility of neural networks can be used effectively to address these problems easily with the modifications necessary to handle longitudinal data.
- As research on Alzheimer's disease has increased over the past two decades, the amount of longitudinal data has also increased. However, the modelling methodologies used in clinical research as well in clinical practice are still very simplistic and linear. MENN present a new modelling paradigm that can be shown to be much more accurate and effective as compared to models in current practice.

Table 4 Comparison of standard NN, linear mixed models and MENN

Model	Linear mixed effects model	Standard NN	MENN
R^2	0.94 (0.03)	0.81 (0.07)	0.95 (0.02)
MSE	0.2 (0.16)	7.06 (7.33)	0.37 (0.34)
Relative MSE	0.4 (0.11)	2.74 (0.77)	0.35 (0.11)
Misclassification rate	0.14 (0.04)	0.34 (0.07)	0.13 (0.04)

Numbers are reported as "mean (standard deviation)" for each model.

Table 5 Comparison of standard NN, linear mixed models and MENN in three experiments for testing the effect of fluctuations

	LMEM	Standard NN	MENN
Experiment 1			
Relative MSE	0.44 (0.71)	2.65 (0.62)	0.41 (0.06)
Misclassification rate	0.11 (0.02)	0.27 (0.05)	0.09 (0.01)
Experiment 2			
Relative MSE	0.43 (0.02)	2.76 (0.41)	0.4 (0.01)
Misclassification rate	0.09 (0.01)	0.25 (0.03)	0.08 (0.01)
Experiment 3			
Relative MSE	0.38 (0.05)	3.1 (1.6)	0.33 (0.04)
Misclassification rate	0.16 (0.29)	0.39 (0.05)	0.15 (0.25)

Numbers are reported as "mean (standard deviation)" for each model. Experiment 1 is when 50% of the fluctuating cases are in the training set and 50% in validation set. Experiment 2 contains all the cases in validation set and Experiment 3 contains all the cases in the training set. LMEN is linear mixed effects models, stdNN stands for standard NN.

Our results were substantiated by extensive experiments using 10-fold cross validation and careful consideration of the issues that arise in Alzheimer's such as fluctuating cognitive status. Cognitive examination is based on written questionnaires and are subject to fluctuations arising from reasons not related to the disease (mood on the particular day, general health on the particular day, etc.). We have shown that careful consideration of fluctuations in this type of cognitive data is necessary in training networks that are more accurate.

While this methodology has been applied to Alzheimer's disease, it is a general research methodology that can be applied to other chronic illnesses such as Parkinson's and even to non-medical applications which have longitudinal data. We are aiming to evaluate this methodology on cancer related data and to extend the current methods to more general and non-linear random effects models.

We are also exploring extensions of MENN to handle discrete responses such as disease stage. This would allow use of neural network classifiers with longitudinal data.

Acknowledgements

The authors gratefully acknowledge the helpful advice and comments of Mr. Sundararajan Ramsubramaniam and Mr. William Gorman.

Supported in part by a grant PHS 5 MO1 RR00334 (Kaye) from the National Institutes of Health, Bethesda, MD and AG 08017 (Kaye) from the National Institute of Aging, Bethesda, MD and the Department of Veterans Affairs Research Service Merit Award (Kaye).

Appendix A

Proof of proposition 2: It follows from Proposition 1 that

$$L(\boldsymbol{\beta}, D, \sigma^2 | \boldsymbol{y}_1, \boldsymbol{y}_2, \ldots, \boldsymbol{y}_M) = \prod_{i=1}^{M} \frac{1}{\sqrt{(2\pi\sigma^2)^{n_i}|D|}} \times \int \frac{\exp(-1/2\sigma^2(\|\boldsymbol{y}_i - f(\mathbf{X}_i, \boldsymbol{\beta})\mathbf{Z}_i\boldsymbol{b}_i\|^2 + \boldsymbol{b}_i' D^{-1}\boldsymbol{b}_i))}{(2\pi\sigma^2)^{q/2}} \mathrm{d}\boldsymbol{b}_i$$

Since $\Delta'\Delta = D^{-1}$ the expression in the exponent becomes

$$\|\mathbf{y}_i - f(\mathbf{X}_i, \boldsymbol{\beta}) - \mathbf{Z}_i\boldsymbol{b}_i\|^2 + \boldsymbol{b}_i' D^{-1}\boldsymbol{b}_i = \|\tilde{\mathbf{y}}_i - \tilde{f}(\mathbf{X}_i, \boldsymbol{\beta}) - \tilde{\mathbf{z}}_i\boldsymbol{b}_i\|^2 \quad (11)$$

where

$$\tilde{\mathbf{y}}_i = \begin{bmatrix} \mathbf{y}_i \\ 0 \end{bmatrix} \tilde{f}(\mathbf{X}_i, \boldsymbol{\beta}) = \begin{bmatrix} f(\mathbf{X}_i, \boldsymbol{\beta}) \\ 0 \end{bmatrix} \tilde{\mathbf{z}}_i = \begin{bmatrix} \mathbf{Z}_i \\ \Delta \end{bmatrix}$$

We perform triangular decomposition of $\tilde{\mathbf{z}}_i$

$$\begin{bmatrix} \mathbf{Z}_i \\ \Delta \end{bmatrix} = Q_{(i)} \begin{bmatrix} R_{11(i)} \\ 0 \end{bmatrix} \quad (12)$$

where $Q_{(i)}$ is a $(n_i + q) \times (n_i + q)$ orthogonal matrix and $R_{11(i)}$ is an upper-triangular 2×2 matrix, then property of orthogonal matrix ensure that

$$\|\tilde{\mathbf{y}}_i - \tilde{f}(\mathbf{X}_i, \boldsymbol{\beta}) - \tilde{\mathbf{z}}_i\boldsymbol{b}_i\|^2 = \|Q_{(i)}'(\tilde{\mathbf{y}}_i - \tilde{f}(\mathbf{X}_i, \boldsymbol{\beta}) - \tilde{\mathbf{z}}_i\boldsymbol{b}_i)\|^2 \quad (13)$$

$Q_{(i)}$ is partitioned as

$$\begin{bmatrix} Q_{11(i)} & Q_{12(i)} \\ Q_{21(i)} & Q_{22(i)} \end{bmatrix}$$

where $Q_{11(i)}$ is a $n_i \times q$ matrix and $Q_{12(i)}$ is a $n_i \times n_i$ matrix.

Substituting the above results in (13), we get,

$$\left\| \begin{bmatrix} Q'_{11(i)}\mathbf{y}_i \\ Q'_{12(i)}\mathbf{y}_i \end{bmatrix} - \begin{bmatrix} Q'_{11(i)}\, f(\mathbf{X}_i,\boldsymbol{\beta}) \\ Q'_{12(i)}\, f(\mathbf{X}_i,\boldsymbol{\beta}) \end{bmatrix} - \begin{bmatrix} R_{11(i)} \\ 0 \end{bmatrix} \boldsymbol{b}_i \right\|^2$$
$$= \|Q'_{11(i)}\mathbf{y}_i - Q'_{11(i)}\, f(\mathbf{X}_i,\boldsymbol{\beta}) - R_{11(i)}\boldsymbol{b}_i\|^2 + \|Q'_{12(i)}\mathbf{y}_i - Q'_{12(i)}\, f(\mathbf{X}_i,\boldsymbol{\beta})\|^2 \quad (14)$$

The integral in Eq. (11); therefore, reduces to

$$\int \frac{\exp(-1/2\sigma^2(\|\mathbf{y}_i - f(\mathbf{X}_i,\boldsymbol{\beta}) - \mathbf{Z}_i\boldsymbol{b}_i\|^2 + \boldsymbol{b}'_i D^{-1}\boldsymbol{b}_i))}{(2\pi\sigma^2)} \mathrm{d}\boldsymbol{b}_i$$
$$= \exp\left[\frac{-\|Q'_{12(i)}\mathbf{y}_i - Q'_{12(i)} f(\mathbf{X}_i,\boldsymbol{\beta})\|^2}{2\sigma^2}\right]$$
$$\times \int \exp \frac{[-1/2\sigma^2 \|Q'_{11(i)}\mathbf{y}_i - Q'_{11(i)} f(\mathbf{X}_i,\boldsymbol{\beta}) - R_{11(i)}\boldsymbol{b}_i\|^2]}{2\pi\sigma^2} \mathrm{d}\boldsymbol{b}_i \quad (15)$$

Because $R_{11(i)}$ is non-singular, we can perform a change of variable

$$\phi_i = \frac{(Q'_{11(i)}\mathbf{y}_i - Q'_{11(i)}\, f(\mathbf{X}_i,\boldsymbol{\beta}) - R_{11(i)}\boldsymbol{b}_i)}{\sigma}$$

with the differential $\mathrm{d}\phi_i = \sigma^{-q/2}\mathrm{abs}|R_{11(i)}|\,\mathrm{d}\boldsymbol{b}_i$ and write the integral as

$$\frac{1}{(2\pi\sigma^2)^{q/2}} \times \int \exp\left[\frac{-1}{2\sigma^2}\|Q'_{11(i)}\mathbf{y}_i - Q'_{11(i)}\, f(\mathbf{X}_i,\boldsymbol{\beta}) - R_{11(i)}\boldsymbol{b}_i\|^2\right] \mathrm{d}\boldsymbol{b}_i$$
$$= \frac{1}{(2\pi)^{q/2}|R_{11(i)}|} \int \exp\left[\frac{-1}{2}\|\phi\|^2\right] d\phi_i = |R_{11(i)}|^{-1} \quad (16)$$

Substituting Eq. (16) in (11) provides likelihood as

$$L(\boldsymbol{\beta}, D, \sigma^2|\mathbf{y}) = \prod_{i=1}^{M} \frac{\mathrm{abs}|R_{11(i)}|^{-1}}{\sqrt{(2\pi\sigma^2)^{n_i}|D|}} \times \exp\left[\frac{-1}{2\sigma^2}\|Q'_{12(i)}\mathbf{y}_i - Q'_{12(i)}\, f(\mathbf{X}_i,\boldsymbol{\beta})\|^2\right] \quad (17)$$

From the definition of *Delta*, it is clear that $\frac{1}{\sqrt{|D|}} = \mathrm{abs}|\Delta|$. Hence, the (17) reduces to

$$L(\boldsymbol{\beta}, D, \sigma^2|\mathbf{y}) = (2\pi\sigma^2)^{-N/2} \times \prod_{i=1}^{M} \mathrm{abs}\left(\frac{|\Delta|}{|R_{11(i)}|}\right) \times \exp\left[\frac{-1}{2\sigma^2}\sum_{i=1}^{M}\|Q'_{12(i)}(\mathbf{y}_i - f(\mathbf{X}_i,\boldsymbol{\beta}))\|^2\right] \quad (18)$$

where $N = \sum_{i=1}^{M} n_i$ is the total number of observations.

References

[1] Baxt WG. Application of artificial neural networks to clinical medicine. Lancet 1995;346:1135–8.

[2] Guigon E, Dorizzi B, Burnod Y, Schultz W. Neural correlates of learning in the prefrontal cortex of the monkey: a predictive model. Cerebr Cortex 1995;5(2):135–47.

[3] Rosenthal DL, Mango LJ, Acosta D, Peters RK. "Negative" Pap smears preceding carcinoma of the cervix rescreening with the PAPNET system. Am J Clin Pathol 1993;100:331.

[4] Rowan KM, Kerr JH, Major E, McPherson K, Short A, Vessey MP. Intensive care Society's Acute Physiology and Chronic Health Evaluation (APACHE II) study in Britain and Ireland: a prospective, multicenter, cohort study comparing two methods for predicting outcome for adult intensive care patients. Am J Respir Crit Care Med 1994;22:1392–401.

[5] Burke HB, Goodman PH, Rosen DB, Henson DE, Weinstein JN, Harrell Jr FE, et al. Artificial neural networks improve the accuracy of cancer survival prediction. Cancer 1997;79(4):857–62.

[6] Defigueiredo RJ, Shankle WR, Maccato A, Dick MB, Mundkur P, Mena I, et al. Neural-network-based classification of cognitively normal, demented, Alzheimer disease and vascular dementia from single photon emission with computed tomography image data from brain. Proc Natl Acad Sci 1995;92(12):5530–4.

[7] Dybowski R. Neural computation in medicine: perspectives and prospects. In: Malmgren H, Borga M, Niklasson L, editors. Proceedings of the ANNIMAB-1 Conference (Artificial Neural Networks in Medicine and Biology), Goteborg. Springer–Verlag; 13–16 May 2000. pp. 26–36.

[8] Lisboa PJ. A review of evidence of health benefit from artificial neural networks in medical intervention. Neural Netw 2002;15(1):11–39.

[9] Papik K, Molnar B, Schaefer R, Dombovari Z, Tulassay Z, Feher J. Application of neural networks in medicine—a review. Diagn Med Tech 1998;4:538–56.

[10] Reggia JA. Neural computation in medicine. Artif Intell Med 1993;5:143–57.

[11] Haykins S. Neural networks: a comprehensive foundation. Pearson Education Asia; 1999.

[12] Zitar Abu R, Cascaded neural network for classification of artificially modeled ECG beats using error signal, proceedings of International Conference on artificial intelligence. Las Vegas, Nevada, USA: Springer–Verlag; June 2000. pp. 1594–98.

[13] Pinheiro JC, Bates DM. Mixed-effects models in S and S-Plus. New York: Springer, 2000.

[14] Vonesh EF, Carter RL. Mixed effects nonlinear for unbalanced repeated measures. Biometrics 1992;48:1–17.

[15] Kukull WA, Higdon R, Bowen JD, McCormick WC, Teri L, Schellenberg GD, et al. Dementia and Alzheimer disease incidence: a prospective cohort study. Arch Neurol 2002;59(11):1737–46.

[16] Brookmeyer R, Gray S, Kawa C. Projections of Alzheimer's disease in the United States and the public health impact of delaying disease onset. Am J Public Health 1998; 88(9):1337–42.

[17] Gelb DJ. Measurement of progression in Alzheimer's disease: a clinician's perspective. Stat Med 2000;19(11–12):1393–400.

[18] Folstein MF, Folstein SE, McHugh PR. "Mini-mental state". A practical method for grading the cognitive state of patients for the clinician. J Psychiatr Res 1975;12(3):189–98.

[19] Morris JC. The Clinical Dementia Rating (CDR): current version and scoring rules. Neurology 1993;43:2412–4.

[20] Lautenschlager NT, Cupples LA, et al. Risk of dementia among relatives of Alzheimer''s disease patients in the MIRAGE study: what is in store for the oldest old? Neurology 1996;46(3):641—50.
[21] Stern Y, Liu X, Albert M, Brandt J, Jacobs DM, Del Castillo-Casteneda C, et al. Application of a growth curve approach to modeling the progression of Alzheimer's disease. J Gerontol Med Sci 1996;51:M179—84.
[22] Doody RS, Massman P, Dunn K. Method for estimating progression rates in Alzheimer disease. Arch Neurol 2001; 58:449—54.
[23] Mendiondo MS, Ashford JW, Kryscio RJ, Schmitt FA. Modelling Mini-Mental State Examination changes in Alzheimer's disease. Stat Med 2000;19(11—12):1607—16.
[24] Milliken JK, Edland SD. Mixed effects models of longitudinal Alzheimer's disease data: cautionary note. Stat Med 2000;1617—29.
[25] Laird NM, Ware JH. Random-effects models for longitudinal data. Biometrics 1982;38:963—74.
[26] Ashford JW, Schmitt FA. Modeling the time-course of alzheimer dementia. Curr Psychiatry Rep 2001;3(1):20—8.
[27] Brooks JO, Kraemer HC, Tanke ED, Yesavage III JA. The methodology of studying decline in Alzheimer's disease. J Am Geriatr Soc 1993;41:623—8.
[28] Ashford JW, Kolm P, Colliver JA, Bekian C, Hsu L-N. Alzheimer patient evaluation and the Mini-Mental State: item characteristic curve analysis. J Gerontol Psychol Sci 1989;5:139—46.
[29] Fillenbaum GG, Wilkinson WE, Welsh KA, Mohs RC. Discrimination between stages of Alzheimer's disease with subsets of mini-mental state examination items. Arch Neurol 1994;51:916—21.
[30] Stern Y, Gurland B, Tatemichi TK, et al. Influence of education and occupation on the incidence of Alzheimer''s disease. JAMA 1994;271(13):1004—10.
[31] Ashford JW, Shan M, Butler S, Rajasekar A, Schmitt FA. Temporal quantification of Alzheimer's disease severity: 'Time-Index' model. Dementia 1995;6:269—80.
[32] Bates DM, Pinheiro JC. Computational methods for multilevel modelling, unpublished. http://cm.bell-labs.com/cm/ms/departments/sia/jcp/pub.html [last accessed October 20, 2005].