Assessing the predictive ability of the UPDRS for falls classification in early stage Parkinson’s disease

Sarini Abdullah¹, Nicole White², James McGree³, Kerrie Mengersen³, Graham Kerr²

¹ Department of Mathematics, University of Indonesia
² Institute of Health and Biomedical Innovation (IHBI), Australia
³ ARC Centre of Excellence for Mathematical and Statistical Frontiers, Queensland University of Technology (QUT), Australia

Abstract

Background: Falling is common for people with Parkinson’s disease (PD), with negative consequences in terms of quality of life. Therefore, the identification of risk factors associated with falls is an important research question. In this study, various ways of utilizing the Unified Parkinson’s Disease Rating Scale (UPDRS) were assessed for the identification of risk factors and for the prediction of falls.

Methods: Three statistical methods for classification were considered: decision trees, random forests, and logistic regression. For variable selection, the stepwise procedure and Bayesian model averaging based on log-marginal likelihood were implemented for logistic regression, and Gini index criterion was used for decision trees and random forests. UPDRS measurements on 51 participants with early stage PD, who completed monthly falls diaries for 12 months of follow-up were analyzed.

Results: All classification methods applied produced similar results in regards to classification accuracy and the selected important variables. The highest classification rates were obtained from model with individual items of the UPDRS with 80% accuracy (85% sensitivity and 77% specificity), higher than in any previous study. A comparison of the independent performance of the four parts of the UPDRS revealed the comparably high classification rates for Parts II and III of the UPDRS. Similar patterns with slightly different classification rates were observed for the 6- and 12-month of follow-up times. Consistent predictors for falls selected by all classification methods at two follow-up times are: thought disorder for UPDRS I, dressing and falling...
for UPDRS II, hand pronate/supinate for UPDRS III, and sleep disturbance and symptomatic orthostasis for UPDRS IV. While for the aggregate measures, subtotal 2 (sum of UPDRS II items) and bradykinesia showed high association with fall/non-fall.

**Conclusions:** Fall/non-fall occurrences were more associated with individual items of the UPDRS than with the aggregate measures. UPDRS parts II and III produced comparably high classification rates for fall/non-fall prediction. Similar results were obtained for modelling data at 6-month and 12-month follow-up times.

*Keywords:* Bayesian model averaging, decision trees, logistic regression, random forests, receiver operating characteristics (ROC), sensitivity, specificity.

---

1. **Introduction**

Falls are a significant and common problem for persons diagnosed with Parkinson’s disease (PD) [4, 31, 40], and are prominent even in the disease’s early stages [3, 24]. Several prospective studies have shown that falls incidence is relatively high among people with Parkinson’s (PWP), with estimates ranging from 46-72%, over three, six and twelve month periods [5, 19, 24, 28, 40]. PD fallers are also more likely to fall in the future [28, 40]. The negative consequences of falls for PWP quality of life [5, 23, 37] and associated health care costs [13] combined with its high prevalence has motivated the investigation of risk factors associated with their occurrence.

Prospective falls methods are the gold standard, in contrast to retrospective falls, for falls prediction and falls risk factors identification. The problem with looking at retrospective falls is that elderly may forget they have fallen. Despite this gold standard, there are relatively few prospective falls studies. Among the few are as in [10, 24, 15]. [24] pointed out the inconsistency in clinically useful falls risk factors in 7 prospective studies. In search for the effective way to predict falls, [24] put more attention to the functional tests and disease-specific clinical assessments and developed a multivariate predictive model. It is inferred that a combination of both disease-specific and balance-and mobility-related measures can accurately predict falls in PWP.

The Unified Parkinson’s Disease Rating Scale (UPDRS) and its recent revision, Movement Disorder Society-UPDRS (MDS-UPDRS) [18], is considered to be the gold standard instrument for the clinical assessment of
PD [34], due to its high reliability [25, 27] and thorough measurement of multiple factors including body structure and function, activity and participation [34]. For these reasons, the UDPRS has been utilised by a number of studies for falls prediction, as an overall measure of disease severity [10, 11, 24, 26, 28, 40]. These studies have consistently shown positive associations between falling and higher UPDRS scores, namely Part II and III subtotals and the overall (Part I-IV) sum. In addition, sums of different combination of UPDRS individual items are often obtained to give a single measure for a symptom, such as: tremor, rigidity, bradykinesia, and Postural Instability and Gait (PIGD). These composite measures also showed a reasonably positive association with falls. The use of these aggregate measures, however, ignores the contribution of individual UPDRS items to the prediction of falls risk, many of which are functionally relevant and potentially managed.

Logistic regression is a popular statistical tool for classification as a function of observed predictors. It has been extensively used for falls classification in PD [2, 24, 28, 35, 40]. The popularity of logistic regression is due to its ease of interpretation via odds ratios and the estimation of individual patient probabilities into faller and non-faller groups, for the purpose of deriving an optimal classification rule. However, the determination of the best subset of predictors for inclusion in logistic regression is challenging and, for this reason, analysis is often restricted to first-order (or linear) effects. Higher order effects, for example the interaction between predictors is ignored. Moreover, one guideline suggests that there should be at least 10 participants for each predictor [1] which is often difficult to fulfill in many PD data applications.

Bayesian model averaging (BMA) offers an appealing solution to optimal model selection by combining predictions from multiple models, with different subsets of predictors [20]. This technique applies concepts from Bayesian inference by weighting different models according to their posterior model probability and producing a consensus prediction for the outcome of interest. The appeal of BMA is driven by accounting for model uncertainty which here relates to whether covariates, interactions and high order terms should appear in the model. Although several health-related studies have implemented this approach [38, 39, 12], to the best of our knowledge, this paper is the first to implement BMA for Parkinson’s related study.

Decision trees (DTs) [29] and random forests (RFs) [8] are examples of nonparametric tree-based classification methods that naturally incorporate variable selection. The natural variable selection is owing to the tree con-
struction mechanism employed in easy method. A tree is grown by first selecting the most discriminating variable (called splitting variable) to partition the data into the target classes (child nodes). Then, the process is repeated in each of the child nodes, until a certain stopping criterion is reached. Thus, only important variables are used in decision making about the predicted classes.

Both methods employ recursive partitioning to automatically determine predictors that best discriminate between classes of the response variable, resulting in tree-like structures. The very nature of these tree-like structures accommodate complex interactions without suffering from the curse of dimensionality [14]. They have also been popularized due to their ease of interpretation. Using these methods, the aim of this paper was to evaluate the utility of the UPDRS for falls classification in people with early stage PD, with a view to identifying key predictors that contribute to this classification. This will provide useful information to clinicians as a more focused attention to the identified factors could provide a better guide in understanding the patients condition. Moreover, a quick and straightforward decision on the likely of falls in patients could be inferred using the “if then rules” in decision trees.

The remainder of this paper is organized as follows. A description of the data and methodology are provided in Section 2. Key results are presented in Section 3, including the comparison of individual UPDRS items versus composite measures, the relative importance of different UDPRS subsections, and the identification of key predictors. A discussion of results and limitations are presented in Section 4 and a summary of overall findings is given in Section 5.

2. Data and Methods

2.1. Participants

Fifty one participants diagnosed with idiopathic PD were recruited for this study, as part of a larger research project conducted by the Institute of Health and Biomedical Innovation in Brisbane, Australia [24]. All participants were classified as early stage PD, determined by a Hoehn and Yahr (HY) score of 3 or less.

Each participant completed a monthly falls diary over a consecutive period. Based on this information, a participant was classified as a faller if they had experienced at least one fall within a defined follow-up period. In
this study, follow-up times were defined at six and twelve months. Successful completion of the diary was monitored by phone calls and mail correspondence.

Disease severity was assessed at baseline using all four parts of the UPDRS: I (mentation, behaviour, mood), II (activities of daily living, ADL), III (motor function), and IV (complications of therapy). Subtotals I-IV were obtained by adding scores of individual items of the UPDRS in Parts I-IV, respectively. Composite UPDRS scores for tremor (items 20 and 21), rigidity (item 22), bradykinesia (items 23, 25, 26, and 31) and Postural Instability and Gait (PIGD, items 13, 14, 15, 27, 28, 29, 30) were also calculated. In this paper, subtotals and composite UPDRS scores are referred to as aggregate measures.

2.2. Classification methods

Four classification methods were evaluated for the identification of UPDRS-related factors associated with falls: decision trees (DTs), random forests (RFs), logistic regression with forward variable selection and logistic regression with Bayesian model averaging (BMA). In this section, key details of each method and selected criteria for model comparison are outlined.

Decision trees apply recursive partitioning to identify the subset of predictors that best discriminates observations into different categories of the outcome variable (faller/non-faller). A decision tree procedure begins by determining with predictor best splits the data into two nodes to minimise classification entropy [9, 17]. For continuous predictors, splits are in the form of an optimal cut-off (≥, ≤). Splits on categorical predictors are in the form of membership to a chosen category. This binary partitioning is then repeated on resulting groups until minimum criteria on node size and/or changes in the chosen misclassification criterion are met. A schematic of this process is provided in Figure 1a. Decision trees results in this paper were obtained using rpart package [36] in R 3.2.5 [30].

Random forests [8] (Figure 1b) offer a robust alternative to decision trees that incorporate bootstrapping and random predictor selection to reduce uncertainty in model predictions. This method fits multiple decision trees, where each tree is fitted to a random subset of the data, sampled with replacement. Within a single tree, splits are determined from a random subset of predictors sampled at each node, as a means of reducing correlation among predictors [3, 7]. A consensus prediction for a single observation is obtained
by combining predictions across trees; for categorical outcomes, the consensus prediction is the most commonly predicted classification over all trees. Random forests are therefore viewed as a form of model averaging [6]. Data processing for RF were conducted using randomForest package [9] in R 3.2.5 [30].

Turning to the regression based methods, logistic regression can be appropriate when the response variable is dichotomous. The underlying relationships between the explanatory and response variables can be explained by the regression model, through the regression coefficients $\beta = (\beta_0, ..., \beta_K)$ and covariate information $x = (x_1, ..., x_K)$, as

$$\logit(\pi_i) = \beta_0 + \beta_1 x_{i1} + ... + \beta_K x_{Ki},$$

where $\pi_i$ is the probability of fall for patient $i$ having $K$ measurement $x_i = (x_{i1}, ..., x_{Ki})$. The relative risk, or odds ratio, of being in one class of the response (i.e. observing a fall) based on a specified value of the explanatory variables, say $x_j$, can be predicted by taking the exponentiation of the corresponding regression coefficient, $e^{\beta_j}$. Odds ratio greater than one would suggest that the explanatory variable being considered is associated with increase of risk of getting an event, and the opposite is for odds ratio less than one. Odds ratio equal to one simply states non-association of the explanatory and response variables.

Within a Bayesian framework, one needs to specify the prior distribution for the model parameters. Here, each regression coefficient is assumed to follow a Gaussian distribution, $\beta_j \sim N(0, v_0)$. To represent a vague prior knowledge, $v_0$ is set to a large value (i.e. $10^3$ in R-INLA [32]).

When fitting a logistic regression model, it is necessary to only include the important explanatory variables in the model. This problem of variable selection is not adequately addressed in many PD studies mentioned earlier [2, 24, 28, 35, 40, 15, 10, 11]. Among the few that raised concern about selection of important variables is [24] that used the total scores (of different clinical instruments) rather than their component (or item) scores to avoid redundancy in the variables used. While, [24] included variables that were significant in the univariate model. However, in the presence of other covariates, the contribution of a predictor variable to the prediction could change from the univariate case. Thus this approach does not ascertain that only important explanatory variables are included in the model.

Determining variables to include in the model is a problem of model choice, and is generally quite a difficult problem to solve in practice due to the
large range of potential models. Here, we tackle this problem via a forward variable selection procedure using the log marginal likelihood (hereafter will be denoted by \( lml \)) to make decisions about whether a variable should be included or excluded from the model. The likelihood is used to measure how the model fits the data \cite{16}. Marginalizing it over the set of parameters produces a marginal likelihood, a well established model selection criterion in Bayesian statistics \cite{22}. By the marginalizing process, it accounts for all the model’s parameters which implies a trivial inbuilt penalty for model complexity. For the numerical stability reason, the marginal likelihood is generally computed in logarithmic scale, resulting an \( lml \). It is difficult (in some cases are impossible) to calculate the marginal likelihood analytically as most of the models contain unknown parameters, and thus an approximation is required. Among many approaches to approximate the marginal likelihood, Integrated nested Laplace approximation (INLA) \cite{32} has become a popular choice for its computationally fast yet still reasonably precise \cite{22}.

In forward variable selection procedure, variables are added to the model one at a time. Starting with a model consisting of an intercept term only, at each step, each variable that is not in the model is tested for inclusion in the model. The variable that results the highest \( lml \) is included in the model, as long as the \( lml \) is higher than that of the current model. The process continues until there is no more increase in the \( lml \). We denote the model consisting these selected variables as the 'preferred' model.

While the forward variable selection procedure should yield a selection of important variables, it ignores model uncertainty. To overcome this problem we consider several potential models, where the prediction is made upon averaging the results from these models. The procedure starts by taking the logistic regression model identified by forward variable selection, then, all models considered in the variable selection procedure are fitted, and predictions are made. A final prediction, called BMA prediction, is then calculated by the weighted average of predictions from all considered models, with the ratio of the model’s \( lml \) to the total \( lml \) of all models as the weights. The logistic regression results in this paper are produced using R-INLA package \cite{32} in R 3.2.5 \cite{30}.

2.3. Model schemes

For each classification method, seven different subsets of predictors (representing seven model schemes) were proposed for the prediction of falls status (faller, non-faller) at both six and twelve months follow-up. In each subset,
several models varied by the selected variables were fit. A collection of these models are contained in the sets of models as listed in (Table 1). The ‘preferred’ model, given by the selected variables producing optimal classification rates, from each set of model is chosen and is used for further analysis.

UPDRS I to UPDRS IV models were fit and compared in order to assess the relative importance of each of the four parts of the UPDRS, and within each part, to identify the relative importance of items. Whereas important items could be identified from these models, this does not necessarily mean that these items play significant role to predict fall/non-fall in the presence of other items from different (probably more important) parts of the UPDRS. Thus UPDRS model consisting combination of all items was also fit.

As to the aggregate measures, similar procedure was done for Subtotal and Composite models, for assessing the relative importance of two different ways of summarizing information from the UPDRS, through subtotal scores and composite measures. Finally, to infer the optimal way of utilizing the UPDRS in predicting falls, a comparison between UPDRS Subtotal and Composite models is conducted.
### Table 1: Subsets of predictors for models fitted at 6-month and 12-month of follow-up.

| Sets of models | Predictor variables          |
|----------------|------------------------------|
| UPDRS I        | UPDRS I items               |
| UPDRS II       | UPDRS II items              |
| UPDRS III      | UPDRS III items             |
| UPDRS IV       | UPDRS IV items              |
| UPDRS          | all UPDRS items             |
| Subtotal       | subtotals 1-4               |
| Composite      | tremor, rigidity, bradykinesia, PIGD |

#### 2.4. Model assessment

All models in each classification method were assessed by their ability to predict new data by the leave-one-out cross validation method. The classification rates were in the form of sensitivity (or true positive rate, TPR), specificity (1-false positive rate, FPR), and accuracy, and calculated as follows:

\[
\text{sensitivity} = \frac{TP}{TP + FN} \tag{1}
\]

\[
\text{specificity} = \frac{TN}{TN + FP} \tag{2}
\]

\[
\text{accuracy} = \frac{TP + TN}{TP + FP + TN + FN} \tag{3}
\]

with

(a) TP, true positives, is the number of patients who actually fell and classified as fallers.

(b) FP, false positives, is the number of patients who actually did not fall and classified as fallers.

(c) TN, true negatives, is the number of patients who actually did not fall and classified as non-fallers.

(d) FN, false negatives, is the number of patients who actually did not fall and classified as non-fallers.
For the logistic regression, once the predicted probabilities are obtained, the classification is based on a chosen threshold. Options for the classification thresholds ranges from 0 to 1. If the predicted probability is greater than the threshold, then it is classified as a faller, and vice versa. The threshold that was actually used was the value jointly optimize the sensitivity and specificity.

In addition, ROC curves and the corresponding area under the ROC curves (AUC) were also presented for models assessment. The graph of ROC reflects the accuracy of the diagnostic test. ROC near the diagonal line means the model is not useful, as the prediction is not different than the random guess. A good model fit will produce ROC close to the upper left corner of the graph, where the TPR (sensitivity) is close to 1 and FPR (1-specificity) is close to 0. Graphs of ROC curves in this paper were produced using ROCR [33] package in R 3.2.5 [30].

3. Results

3.1. Participants description

Table 2 summarizes the subjects classified by fallers (those who experienced at least 1 fall during the follow-up period) and non-fallers (those who did not fall), at 6-month and 12-month of follow-up period.

The majority of the participants were males. Proportion of fallers and non-fallers is around the same for males, while for females, proportion of fallers is around twice of non-fallers. However, this difference is not statistically significant, as implied by the chi-squared independence test. Similarly, there is no significant difference between age, on the average, between the two groups. Although the number is small, the relative proportion of fallers was higher in people who lived alone than in those who lived with family. Yet, the difference is not significant. Overall, there are no significant differences between fallers and non-fallers based on their demographic information.

As for the disease specific measurements, UPDRS sub-totals and total scores (except for Subtotal 1 at 6 month and Subtotal 1 and Subtotal 2 at 12 month of follow up) discriminate the two groups. Bradykinesia shows consistency of discriminating fall/non-fall groups at the two follow-up times, while rigidity shows the opposite. Prospective fallers had been diagnosed with the disease for a slightly longer time and had more falls prior to the participation in the study, compared to the non-fallers.
Table 2: Descriptive statistics for study cohort classified by fallers and non-fallers at 6-month and 12-month of follow-ups. Each categorical variable is summarized by the frequency (%). Numerical variables are summarized by mean (standard deviation). p-value is the statistical significance for Mann-Whitney-Wilcoxon test (for quantitative variables) and chi-square test (for categorical variable).

|                        | Fallers | Non-Fallers | p-value | Fallers | Non-Fallers | p-value |
|------------------------|---------|-------------|---------|---------|-------------|---------|
| **Demographic**        |         |             |         |         |             |         |
| Gender                 |         |             |         |         |             |         |
| Male                   | 16 (42%)| 22 (58%)    | 0.21    | 19 (50%)| 19 (50%)    | 0.45    |
| Female                 | 9 (69%) | 4 (31%)     |         | 9 (69%) | 4 (31%)     |         |
| Age (year)             | 65.9 (8.2) | 67.2 (7.8) | 0.90    | 67.1 (7.9) | 65.8 (6.9) | 0.70    |
| **Living arrangement**|         |             |         |         |             |         |
| Alone                  | 3 (60%) | 2 (40%)     | 0.90    | 3 (60%) | 2 (40%)     | 0.98    |
| With family            | 22 (48%)| 24 (52%)    |         | 25 (54%)| 21 (46%)    |         |
| **Disease specific**   |         |             |         |         |             |         |
| UPDRS scores           |         |             |         |         |             |         |
| Subtotal 1             | 2.9 (2.5) | 1.9 (1.7)  | 0.20    | 2.9 (2.5) | 1.8 (1.7)  | 0.15    |
| Subtotal 2             | 14.4 (6.1)| 9.3 (3.9)  | 0.03    | 12.8 (5.9)| 9.6 (3.9)  | 0.13    |
| Subtotal 3             | 21.8 (9.4)| 15.1 (8.9) | 0.01    | 20.5 (10.2)| 16.2 (9.1) | 0.09    |
| Subtotal 4             | 2.7 (2.3)| 1.3 (1.8)  | 0.08    | 2.5 (2.5) | 1.2 (1.5)  | 0.05    |
| Total                  | 41.5 (13.7)| 27.2 (12.3)| 0.00    | 38.7 (14.8)| 28.7 (12.1)| 0.03    |
| Tremor                 | 3.0 (4.2)| 3.1 (2.2)  | 0.20    | 2.6 (3.9) | 3.3 (2.2)  | 0.06    |
| Rigidity               | 3.52 (3.1)| 3.1 (3.3)  | 0.40    | 3.6 (2.9) | 3.3 (3.5)  | 0.47    |
| PIGD                   | 4.9 (2.9)| 3.1 (2.5)  | 0.05    | 4.3 (2.9) | 3.5 (2.5)  | 0.37    |
| Bradykinesia           | 7.7 (4.1)| 4 (3.3)    | 0.01    | 7.2 (4.3) | 4.2 (3.5)  | 0.01    |
| Duration               | 6.7 (4.8)| 4.6 (2.7)  | 0.40    | 6.7 (4.2) | 4.4 (2.4)  | 0.08    |
| Previous falls         | 1.7 (1.6)| 0.5 (1.1)  | 0.01    | 1.4 (1.5) | 0.5 (1.3)  | 0.02    |
3.2. *Univariate logistic regression models.*

Further exploratory on individual items of the UPDRS were conducted through fitting univariate logistic regression model. Each item is used as an explanatory for fall/non-fall prediction. The results were summarized in the form of an odds ratio (OR, with 95% confidence intervals) of falling (at 6-month) given the item measurements (Figure 2). Several items produced non-unity ORs: swallowing, dressing, falling (unrelated to freezing) and freezing for UPDRS II, and hand pronate/supinate and leg agility for UPDRS III. This indicates the usefulness of these items in explaining falls.

Similar exploratory were also conducted for the aggregate measures, and is shown in Figure 3. Subtotal 2 and bradykinesia produced ORs greater than 1, indicating higher risk of falls for patients with higher score of these aggregate measures. Graphs for univariate tests at 12-month are given in Appendix A.

Overall, the plots suggest the usefulness of UPDRS individual items, as well as the composite measures, for falls prediction. However, this result should be used just as such motivate a modelling approach. The condition of people with PD is affected by interdependent factors, some were measured by the UPDRS, and thus the associations between items and falls occurrences might change when other measures are taken into account. Thus, a multivariate model is preferred rather than univariate analysis.
Figure 2: Odds ratio (with 95% CI) of falls classification at 6-month of follow-up, using univariate logistic regression with individual items of the UPDRS as the predictor.
3.3. Relative importance of UPDRS Parts I - IV

For the interpretation, the results of logistic regression from a forward variable selection procedure will be referred to as LOGIT, and the results from model averaging procedure will be referred to as BMA. Table 3 presents classification rates (accuracy, sensitivity, and specificity) and AUC for UPDRS I - IV models. ROC curves are depicted in Figure 4. In general, all methods agree that items of UPDRS II and UPDRS III can classify participants into fall/non-fall groups better than UPDRS I or UPDRS IV, as these two models produce high accuracy, sensitivity and specificity, at 6- and 12-month follow-up times. This is also confirmed by the high values of AUC for UPDRS II and III models. Between UPDRS I and IV, items of the latter part are more informative than UPDRS I items, as implied by higher classification rates for UPDRS IV model than that for UPDRS I model.

More insight into UPDRS II and UPDRS III models at 6-month of follow-up, the accuracy is not significantly different for these two models. All the methods produce almost the same classification rates, with the range between 71% to 75%. DT and RF yield the same sensitivity at 75%, while LOGIT and BMA produce higher sensitivity for UPDRS III model, at 82% and 78%.
Table 3: Classification rates (accuracy, sensitivity, specificity) and AUC of models with individual items of UPDRS Part I-IV as the explanatory variables. Highest values among the four models within each method are in bold. LOGIT is the logistic regression with forward variable selection, and BMA is the logistic regression with the Bayesian model averaging.

(a) At 6-month of follow-up

| Model  | Accuracy | Sensitivity | Specificity | AUC    |
|--------|----------|-------------|-------------|--------|
|        | DT       | RF          | LOGIT       | BMA    |
|        | DT       | RF          | LOGIT       | BMA    |
| UPDRS I| 0.52     | 0.52        | 0.52        | 0.52   |
| UPDRS II| 0.71    | 0.70        | 0.75        | 0.71   |
| UPDRS III| 0.71    | 0.71        | 0.73        | 0.73   |
| UPDRS IV| 0.65    | 0.65        | 0.65        | 0.65   |

(b) At 12-month of follow-up

| Model  | Accuracy | Sensitivity | Specificity | AUC    |
|--------|----------|-------------|-------------|--------|
|        | DT       | RF          | LOGIT       | BMA    |
|        | DT       | RF          | LOGIT       | BMA    |
| UPDRS I| 0.50     | 0.57        | 0.50        | 0.50   |
| UPDRS II| 0.73    | 0.73        | 0.73        | 0.73   |
| UPDRS III| 0.71    | 0.79        | 0.79        | 0.71   |
| UPDRS IV| 0.63    | 0.63        | 0.65        | 0.65   |

respectively. The opposite is observed for the specificity, with the higher values for UPDRS II model, at 73% to 79% as produced by RF, LOGIT, and BMA. Similar with accuracy, the difference in AUC is negligible (only 2%) -as also implied by overlap ROC curves in Figures 4(b) and 4(c)- thus confirms the comparable results of UPDRS II and UPDRS III models.

Similar trend is also obtained for the 12-month follow-up. The accuracy of UPDRS II and UPDRS III models are between 73% to 79%, with the highest is 79% for UPDRS III model using RF. All methods consistently produce higher sensitivity for UPDRS III model than that for UPDRS II model, with the highest at 83% produced by DT and RF. While, the specificity for UPDRS II is consistently higher (for the 4 implemented methods) than that of UPDRS III, with the highest specificity at 81% produced by BMA. The difference in AUC for UPDRS II and UPDRS III models is only 2% at 6-month period, and more varied (from 2 to 7%) at 12-month period. However, AUC does not clearly differentiate the two models. Despite higher AUC for UPDRS II model produced by 3 methods other than RF, the highest AUC at 0.84 is obtained from UPDRS III model using RF.

Comparing the two follow-up times, the differences in accuracy at 6-month and 12-month are not significant for all models. There are variations in sensitivity, specificity, and AUC values produced by the same method for the same model. For example, sensitivity for UPDRS III model is higher at
(a) UPDRS I model.  
(b) UPDRS II model.  
(c) UPDRS III model.  
(d) UPDRS IV model.  

Figure 4: ROC curves for falls classification at 6 month of follow-up using items of each part of the UPDRS. Classification methods employed are: Decision Tree (solid), Random Forest (dashed), logistic regression with stepwise (dotted), and logistic regression with BMA (dashed dotted).

12-month than that at 6-month based on DT and RF, and the opposite is for UPDRS II model. However, LOGIT and BMA produced lower sensitivity at 12-month than that at 6-month for these models. Yet, the relatively small differences can be neglected. Similar variations also obtained for specificity
and AUC. Thus, it can be inferred that 6-month and 12-month time periods produce comparably similar results with regards to falls prediction based on items of UPDRS parts I - IV.

Comparing the methods, all produced relatively the same results, with slight variations for UPDRS II and UPDRS III models. The accuracy is slightly higher -on average- for LOGIT and BMA in UPDRS II and UPDRS III models at 6-month and UPDRS II model at 12-month period. On the other hand, DT and RF produce a slightly higher sensitivity than LOGIT and BMA, as can be seen in UPDRS II model at 6-month and UPDRS II and UPDRS III models at 12-month. The specificity from DT and RF are less than that from LOGIT and BMA only -on average- for UPDRS II model at 12-month period. However, all the differences from these comparison are relatively small, less than 5% on average. Thus, it can be inferred that among the 4 methods implemented for this study, none is significantly produce different results.

Overall, individual items of parts II and III of the UPDRS are useful explanatory for falls classification. It cannot be decided clearly which of these two parts is more accurate to predict fall/non-fall, as the differences in accuracy of UPDRS II model is not distinct from that of UPDRS III model. However, further checking indicates that UPDRS III items tend to be more sensitive than UPDRS II items in identifying fallers for their higher sensitivity. While if the focus is in identifying non-faller, then UPDRS II model is preferred as its specificity is higher than that of UPDRS III model.

3.4. Predictive comparison between the individual items and summary quantities of UPDRS

The classification rates for UPDRS items, Subtotal, and Composite models are presented in Table 4 and the corresponding ROC curves are depicted in Figure 5.

In general, UPDRS items are more informative than the aggregate measures as the accuracy and sensitivity of UPDRS model are higher than that of Subtotal and Composite models. An exception is the accuracy produced by RF, which yield higher values for Subtotal model, 84% at 6-month and 83% at 12-month periods. The sensitivity of UPDRS items model range from 80% to 85%, almost double the sensitivity of Subtotal model for LOGIT and BMA at 6-month period. Although lower than that of Subtotal model, the specificity of UPDRS items model are still reasonably high, at 74% to 82% (except for DT is 61% at 6-month period). The relatively higher AUC values
Table 4: Classification rates (accuracy, sensitivity, specificity) and AUC of models with individual items and aggregate measures of UPDRS as the explanatory variables. Highest values among the four models within each method are in bold. LOGIT is the logistic regression with forward variable selection, and BMA is the logistic regression with the Bayesian model averaging.

(a) At 6-month of follow-up

| Model     | Accuracy DT | DT  | Accuracy RF | RF  | Accuracy LOGIT | LOGIT | Accuracy BMA | BMA |
|-----------|-------------|-----|-------------|-----|----------------|-------|--------------|-----|
| UPDRS     | 0.72        | 0.78| 0.80        | 0.78| 0.81           | 0.85  | 0.85         | 0.80| 0.61          | 0.74  | 0.77         | 0.82  | 0.77         | 0.83  | 0.85         | 0.77  |
| Subtotal  | 0.67        | 0.84| 0.65        | 0.61| 0.73           | 0.82  | 0.40         | 0.49| 0.61          | 0.85  | 0.85         | 0.72  | 0.72         | 0.77  | 0.63         | 0.65  |
| Composite | 0.69        | 0.80| 0.68        | 0.68| 0.63           | 0.78  | 0.78         | 0.78| **0.76**      | 0.82  | 0.58         | 0.58  | 0.73         | **0.87**| 0.70         | 0.70  |

(b) At 12-month of follow-up

| Model     | Accuracy DT | DT  | Accuracy RF | RF  | Accuracy LOGIT | LOGIT | Accuracy BMA | BMA |
|-----------|-------------|-----|-------------|-----|----------------|-------|--------------|-----|
| UPDRS     | 0.74        | 0.79| 0.76        | 0.79| 0.83           | 0.83  | 0.80         | 0.84| 0.71          | 0.76  | 0.74         | 0.78  | 0.77         | 0.84  | 0.79         | 0.81  |
| Subtotal  | 0.60        | 0.83| 0.49        | 0.49| 0.61           | 0.82  | 0.51         | 0.51| 0.60          | 0.85  | 0.48         | 0.48  | 0.64         | 0.82  | 0.55         | 0.55  |
| Composite | 0.68        | 0.77| 0.63        | 0.63| **0.85**       | 0.85  | 0.69         | 0.69| 0.48          | 0.66  | 0.55         | 0.55  | 0.67         | 0.83  | 0.63         | 0.63  |

-as confirmed by ROC curves in Figure 5(a) is more to the upper left corner than ROC curves in Figure 5(b),(c)- also support that UPDRS items outperform UPDRS subtotals and composite measures to predict fall/non-fall.

Further comparison of the aggregate measures, at 6-month period, the difference in accuracy between subtotal scores and composite measures is relatively small. However, based on LOGIT and BMA, the sensitivity of Subtotal model is far less than that of Composite model and the opposite is for the specificity. While of based on DT and RF, the sensitivity is higher for Subtotal model than for Composite model. Furthermore, the difference between sensitivity and specificity is not as large as that based on LOGIT and BMA. The AUC also suggests similar results for Subtotal and Composite models.

While at 12-month period, despite the relatively low values, the accuracy of Composite model is substantially higher by 14% than Subtotal model (based on LOGIT and BMA). Its sensitivity, specificity, and AUC are also higher than the Subtotal model. This implies that based on LOGIT and BMA, the composite measures are relatively more informative than the subtotals in predicting fall/non-fall. However, composite model has higher difference between sensitivity and specificity than the Subtotal model, and the sensitivity is higher than the specificity. This implies that making a decision as to whether a patient will fall or not is easier using the composite measures than the subtotals. On the other hand, DT and RF yield higher
classification rates -on average- for these two models.

In summary, UPDRS items are shown to outperform the aggregate measures in predicting fall/non-fall. As for the aggregate measures, the difference in classification rate between Subtotal model and Composite model is relatively small. Composite measures are more sensitive in identifying fallers than identifying non-fallers, compared to the subtotals. As for the methods, the tree-based methods (DT and RF) provide higher classification rates than the regression-based methods (LOGIT and BMA) when aggregate measures are used instead of the UPDRS items.

3.5. Important risk factors related to falls

Important explanatory variables extracted from the models for DT, RF, and logistic regression with forward variable selection (LOGIT) are listed in Table 5 for the two follow-up times. Variables in BMA for the selected models are listed in Appendix C. In general, the four methods selected similar set of variables as shown by several common important variables in each model.

At 6-month of follow-up, by modelling each of the four parts of the UPDRS separately several items were consistently selected: thought disorder in UPDRS I, dressing and falling in UPDRS II, hand pronate/supinate in UPDRS III, and symptomatic orthostasis and sleep disturbance in UPDRS IV. In addition for LOGIT, freezing was also a significant item in UPDRS Part II, and leg agility in UPDRS Part III. Similar set of items were also
selected at the 12-month period, indicating the consistency of the models and methods performance in the two follow-up times.

When all items were combined together, as in all UPDRS items model, dressing and hand pronate/supinate were always selected by all classification methods, in both time periods. In addition, speech, sleep disturbance, and symptomatic orthostasis were also regarded important in regression.

Figure 6: Falls prediction using Decision Tree with items of the UPDRS as the predictor variables. Items scoring are based on UPDRS questionnaire scoring. Values in the final nodes are the purity (homogeneity) of the nodes (proportion of correctly classified patients relative to all patients in that node).

As for the aggregate measures, Subtotal 2 and bradykinesia are selected by all the methods at 6-month and 12-month periods. In addition, subtotal 3 was also considered important at 6-month but not at 12-month if the models were based on DT and LOGIT. It it worth to note that at 6-month period,
PIGD is selected in the model but is replaced by tremor at 12-month period. Among the 4 methods used, DT is appealing for its ease of visualization and interpretation. Figure 6 and Figure 7 exemplify falls prediction at 6-month period using for UPDRS items and aggregate measures as the predictor variables. Prediction using UPDRS I model is not displayed, as it has only 1 predictor variable, thought disorder, and the classification rates are the lowest amongst all models. Through Figure 6 we can infer that examination on several selected (targeted) items could provide information on deciding whether a patient will likely to fall or not-fall based on the cut-offs given from the model without the need to calculate some derived scores such as odds ratios as in logistic regression.
Table 5: Selected variables for falls prediction using the classification methods: Decision Tree (DT), Random Forest (RF), and logistic regression with forward variable selection (LOGIT).

(a) At 6-month period

| Method | UPDRS I       | UPDRS II     | UPDRS III                  | UPDRS IV                  | All UPDRS items | UPDRS subtotals | Composite measures |
|--------|---------------|--------------|----------------------------|---------------------------|-----------------|-----------------|-------------------|
| DT     | Thought disorder | Falling   | Hand pronate/supinate      | Sleep disturbance         | Dressing        | Subtotal 2      | Bradykinesia      |
|        |                |             |                            | Symptomatic orthostasis  | Hand pronate/supinate | Subtotal 3   |                   |
| RF     | Thought disorder | Falling   | Hand pronate/supinate      | Sleep disturbance         | Dressing        | Subtotal 2      | Bradykinesia      |
|        |                |             |                            | Symptomatic orthostasis  | Hand pronate/supinate | Subtotal 3   |                   |
| LOGIT  | Thought disorder | Falling   | Hand pronate/supinate      | Sleep disturbance         | Dressing        | Subtotal 2      | Bradykinesia      |
|        | Freezing       |             | Leg agility                | Symptomatic orthostasis  | Speech          | Subtotal 3      |                   |
|        | Dressing       |             |                            | Early morning dystonia   | Hand pronate/supinate |             |                   |
|        | Handwriting    |             |                            |                            | Sleep disturbance|             |                   |
|        |                |             |                            |                            | Symptomatic orthostasis |             |                   |

(b) At 12-month period

| Method | UPDRS I       | UPDRS II     | UPDRS III                  | UPDRS IV                  | All UPDRS items | UPDRS subtotals | Composite measures |
|--------|---------------|--------------|----------------------------|---------------------------|-----------------|-----------------|-------------------|
| DT     | Thought disorder | Falling   | Action tremor              | Sleep disturbance         | Dressing        | Subtotal 1      | Tremor            |
|        |                |             |                            | Symptomatic orthostasis  | Hand pronate/supinate | Subtotal 2   |                   |
| RF     | Thought disorder | Falling   | Tremor at rest             | Sleep disturbance         | Dressing        | Subtotal 2      | Tremor            |
|        | Freezing       |             |落入字| Symptomatic orthostasis  | Falling              | Subtotal 3      | Bradykinesia      |
|        | Dressing       |             |落入字|                            | Hand pronate/supinate |             |                   |
| LOGIT  | Thought disorder | Falling   | Hand pronate/supinate      | Dyskinesia (disability)  | Dressing        | Subtotal 2      | Bradykinesia      |
|        | Freezing       |             |落入字| Anorexia, nausea, vomiting| Speech          | Subtotal 3      |                   |
|        | Dressing       |             |落入字| Sleep disturbance          | Hand pronate/supinate |             |                   |
|        | Walking        |             |落入字| Symptomatic orthostasis  | Sleep disturbance  |             |                   |
|        |                |             |落入字|                            | Symptomatic orthostasis |             |                   |
4. Discussion

This study demonstrated different ways of utilizing the UPDRS measurements for classifying people at the early stages of PD into fallers/non-fallers: using the individual items of the UPDRS, and aggregate measures derived from them. It is shown that the selected individual items of the UPDRS were more informative than the aggregate measures.

It is worth noting that although UPDRS IV was often overlooked in similar studies for falls prediction, the odds ratio for falls prediction using symptomatic orthostasis and sleep disturbance were greater than 1 in the univariate model. This is also confirmed by a comparably high purity of the classification results using UPDRS IV model as shown in Figure 6(c). Yet, it has a relatively higher reduction in the classification rates compared to UPDRS II and UPDRS III when applied to cross-validation data.

There is no clear difference between the performance of the methods used in this paper, with regards to the classification rates. Logistic regression provided higher accuracy in some models, yet there were cases where tree-based methods provide higher sensitivity, and vice versa. Logistic regression is attractive for its odds ratio interpretation for the effect, or contribution, of predictor variables in prediction. It is also less prone to over-fitting given the variables in the model are appropriately selected. Nevertheless, decision trees are more appealing since their visualization is easily interpretable, without the need for further calculation. The non-parametric approach of decision trees also offer the flexibility to handle a large number of variables, as demonstrated in this paper for models with UDPSR items as the predictor variables. While DT is often regarded to be more liable to over-fit, comparing the classification rates and selected variables with RF, the more robust method, the results are not greatly different for our data. Thus, we prefer and presented the fall/non-fall prediction rules based on DT, as in Section 3.5.

Models were fit at two follow-up times to assess the effect of time to fall/non-fall prediction. The classification rates were varied at the two times, but the differences were not significant. There were also variations in variables being selected for the two time periods, but the differences were minor. So, overall it seemed that there was little change over time. This may be due to the relatively short differences in follow up times. However, on another perspective, this might imply that a shorter study time (6 months) could provide similar information than a longer study time (12 months).
Regarding the measurements, the Movement Disorder Society (MDS) has released the improved version of the UPDRS called the MDS sponsored UPDRS (MDS-UPDRS). Yet, the data used in this study were based on the UPDRS measurements. However, a reasonably high classification rates obtained showed that the UPDRS provided a useful information for fall/non-fall prediction. Moreover, results from models using only the UPDRS were comparable to those using additional information from other instruments (results not shown); suggesting that numerous measurements from many different instruments are not needed when information from a few particular instruments is used in an optimal way.

5. Summary

Through this study, we have provided empirical evidence that in the early stages of PD, fall/non-fall occurrences were better explained using items of the UPDRS than using the composite measures. The highest classification rates for this model are: 80% accuracy, 85% sensitivity, and 77% specificity, higher than previous studies.

Among the four parts of the UPDRS, selected items from UPDRS Parts II and III produce a reasonably high classification rates compared to the other parts. The classification rates from all 4 methods at 2 time periods for UPDRS II items varied within these range: $70 - 75\%$ accuracy, $70 - 75\%$ sensitivity, and $68 - 71\%$ specificity. While for UPDRS III items, the range for the classification rates are $71 - 79\%$ accuracy, $73 - 83\%$ sensitivity, and $58 - 74\%$ specificity.

We also identified variables that best predict fall/non-fall. It was also inferred that results from a 6-month follow-up time were not greatly different to that from a 12-month follow-up time, suggesting a shorter study time (6 months) could replace the longer study time (12 months).

Identification of the UPDRS items that are highly associated with falls offers several advantages. From a practical point of view, adjustments to treatment might be developed for PD patients to prevent falls. Focusing assessment based on the identified risk factors may provide more reliable responses, which will be advantageous for building a more informative model.

References

[1] Alan Agresti. An introduction to categorical data analysis, 2nd ed. Hoboken. NJ: John Wiley & Sons, Inc, 2007.
[2] LM Allcock, EN Rowan, IN Steen, K Wesnes, RA Kenny, and DJ Burn. Impaired attention predicts falling in Parkinson’s disease. *Parkinsonism & Related Disorders*, 15(2):110–115, 2009.

[3] Yali Amit and Donald Geman. Shape quantization and recognition with randomized trees. *Neural Computation*, 9(7):1545–1588, 1997.

[4] Yacov Balash, Ch Peretz, G Leibovich, T Herman, JM Hausdorff, and Nir Giladi. Falls in outpatients with Parkinson’s disease. *Journal of Neurology*, 252(11):1310–1315, 2005.

[5] Bastiaan R Bloem, Yvette AM Grimbergen, Monique Cramer, Mirjam Willemsen, and Aeilko H Zwinderman. Prospective assessment of falls in Parkinson’s disease. *Journal of Neurology*, 248(11):950–958, 2001.

[6] H. Bostrom. Estimating class probabilities in random forests. In *Sixth International Conference on Machine Learning and Applications (ICMLA 2007)*, pages 211–216, Dec 2007.

[7] Leo Breiman. Some infinity theory for predictor ensembles. Technical report, Technical Report 579, Statistics Dept. UCB, 2000.

[8] Leo Breiman. Random forests. *Machine Learning*, 45(1):5–32, 2001.

[9] Leo Breiman. randomforest: Breiman and Cutler’s random forests for classification and regression, 2006.

[10] Sophie Drapier, Sylvie Raoul, Dominique Drapier, Emmanuelle Leray, Francois Lallement, Isabelle Rivier, Paul Sauleau, Youen Lajat, Gilles Edan, and Marc Vérin. Only physical aspects of quality of life are significantly improved by bilateral subthalamic stimulation in Parkinson’s disease. *Journal of Neurology*, 252(5):583–588, 2005.

[11] S Fahn, D Oakes, Ira Shoulson, K Kieburtz, A Rudolph, A Lang, CW Olanow, C Tanner, and K Marek. Levodopa and the progression of Parkinson’s disease. *The New England Journal of Medicine*, 351(24):2498–2508, 2004.

[12] Xin Fang, Runkui Li, Haidong Kan, Matteo Bottai, Fang Fang, and Yang Cao. Bayesian model averaging method for evaluating associations between air pollution and respiratory mortality: a time-series study. *BMJ open*, 6(8):e011487, 2016.
[13] Clement François, Italo Biaggioni, Cyndya Shibao, Augustina Ogbonnaya, Huai-Che Shih, Eileen Farrelly, Adam Ziemann, and Amy Duhig. Fall-related healthcare use and costs in neurogenic orthostatic hypotension with Parkinson’s disease. *Journal of Medical Economics*, 20(5):525–532, 2017. PMID: 28125950.

[14] Jerome H Friedman. On bias, variance, 0/1-loss, and the curse-of-dimensionality. *Data Mining and Knowledge Discovery*, 1(1):55–77, 1997.

[15] Tatjana Gazibara, Tatjana Pekmezovic, Darija Kisić-Tepavcevic, Marina Svetel, Aleksandra Tomic, Iva Stankovic, and Vladimir S Kostic. Incidence and prediction of falls in Parkinson’s disease: a prospective cohort study. *European Journal of Epidemiology*, 30(4):349–352, 2015.

[16] Andrew Gelman, John B Carlin, Hal S Stern, and Donald B Rubin. *Bayesian data analysis*, volume 2. Chapman & Hall/CRC Boca Raton, FL, USA, 2014.

[17] CW Gini. Variability and mutability, contribution to the study of statistical distribution and relations. *Studi Economico-Giuricici della R*, 1912.

[18] Christopher G. Goets, Poewe W., B. Dubois, A Schrag, M. B. Stern, A Lang, P. A. LeWitt, S Fahn, J. Jankovic, C. Olanow, P. Martinez-Martín, G. T. Stebbins, R. Holloway, D. Nyenhuis, C. Sampaio, R. Dodel, J. Kulisevsky, B. Tilley, S. Leurgans, J. Teresi, S. R. Shaftman, and N. LaPelle. MDS-UPDRS, July 2008.

[19] Peggy Gray and Kathleen Hildebrand. Fall risk factors in Parkinson’s disease. *Journal of Neuroscience Nursing*, 32(4):222–228, 2000.

[20] Jennifer A Hoeting, David Madigan, Adrian E Raftery, and Chris T Volinsky. Bayesian model averaging: A tutorial. *Statistical Science*, 15:382–401, 1999.

[21] Martina Hoskovcová, Petr Dušek, Tomáš Sieger, Hana Brožová, Kateřina Zárubová, Ondřej Bezdíček, Otakar Šprdlík, Robert Jech, Jan Štochl, Jan Roth, et al. Predicting falls in Parkinson disease: what is the value of instrumented testing in off medication state? *PloS one*, 10(10):e0139849, 2015.
[22] Aliaksandr Hubin and Geir Storvik. Estimating the marginal likelihood with Integrated nested Laplace approximation (INLA). *arXiv preprint arXiv:1611.01450*, 2016.

[23] Karen H Karlsen, Elise Tandberg, Dag Årsland, and Jan P Larsen. Health related quality of life in Parkinson’s disease: a prospective longitudinal study. *Journal of Neurology, Neurosurgery & Psychiatry*, 69(5):584–589, 2000.

[24] GK Kerr, Charles J Worringham, Michael H Cole, Philippe F Lacherez, Joanne M Wood, and PA Silburn. Predictors of future falls in Parkinson’s disease. *Neurology*, 75(2):116–124, 2010.

[25] P1 Martínez-Martín, A Gil-Nagel, L Morlán Gracia, J Balseiro Gómez, J Martínez-Sarriés, and F Bermejo. Unified Parkinson’s disease rating scale characteristics and structure. *Movement Disorders*, 9(1):76–83, 1994.

[26] Elena Moro, Clement Hamani, Yu-Yan Poon, Thamar Al-Khairallah, Jonathan O Dostrovsky, William D Hutchison, and Andres M Lozano. Unilateral pedunculopontine stimulation improves falls in Parkinson’s disease. *Brain*, 133(1):215–224, 2010.

[27] Movement Disorder Society Task Force on Rating Scales for Parkinson’s Disease et al. The unified Parkinson’s disease rating scale (UPDRS): status and recommendations. *Movement Disorders: Official Journal of the Movement Disorder Society*, 18(7):738, 2003.

[28] Ruth M Pickering, Yvette AM Grimbergen, Una Rigney, Ann Ashburn, Gordon Mazibrada, Brian Wood, Peggy Gray, Graham Kerr, and Bastiaan R Bloem. A meta-analysis of six prospective studies of falling in Parkinson’s disease. *Movement Disorders*, 22(13):1892–1900, 2007.

[29] J. Ross Quinlan. Induction of decision trees. *Machine Learning*, 1(1):81–106, 1986.

[30] R Core Team. *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria, 2016.
[31] William C Roller, Sander Glatt, Bridget Vetere-Overfield, and Ruth Hassanein. Falls and Parkinson’s disease. *Clinical Neuropharmacology*, 12(2):98–105, 1989.

[32] Håvard Rue, Sara Martino, and Nicolas Chopin. Approximate Bayesian inference for latent Gaussian models by using integrated nested Laplace approximations. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 71(2):319–392, 2009.

[33] Tobias Sing, Oliver Sander, Niko Beerenwinkel, and Thomas Lengauer. ROCR: Visualizing classifier performance in R. *Bioinformatics*, 21(20):3940–3941, 2005.

[34] Jooeun Song, Beth E Fisher, Giselle Petzinger, Allan Wu, James Gordon, and George J Salem. The relationships between the unified Parkinson’s disease rating scale and lower extremity functional performance in persons with early-stage Parkinson’s disease. *Neurorehabilitation and Neural Repair*, 23(7):657–661, 2009.

[35] Yanmin Sun, Andrew KC Wong, and Mohamed S Kamel. Classification of imbalanced data: A review. *International Journal of Pattern Recognition and Artificial Intelligence*, 23(04):687–719, 2009.

[36] Terry M Therneau, Beth Atkinson, and Brian Ripley. rpart: Recursive partitioning. *R package version*, 3(3.8), 2010.

[37] Karen E Thomas, Judy A Stevens, K Sarmiento, and Marlena M Wald. Fall-related traumatic brain injury deaths and hospitalizations among older adults in United States, 2005. *Journal of Safety Research*, 39(3):269–272, 2008.

[38] Valerie Viallefont, Adrian E Raftery, and Sylvia Richardson. Variable selection and Bayesian model averaging in case-control studies. *Statistics in Medicine*, 20(21):3215–3230, 2001.

[39] Duolao Wang, Wenyang Zhang, and Ameet Bakhai. Comparison of Bayesian model averaging and stepwise methods for model selection in logistic regression. *Statistics in Medicine*, 23(22):3451–3467, 2004.

[40] BH Wood, JA Bilclough, A Bowron, and RW Walker. Incidence and prediction of falls in Parkinson’s disease: A prospective multidisciplinary
study. *Journal of Neurology, Neurosurgery & Psychiatry, 72*(6):721–725, 2002.
Appendix A. Falls odds ratio for the univariate logistic regression model at the 12-month of follow up.

Figure A.8: Odds ratio (with 95% CI) for the univariate logistic regression model using the UPDRS individual items as the explanatory variable.
Figure A.9: Odds ratio (with 95% CI) of falls classification at 12-month of follow-up, using univariate logistic regression model with aggregate measures of the UPDRS as the explanatory variable. Subtotals 1-4 are the sums of item scores in UPDRS Parts I - IV, respectively.

Appendix B. ROC at 12 month of follow-up
Figure B.10: ROC curves for falls classification at 12 month of follow-up using items of each part of the UPDRS. Classification methods employed are: Decision Tree (solid), Random Forest (dashed), logistic regression with stepwise (dotted), and logistic regression with BMA (dashed dotted).
Figure B.11: ROC curves for falls classification at 12 month of follow-up using individual items (a), composite measures (b) and individual items and composite measures of the UPDRS. Classification methods employed are: Decision Tree (solid), Random Forest (dashed), logistic regression with stepwise (dotted), and logistic regression with BMA (dashed dotted).

Appendix C. Results of logistic regression with BMA

No table is produced for UPDRS I model, as only item number 2, Thought disorder, was selected in the model for both 6-month and 12-month period. Also for Composite model, no table is produced as the only variable chosen for predictor is Bradykinesia, both at 6-month and 12-month periods.

Table C.6: BMA results for logistic regression using UPDRS Part II items as predictor variables (UPDRS II models).

| (a) At 6-month period                | (b) At 12-month period            |
|--------------------------------------|----------------------------------|
| **Model**                            | **Model**                        |
| 1                                    | 1                                |
| 2                                    | 2                                |
| 3                                    | 3                                |
| 4                                    | 4                                |
| 5                                    | 5                                |
| Falling                              | Falling                          |
| Freezing                             | Freezing                        |
| Dressing                             | Dressing                        |
| Handwriting                          | Handwriting                     |
| Weight                               | Weight                          |
| 0.27                                 | 0.19                            |
| 0.23                                 | 0.16                            |
| 0.20                                 | 0.13                            |
| 0.14                                 | 0.13                            |
| 0.13                                 | 0.11                            |
Table C.7: BMA results for logistic regression using UPDRS Part III items as predictor variables (UPDRS III models).

|          | (a) At 6-month period | (b) At 12-month period |
|----------|-----------------------|------------------------|
|          | Model 1 | Model 2 | Model 3 | Model 1 | Model 2 |
| Hand pronate/supinate | | | | | |
| Leg agility | | | | | |
| Weight | 0.50 | 0.47 | 0.04 | 0.81 | 0.10 |

Table C.8: BMA results for logistic regression using UPDRS Part IV items as predictor variables (UPDRS IV models).

|          | (a) At 6-month period | (b) At 12-month period |
|----------|-----------------------|------------------------|
|          | Model 1 | Model 2 | Model 3 | Model 1 | Model 2 | Model 3 | Model 4 | Model 5 |
| Sleep disturbance | | | | | | | | |
| Symptomatic orthostasis | | | | | | | | |
| Early morning dystonia | | | | | | | | |
| Weight | 0.76 | 0.17 | 0.08 | 0.35 | 0.18 | 0.17 | 0.10 | 0.06 |

Table C.9: BMA results for logistic regression using all items of the UPDRS as predictor variables (UPDRS models).

|          | (a) At 6-month period | (b) At 12-month period |
|----------|-----------------------|------------------------|
|          | Model 1 | Model 2 | Model 3 | Model 4 | Model 5 | Model 1 | Model 2 | Model 3 | Model 4 | Model 5 |
| Dressing | | | | | | | | | | |
| Speech | | | | | | | | | | |
| Hand pronate/supinate | | | | | | | | | | |
| Sleep disturbance | | | | | | | | | | |
| Symptomatic orthostasis | | | | | | | | | | |
| Weight | 0.25 | 0.20 | 0.17 | 0.13 | 0.11 | 0.54 | 0.19 | 0.12 | 0.12 | 0.04 |
Table C.10: BMA results for logistic regression using subtotals of the UPDRS as predictor variables (Subtotal models) at 6-month period. At 12-month, only Subtotal 2 was selected to include in the model.

|          | Model |   |   |   |
|----------|-------|---|---|---|
|          | 1     | 2 | 3 |
| Subtotal 2 |       |   |   |
| Subtotal 3 |       |   |   |
| Weight    | 0.89  | 0.06 | 0.05 |