University of Exeter
Department of Computer Science

Machine Learning for Classification and Clustering of Dementia Data

Felicity Louise Guest

Submitted by Felicity Louise Guest to the University of Exeter as a thesis for the degree of Doctor of Philosophy in Computer Science, January 2021.

I certify that all material in this thesis which is not my own work has been identified and that any material that has previously been submitted and approved for the award of a degree by this or any other University has been acknowledged.

Signed:  .....

To my grandad,

who will always be my biggest inspiration.

Abstract

Dementia is a term used to describe heterogeneous diseases that can generally be characterised by a decline in cognitive ability that affects daily living. It has been predicted that the prevalence of dementia will increase significantly over the coming years, thus it is a priority worldwide. This thesis discusses research conducted with two primary aims. They were to investigate the use of machine learning for distinguishing between people with and without dementia, as well as differentiating between key dementia subtypes where appropriate; and to gain an understanding of the inherent structure of dementia data, to ultimately investigate disease signatures.

Data was acquired from the National Alzheimer's Coordinating Center in the United States, and a data set comprising 32,573 observations and 260 features of mixed type was utilised. It included features whose values were constrained by relations with others, as well as two types of missingness which arose when data was unexpectedly not recorded and when the information was irrelevant or unobtainable for a known reason, respectively. Notably, the former genuinely missing values were imputed where possible, whilst the latter conditionally missing values were handled.

An imputation approach was developed, which simultaneously builds a random forest classifier while handling conditionally missing values. It maintained the known relations in the data set, so far as possible. A clustering approach was also developed that ultimately measures the similarity of observations based on the similarity of their paths through the trees of an isolation forest before employing spectral clustering. Crucially, it can naturally draw on variables of mixed type.

A dementia classifier with an area under the receiver operating characteristic curve (AUC) of 0.99 and 10 pairwise dementia subtype classifiers with AUCs ranging from 0.88 to 1.0 (rounded) were produced, suggesting machine learning could be a

useful tool for diagnosing dementia and differentiating between the main subtypes. Key features were identified using these classifiers and were markedly different for the two types of diagnosis. Furthermore, preliminary experiments conducted using the clustering approach suggested that mild cognitive impairment may be a mild form of dementia as opposed to a clinical entity, over which there is much debate; and there could be evidence for the current subtypes. Ultimately, these findings have the potential to transform the way dementia is diagnosed.

Acknowledgements

Differences of habit and language are nothing at all if our aims are identical and our hearts are open.

— J. K. Rowling, *Harry Potter and the Goblet of Fire*

I must begin by thanking my supervisors: Richard Everson, David Llewellyn and Elżbieta Kuźma. This thesis would not have been possible without their expertise, advice and support. I would also like to thank my family who believed in me when I did not; their continued encouragement kept me going and enabled me to complete this work.

I am extremely grateful to the Alzheimer's Society and Garfield Weston Foundation for providing funding for this research, and to the University of Exeter for supporting my studies at both undergraduate and postgraduate level. I am also indebted to the National Alzheimer's Coordinating Center for providing the data that was a fundamental part of this work.[1]

[1] The NACC database is funded by NIA/NIH Grant U01 AG016976. NACC data are contributed by the NIA-funded ADCs: P30 AG019610 (PI Eric Reiman, MD), P30 AG013846 (PI Neil Kowall, MD), P30 AG062428-01 (PI James Leverenz, MD), P50 AG008702 (PI Scott Small, MD), P50 AG025688 (PI Allan Levey, MD, PhD), P50 AG047266 (PI Todd Golde, MD, PhD), P30 AG010133 (PI Andrew Saykin, PsyD), P50 AG005146 (PI Marilyn Albert, PhD), P30 AG062421-01 (PI Bradley Hyman, MD, PhD), P30 AG062422-01 (PI Ronald Petersen, MD, PhD), P50 AG005138 (PI Mary Sano, PhD), P30 AG008051 (PI Thomas Wisniewski, MD), P30 AG013854 (PI Robert Vassar, PhD), P30 AG008017 (PI Jeffrey Kaye, MD), P30 AG010161 (PI David Bennett, MD), P50 AG047366 (PI Victor Henderson, MD, MS), P30 AG010129 (PI Charles DeCarli, MD), P50 AG016573 (PI Frank LaFerla, PhD), P30 AG062429-01 (PI James Brewer, MD, PhD), P50 AG023501 (PI Bruce Miller, MD), P30 AG035982 (PI Russell Swerdlow, MD), P30 AG028383 (PI Linda Van Eldik, PhD), P30 AG053760 (PI Henry Paulson, MD, PhD), P30 AG010124 (PI John Trojanowski, MD, PhD), P50 AG005133 (PI Oscar Lopez, MD), P50 AG005142 (PI Helena Chui, MD), P30 AG012300 (PI Roger Rosenberg, MD), P30 AG049638 (PI Suzanne Craft, PhD), P50 AG005136 (PI Thomas Grabowski, MD), P30 AG062715-01 (PI Sanjay Asthana, MD, FRCP), P50 AG005681 (PI John Morris, MD), P50 AG047270 (PI Stephen Strittmatter, MD, PhD).

Contents

List of Figures

List of Tables

Abbreviations

AD	Alzheimer's disease
ADC	Alzheimer's Disease Center
ANOVA	Analysis of variance
ARI	Adjusted Rand index
AUC	Area under the receiver operating characteristic curve
BMI	Body mass index
CDR	Clinical Dementia Rating
CNS	Central nervous system
DLB	Dementia with Lewy bodies
EEG	Electroencephalography
FAQ	Functional Activities Questionnaire
FCS	Fully conditional specification
FNR	False negative rate
FPR	False positive rate
FTD	Frontotemporal dementia
FTLD	Frontotemporal lobar degeneration
GDS	Geriatric Depression Scale
HIS	Hachinski ischemic score
IADL	Instrumental activity of daily living
IQR	Interquartile range
MAR	Missing at random
MCAR	Missing completely at random
MCI	Mild cognitive impairment
MDS	(Metric) multidimensional scaling

MIA	Missingness incorporated in attributes
MMSE	Mini-Mental State Examination
MNAR	Missing not at random
NACC	National Alzheimer's Coordinating Center
NIH	National Institutes of Health
NMI	Normalised mutual information
NPI-Q	Neuropsychiatric Inventory Questionnaire
NRMSE	Normalised root mean squared error
OOB	Out-of-bag
PD	Parkinson's disease
PFC	Proportion of falsely classified entries
ROC	Receiver operating characteristic
SMC-FCS	Substantive model compatible fully conditional specification
TBI	Traumatic brain injury
TIA	Transient ischemic attack
TNR	True negative rate
TPR	True positive rate
UDS	Uniform Data Set
UPDRS	Unified Parkinson's Disease Rating Scale
VD	Vascular dementia
VIMP	Variable importance
WAIS-R	Wechsler Adult Intelligence Scale (Revised)
WCSS	Within-cluster sum of squares

Notation

X	Data set	6
Y	Classification targets	6
N	Number of subjects/visits/observations	6
F	Number of variables/features	6
X_n	A subject in the data set	6
X^f	A variable in the data set	6
X_n^f	A value for a subject and variable in the data set	6
Y_n	Class of a subject in the data set	6
$Y^{(0)}$	Instances of class 0	6
$Y^{(1)}$	Instances of class 1	6
X_{true}^f	True values for a variable	47
X_{pred}^f	Predicted values for a variable	47
$\tilde{X}$	Imputed data set	62
$\boldsymbol{\phi}(\cdot)$	Missing value predicate	62
$\boldsymbol{\psi}(\cdot)$	Conditionally missing value predicate	64
$\boldsymbol{\gamma}(\cdot)$	Conditionally missing fill value legitimate predicate	70
X_{obs}^f	Observed values for a variable	64
$X_{obs\,\vert\,0}^f$	Observed variable values associated with class 0	62
$X_{obs\,\vert\,1}^f$	Observed variable values associated with class 1	62
X_{c-mis}^f	Conditionally missing values for a variable	71
X^{true}	Values replaced by the 1,000 additional missing values	85

X^{imp}	Values imputed for the 1,000 additional missing values	85
n_{val}	Number of true-imputed value pairs considered for the PFC	85
t	Decision tree	6
η	Internal splitting node	6
ℓ	Terminal/Leaf node	6
t_L	Left subtree	7
t_R	Right subtree	7
T	Ensemble of decision trees (random forest or isolation forest)	8
M	Number of trees in the ensemble	8
n_{min}	Number of observations required for splitting	11
K	Number of variables considered for splitting	8
	* Also, number of clusters (chapter 5)	
κ	Number of splits typically generated for a variable	59
S	A split on a variable	6
f_{cp}	Random cut-point for a continuous variable	9
f	Set of possible values for a categorical variable	9
$f^{(in)}$	Set of values present for a categorical variable	9
f_1	Proper nonempty subset of values present for a categorical variable	9
f_2	Subset of values absent for a categorical variable	9
X_L	Subset of observations sent to the left child node	6
X_R	Subset of observations sent to the right child node	6
Y_L	Subset of classification targets sent to the left child node	6
Y_R	Subset of classification targets sent to the right child node	6
$\mathcal{I}_S(X)$	Information gain resulting from a split	7
N_S	Number of observations split on	66
$S_{MIA_{1-3}}$	Three possible MIA splits	59

X_{mis}	Subset of observations with a (conditionally) missing variable value * X_{c-mis} used where necessary	60
X_{obs}	Subset of observations with an observed variable value	60
Y_{mis}	Subset of targets for the (conditionally) missing value observations	60
Y_{obs}	Subset of targets for the observed value observations	60
P	Proximity matrix for an ensemble of trees	68
ρ_n	Path of an observation through a tree	66
Y_ℓ	Classification targets for a terminal node	66
$Y_\ell^{(c)}$	Instances of a class for a terminal node	66
d_{ij}	Depth of last common node	146
P_t	Proximity matrix for a tree	67
$\bar{P}_t$	Normalised proximity matrix for a tree	67
L_0	Stacking level zero	125
L_1	Stacking level one	128
C_{1-5}	Level zero (subtype) classifiers	125
SC	Level one (stacking) classifier	125
$\mathcal{L}$	(Unnormalised) Laplacian matrix	101
$\mathcal{L}_{rw}$	Normalised Laplacian matrix (random walk)	149
D	Diagonal/Degree matrix	101
λ_n	A generalised eigenvalue	150
u_n	A generalised eigenvector	150
U	Projection of the data (eigenvector matrix)	150
U_n	Projected data point	150
$\mathcal{P}$	Partition (set of clusters)	151
c_k	A cluster	150
μ_k	Centroid of a cluster	150

$\mathcal{P}_U$	Partition of the projected data	150
$\mathcal{P}_X$	Partition of the original data	150
Ω	True classes	158

Chapter 1

Introduction

Dementia is a term used to describe heterogeneous diseases that can generally be characterised by a decline in cognitive ability that affects daily living. It mainly affects older people, although it is not a normal part of ageing (World Health Organisation, 2020). There are thought to be numerous types (or subtypes) of dementia, for which the signs and symptoms vary. The four main subtypes are Alzheimer's disease (AD), vascular dementia (VD), dementia with Lewy bodies (DLB) and frontotemporal dementia (FTD). However, dementia can also manifest differently from person to person. Furthermore, it is possible, if not common, for people to be affected by more than one subtype; the term *mixed dementia* is used to describe this eventuality (National Institute on Aging, 2017).

Prince et al. (2015) reported that 46.8 million people worldwide were living with dementia in 2015 and estimated that this number would increase to 131.5 million by 2050. Prince et al. (2016) and Wu et al. (2017) highlight, however, that predicting the prevalence (proportion of the population affected at a certain time) and incidence (occurrence of new cases during a period of time) of dementia is challenging and suggest that this forecast may not be accurate. Nevertheless, the predicted increase in the prevalence of dementia is concerning, along with the the considerable economic and social burden associated with dementia.

There are diagnostic criteria for dementia, as well as its subtypes, but important assessments that facilitate an accurate diagnosis are not included; the lack of knowledge of the signs and symptoms could be considered a contributing factor.

Consequently, it is currently difficult and time consuming to diagnose dementia reliably. In short, this reinforces that conducting research into dementia is vital. It also provides some insight into the motivations behind this research and the two primary aims that are outlined in the following section.

As the four main dementia subtypes are considered throughout the thesis, the key criteria that characterise them are summarised. Firstly, AD is diagnosed when someone is exhibiting deficits in at least two cognitive domains, specifically progressive deterioration of memory and other cognitive functions (e.g. language, motor skills and perception) (McKhann et al., 1984). VD is diagnosed when there is evidence of cognitive decline and cerebrovascular disease (e.g. stroke), as well as a relationship between them (Román et al., 1993). DLB is diagnosed when there is progressive cognitive decline and symptoms such as fluctuations in attention and alertness, recurrent visual hallucinations and spontaneous movement abnormalities associated with Parkinson's disease (McKeith et al., 2005). Lastly, FTD is diagnosed when there is a gradual decline in someone's cognition that affects their personality and social conduct; their memory is relatively well preserved (Neary et al., 1998).

1.1 Research Aims

As indicated in the previous section, there were two primary aims for this research.

1. Investigate the use of machine learning for distinguishing between people with and without dementia, as well as differentiating between key dementia subtypes (AD, VD, DLB and FTD) where appropriate.

2. Gain an understanding of the inherent structure of dementia data, to ultimately investigate disease signatures.

The latter aim, in particular, allowed for some investigation into whether the prevailing diagnostic criteria accurately reflect the nature of dementia and its subtypes. Incidentally, a disease signature attempts to characterise a disease. Stemmer et al. (2019) proffer a detailed definition which specifies different aspects that should be considered, including causes and undesired effects of the disease.

These aims were tackled using classification and clustering respectively. Classification is the process of individually assigning new (unseen) observations to one of a number of classes. A classifier is constructed in order to achieve this, specifically by drawing on labels, relating to the classes, that are associated with observations comprising a training set; this is an example of supervised learning. Clustering, on the other hand, is an unsupervised learning technique that aims to discover groups (or clusters) of similar observations without utilising any associated labels. Information as to how classification and clustering was performed is included in section 1.3.

1.2 Data

Data was obtained from the National Alzheimer's Coordinating Center (NACC), ultimately due to it being one of the biggest and most comprehensive sources of its kind (National Alzheimer's Coordinating Center, 2019). In particular, the Uniform Data Set (UDS) was acquired, comprising data from visits to Alzheimer's Disease Centers (ADCs) situated across the United States. During a visit, the visitor (or subject) is assessed according to a standardised evaluation, administered using a number of different forms, in order to ascertain a diagnosis that essentially indicates whether they have dementia, along with the type of dementia if appropriate. 112,719 visits from September 2005 to February 2016, namely to 35 ADCs, were included in the data set obtained. More specifically, the data described 33,415 subjects, the majority of which had visited an ADC on more than one occasion, by means of 755 variables (or features).

The variables concerning diagnosis were extracted from the data set so labels (or classification targets) could be generated. The remaining data was cleansed, which involved identifying the variables of interest, along with the subjects to be analysed, for which only initial visits were considered. This resulted in a data set composed of 32,573 visits/subjects/observations and 260 variables; the latter of which were of continuous, categorical, ordinal and binary type. Notably, ordinal and binary variables can be considered to be types of categorical variables. Additionally, it involved improving the representation of missingness, of which there were two types,

as well as identifying and verifying relationships between variables. In particular, missingness arose when data was unexpectedly not recorded, or when the information was irrelevant or unobtainable for a known reason. The former genuinely missing values were imputed where possible, using the approach detailed in chapter 3. The latter values, however, were handled during classification and clustering, and are instances of what is termed *conditional missingness*. To illustrate, a conditionally missing value could arise from the question "In the past four weeks, did the subject have difficulty or need help with preparing a balanced meal?" (ADC Clinical Task Force and National Alzheimer's Coordinating Center, 2017c) if the subject had never performed this task. Furthermore, training and test sets were formed in preparation for classification. For more details regarding the data, the reader is directed to chapter 2.

1.3 Decision Tree Learning

Decision tree learning is fundamental to this thesis, which is why it is discussed here at the outset. A decision tree consists of two types of nodes, namely internal splitting nodes and terminal (or leaf) nodes, together with edges (or branches). Decision trees are essentially used to make decisions, but can be learnt for classification, along with regression, as explained by Breiman et al. (1984). As a result, there are two main types of decision trees within this field, specifically classification and regression trees. Due to the nature of the research conducted, the former (i.e. classification trees) are focused on here. There are a variety of algorithms which can be used to construct classification trees and Kotsiantis (2013) discusses a number of them. Nonetheless, the basic principles are largely the same. In brief, a tree is built from top to bottom by recursively partitioning a set of observations, forming a data set, based on the values of a variable. The variable can change over the course of construction and the feature space, defined by the set of variables, is partitioned. Each region of the partitioned feature space corresponds to one of the terminal nodes found at the base of the tree and it is these nodes that enable class predictions to be made.

Decision trees were chosen to aid in the investigation of the classification of

dementia and its main subtypes for a number of reasons. Firstly, they are able to handle categorical variables, along with continuous variables, without needing the former to be transformed in some way (James et al., 2017). Many alternative classifiers, such as artificial neural networks, are unable to do so. In fact, it is common practice to transform each categorical variable into a set of binary variables using one-hot encoding. By using decision trees and keeping the categorical variables intact, the number of variables is kept to a minimum; this helps to avoid the *curse of dimensionality* (Keogh and Mueen, 2017). Secondly, decision trees are interpretable and allow the importance of each variable for the classification task to be ascertained. It should be noted that variable importances enabled the key features for diagnosing dementia and differentiating between the main subtypes to be identified, and being able to maintain the identity of each of the categorical variables aided the process. Thirdly, powerful classifiers can be constructed by aggregating multiple trees, which are all different to some extent, to form an ensemble (or forest) (James et al., 2017). Ho (1995) demonstrates how generalisation, or the classifier's ability to handle new (unseen) data, can be improved if multiple trees as opposed to a single tree are utilised. Criminisi, Shotton and Konukoglu (2011) highlight that the popularity of decision trees can be attributed to the performance of ensembles.

The Extra-Trees algorithm (Geurts, Ernst and Wehenkel, 2006), which is closely related to the well-known Random Forests algorithm (Breiman, 2001), generates an ensemble of extremely randomised trees (or random forest). The algorithm was used, in conjunction with missingness incorporated in attributes (MIA) (Twala, Jones and Hand, 2008), to generate random forest classifiers for the NACC data. It was chosen for its accuracy, as well as its computational efficiency, as evidenced by Geurts, Ernst and Wehenkel (2006). The rest of this section discusses the Extra-Trees algorithm and provides insight into how it was employed for the NACC data. Prior to this, however, the section offers further explanation as to how a decision tree can be constructed, along with a brief overview of the Random Forests algorithm. Notably, chapter 3 describes how Extra-Trees and MIA were employed together during imputation, whilst chapter 4 presents the various results obtained for the

NACC data.

Before defining the notation pertaining to the data, which is utilised in the discussion that follows and beyond, it should be highlighted that decision trees were also employed for clustering. More specifically, an isolation forest (Liu, Ting and Zhou, 2008), consisting of unsupervised decision trees, was constructed to ascertain the similarity between observations; thus, enabling them to be clustered. The reader is directed to chapter 5 for details.

With regards to notation, the data set is denoted by X and the classification targets by Y. In particular, X is a design matrix of size $N-\text{by}-F$, where N is the number of observations (i.e. subjects) and F is the number of features (or variables). Each row of X pertains to a single subject, which is denoted by X_n, whilst each column of X is a variable X^f; thus, a value for a specific subject and variable can be designated as X_n^f. Y, however, is the column of targets or class labels. As the focus is on two-class classification, a subject X_n can be assigned one of two classes, specifically 0 or 1; Y_n is used to refer to the class of X_n. $Y^{(0)}$ and $Y^{(1)}$ denote, more generally, all the instances of class 0 and 1 respectively.

1.3.1 Decision Tree

A very high-level description of how a decision tree is built has been provided, but this section outlines, in more detail, how a standard binary decision tree can be constructed. A binary tree, in particular, splits the set of observations at each internal splitting node in two; thus, each of the nodes has two child nodes. The basic notation used for a decision tree is as follows: t represents the tree itself, η denotes an internal splitting node and ℓ corresponds to a terminal node.

The data set (or training set) X, as well as the classification targets Y, are required to construct a tree t. Initially, the complete set of observations is considered; and a split S on a variable X^f must be chosen which partitions X into X_L and X_R, along with Y into Y_L and Y_R (algorithm 1 line 15). The objective is to choose the best split, or the purest with regards to Y_L and Y_R, of all the possible splits on every variable. There are a number of metrics to assess the quality of a split, but S can be

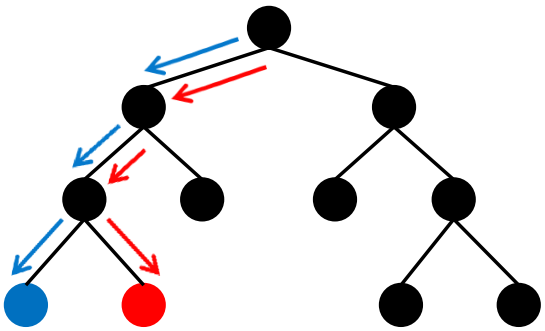


Figure 1.1: A simple diagram of a decision tree marked with two possible paths. One path is shown by blue arrows; the other by red arrows.

chosen so as to maximise the information gain $\mathcal{I}_S(X)$ which is described in detail below (equation 1.1). Generally, S is formed using a cut-point if X^f is continuous, or a subset if X^f is categorical. This particular split is associated with the first internal splitting node, which is also known as the root node. X_L and X_R, which satisfy $X_L \cup X_R = X$ and $X_L \cap X_R = \varnothing$, as well as Y_L and Y_R, are subsequently used to construct the left and right subtrees (algorithm 1 lines 16–19). In particular, X_L and Y_L enable the left subtree t_L to be built, whilst X_R and Y_R enable the right subtree t_R to be built. The process is repeated for each subtree (or child node), but a terminal node ℓ must be formed if certain criteria are met (algorithm 1 lines 8–12). Each terminal node can be labelled with the class frequencies for the set of observations that reach it.

A new (unseen) observation is classified by passing it through t, once the latter is fully formed. Figure 1.1 provides a visual representation of two paths through a simple decision tree. The paths are identical up to, but not including, the terminal node; the terminal node for each path is coloured accordingly to highlight this. As previously discussed, it is the terminal nodes that enable classifications to be made. In fact, an observation is assigned a class (0 or 1) in accordance with the majority of the observations that reached the same terminal node when constructing t. If it could be assigned more than one class, one can be chosen at random from the set of possible classes.

1.3.2 Random Forests

As previously explained, an ensemble (or forest) T is an aggregation of multiple (or M) different decision trees (algorithm 1 lines 1–6). As Breiman (2001) highlights, using an ensemble of trees can significantly improve classification accuracy; and predictions are made by allowing the trees to vote, in the manner in which a terminal node of a tree makes a prediction (i.e. by majority). A decision tree, as described in section 1.3.1, is built using all the observations in X, whilst the best split is selected for each internal node based on the complete set of variables. Constructing each member of a forest in this way would prove problematic as there would be no variability, but injecting some randomness into the process would allow a viable ensemble to be produced. The ensemble of trees generated, as a result, can be referred to as a random forest.

The Random Forests algorithm, originally termed Forest-RI, introduces randomness in two ways. Firstly, each tree is constructed using a different bootstrap sample of size N, which is generated by randomly sampling the observations in X with replacement; this is known as bootstrap aggregating (or bagging) (Breiman, 1996). Secondly, a different random subset of the variables is used in order to choose a split S for each internal node (algorithm 1 line 13). The size of the random subset K is decided on at the outset. Combining bagging with random feature selection ensures there is enough variation between the trees for the ensemble to be effective, whilst speeding up construction (James et al., 2017; Breiman, 2001). Bagging is also advantageous, however. In fact, it was found to improve the accuracy of the ensemble when random feature selection was employed, according to Breiman (2001); and it enables out-of-bag (OOB) estimates to be calculated, which eliminate the need for a test set. An OOB estimate of the classification error, for example, can be calculated by obtaining a prediction for every observation in X using only those trees trained on a bootstrap sample for which it was omitted.

1.3.3 Extra-Trees

Breiman (2001) suggests, in his seminal paper on random forests, that it may be possible to improve upon the Random Forests algorithm by introducing randomness in alternative ways; Geurts, Ernst and Wehenkel (2006) did just that with their Extra-Trees algorithm. The two algorithms are closely related, as highlighted in section 1.3, but there are two fundamental differences. Firstly, all the observations in X are used to train each tree instead of a bootstrap sample. Secondly, a single split is generated at random, as opposed to all the possible splits, for each of the variables considered for splitting (i.e. constituting the random subset) (algorithm 1 line 14). In short, Extra-Trees makes use of random feature selection, along with what could be termed *random split selection*.

Similarly to Random Forests, K variables are randomly selected to generate potential splits for an internal node. Geurts, Ernst and Wehenkel (2006) emphasise the fact that these variables must be chosen without replacement and should be inconstant, that is, the observations do not all have the same value. As a result, it is possible for less than K variables to be considered at any one time. The default value for K is $\sqrt{F}$ (rounded), where F is the number of variables in X. According to Geurts, Ernst and Wehenkel (2006), this value is generally deemed suitable for data sets with features of variable importance.

In contrast to Random Forests, a single random split is generated for every variable within the random subset. The manner in which each split is generated, however, is dependent on the type of the variable (algorithm 1 lines 21–35). If X^f is continuous, a cut-point f_{cp} is uniformly drawn between the minimum and maximum of X^f, and then used to split the set of observations in two (algorithm 1 lines 23–25). If X^f is categorical, the process is a little more involved (algorithm 1 lines 26–31). Initially, all the possible values present in X^f are identified. f is used to denote the set of possible values, whilst $f^{(in)}$ represents the set of values present. A proper nonempty subset of $f^{(in)}$ is randomly drawn (f_1), along with a subset of the values absent from X^f (f_2). f_1 and f_2 are subsequently combined, and the resultant subset is used to split the set of observations in two.

Algorithm 1 Pseudocode for Extra-Trees

function build_ensemble(X,Y)
  for $i \leftarrow 1, \ldots, M$ **do**
    $t_i \leftarrow$ **build_tree**(X,Y)
  end for
  return $T \leftarrow \{t_1, \ldots, t_M\}$
end function

function build_tree(X,Y)
  if $X^i \; \forall i \leftarrow 1, \ldots, F$ constant or Y constant or $|X| < n_{min}$ **then**
    $Y^{(0)} \leftarrow \{Y_n \in Y \,|\, Y_n = 0\}$
    $Y^{(1)} \leftarrow \{Y_n \in Y \,|\, Y_n = 1\}$
    return $\ell \leftarrow \{|Y^{(0)}|, |Y^{(1)}|\}$
  end if
  Randomly select K inconstant variables $\{X^{\varphi_1}, \ldots, X^{\varphi_K}\}$ without replacement
  Generate K splits $\{S_1, \ldots, S_K\}$
    where $S_i \leftarrow$ **generate_split**(X,X^{φ_i},Y) $\forall i \leftarrow 1, \ldots, K$
  Choose a split $S \triangleq \{(X_L, Y_L), (X_R, Y_R)\}$
    such that $\mathcal{I}_S(X) \leftarrow \max_{i \leftarrow 1,\ldots,K} \mathcal{I}_{S_i}(X)$ using equation 1.1
  $t_L \leftarrow$ **build_tree**(X_L,Y_L)
  $t_R \leftarrow$ **build_tree**(X_R,Y_R)
  Create η for S and attach t_L and t_R to form t
  return t
end function

function generate_split(X,X^f,Y)
  $\Lambda \leftarrow \{1, \ldots, |X|\}$
  if X^f continuous **then**
    Uniformly draw a cut-point f_{cp} in $(\min X^f, \max X^f)$
    $\Lambda_L \leftarrow \{i \in \Lambda \,|\, X_i^f < f_{cp}\}$
  else if X^f categorical **then**
    Identify all possible values present in X^f ($f^{(in)} \subseteq f$)
    Randomly draw $f_1 \subset f^{(in)}$ where $f_1 \neq \varnothing$
    Randomly draw $f_2 \subseteq f \setminus f^{(in)}$
    $\Lambda_L \leftarrow \{i \in \Lambda \,|\, X_i^f \in f_1 \cup f_2\}$
  end if
  $X_L \leftarrow \{X_i \; \forall i \in \Lambda_L\}$; $X_R \leftarrow X \setminus X_L$
  $Y_L \leftarrow \{Y_i \; \forall i \in \Lambda_L\}$; $Y_R \leftarrow Y \setminus Y_L$
  return $S \leftarrow \{(X_L, Y_L), (X_R, Y_R)\}$
end function

For each internal node, S is chosen from the pool of potential splits so as to maximise the information gain $\mathcal{I}_S(X)$. Geurts, Ernst and Wehenkel (2006) advocate the use of a normalisation of the information gain formulated by Wehenkel and Pavella (1991), which can be defined as

$$\mathcal{I}_S(X) = \frac{2(H_c(X) - H_{c\,|\,S}(X))}{H_c(X) + H_S(X)}\,, \tag{1.1}$$

where $H_c(X)$ is the classification entropy of X, $H_S(X)$ is the split entropy of X and $H_{c\,|\,S}(X)$ is the classification entropy of X given S. More specifically, $H_c(X)$ measures the uncertainty associated with classifying an observation in X, as does $H_{c\,|\,S}(X)$ given that X has been split according to S, whilst $H_S(X)$ assesses the impurity of S for X (Wehenkel and Pavella, 1991). These three terms can also be defined mathematically as follows:

$$H_c(X) = -[p(Y^{(1)}\,|\,X)\log_2 p(Y^{(1)}\,|\,X) + p(Y^{(0)}\,|\,X)\log_2 p(Y^{(0)}\,|\,X)]\,, \tag{1.2a}$$

$$H_S(X) = -[p(X_L\,|\,X)\log_2 p(X_L\,|\,X) + p(X_R\,|\,X)\log_2 p(X_R\,|\,X)]\,, \tag{1.2b}$$

$$H_{c\,|\,S}(X) = p(X_L\,|\,X)H_c(X_L) + p(X_R\,|\,X)H_c(X_R)\,. \tag{1.2c}$$

To put these equations into context, $p(Y^{(1)}\,|\,X)$ is the probability an observation in X is a member of class 1 (in equation 1.2a), and $p(X_L\,|\,X)$ is the probability an observation in X is sent to the left child node (in equations 1.2b and 1.2c).

In section 1.3.1 it was stated that a terminal node ℓ must be formed if certain criteria are met. Geurts, Ernst and Wehenkel (2006) specify three criteria which trigger a terminal node, namely all the variables in X are constant (i.e. all the observations are equivalent), the classification targets constituting Y are constant (or all equal) and the number of observations in X is less than n_{min} (algorithm 1 line 8). Only one criterion needs to be satisfied for a terminal node to be formed and setting $n_{min} = 2$ is a robust choice, according to Geurts, Ernst and Wehenkel (2006).

Extra-Trees was chosen not only for its accuracy but also its computational efficiency, as explained in section 1.3. In fact, Geurts, Ernst and Wehenkel (2006) show that their algorithm is faster than Random Forests and bagging, fundamentally

due to its use of random split selection which avoids an expensive search for the best possible split. A disadvantage of using Extra-Trees, however, is that OOB estimates cannot be calculated, meaning a test set is required. The authors note that bagging could be incorporated into the algorithm, enabling OOB estimates to be computed, but highlight that it typically reduces the accuracy of the algorithm. Bagging, in general, can also reduce interpretability (James et al., 2017).

The NACC data included variables of mixed type (continuous, categorical, ordinal and binary), as explained in section 1.2, but Geurts, Ernst and Wehenkel (2006) fail to explicitly specify how ordinal and binary variables should be handled. In short, ordinal variables were treated as if they were continuous, as their values had inherent order, along with binary variables. Binary variables could have been regarded as continuous or categorical in this context, but it is much simpler to generate a cut-point than a subset to split on. The NACC data also included conditionally missing values, which were handled using MIA (missingness incorporated in attributes) in conjunction with Extra-Trees. As stated in section 1.3, chapter 3 describes how Extra-Trees and MIA were employed together during imputation.

1.4 Contributions

Two machine learning approaches were developed for the purposes of this research. Firstly, an imputation approach was developed, which simultaneously builds a random forest classifier whilst handling conditionally missing values; it is termed *proximity imputation with MIA*. In particular, it can deal with mixed data and maintain the known relations between variables in the (NACC) data set, so far as possible. Secondly, a clustering approach was developed that ultimately measures the similarity of observations by means of an isolation forest, and is able to naturally draw on variables of mixed type. Notably, three (novel) isolation forest proximity (distance or similarity) measures were considered.

Of course, there are also contributions to dementia research. To summarise, a dementia classifier was constructed with an accuracy of 94.21%, a sensitivity of 0.93, a specificity of 0.95 and an area under the receiver operating characteristic curve

(AUC) of 0.99, suggesting machine learning could be a useful tool for diagnosing dementia. 10 pairwise dementia subtype classifiers were also generated with AUCs ranging from 0.88 to 1.0 (rounded to two decimal places), indicating machine learning could be used to differentiate between the main dementia subtypes. Using these classifiers, it was possible to identify the key features for diagnosing dementia, as well as differentiating between the main subtypes of dementia. Crucially, there is a clear difference between the important features for the two types of diagnosis. Last but not least, preliminary experiments conducted using the clustering approach developed suggested that mild cognitive impairment (MCI) may be a mild form of dementia as opposed to a clinical entity (i.e. a condition in its own right), over which there is much debate. They also suggested that there could be evidence for the current subtypes (AD, VD, DLB and FTD).

In conducting this research, numerous possible avenues for future research have been revealed, some of which are already being explored by another researcher. Most importantly, however, its findings have the potential to transform the way in which dementia is diagnosed. As a matter of fact, the key features identified, for both diagnosing dementia and differentiating between the main subtypes, could prove useful in redesigning and streamlining routine clinical practice. They may also help to improve dementia diagnosis, in more general terms, if the diagnostic criteria were updated accordingly. Furthermore, there is the potential to develop a diagnostic aid from the classifiers constructed. To clarify, this research is not immediately changing dementia diagnosis practice but is a foundation for change.

1.5 Thesis Overview

The thesis is organised into four main chapters.

- Chapter 2 discusses the NACC UDS, expanding on the introduction to the data set provided in section 1.2.
- Chapter 3 explains the imputation approach developed, as well as the experimental work carried out which helped to inform the parameters employed

and ascertained the effects of additional missingness on the imputation and classification performance. In addition, it provides a brief overview of related literature, specifically concerning methods for handling missing data, to put the work into context.

- Chapter 4 presents results from the imputation and classification of the NACC data. In particular, results pertaining to the dementia classifier and pairwise dementia subtype classifiers are included, along with those concerning a stacking classifier. It explains how these classifiers were built, whilst it also puts the work into context by discussing related literature and the clinical implications of the findings.

- Chapter 5 describes the clustering approach developed, as well as how it was tested using a variety of data sets, following a brief discussion of related work predominantly on clustering categorical and mixed data. It also gives an account of the preliminary experiments conducted on NACC data, for which the results are provided.

Finally, chapter 6 concludes the thesis and highlights some potential future research.

Chapter 2

National Alzheimer's Coordinating Center Uniform Data Set

This chapter expands on the introduction to the data set provided in chapter 1. It describes the variety of data constituting the data set; and gives an overview of the data cleansing process performed, including a discussion of the missing data present. It also outlines how labels were generated using the diagnosis data, primarily for the purposes of classification, as well as how training and test sets were created. Finally, it explains how each variable's predictive capacity was investigated.

2.1 Overview

The National Alzheimer's Coordinating Center Uniform Data Set (NACC UDS) contains data pertaining to Alzheimer's Disease Center (ADC) visits. The ADCs are situated at major medical institutions across the United States and are funded by the National Institute on Aging to carry out research into dementia (National Institute on Aging, 2019). Those visiting an ADC will typically have been referred for a clinical evaluation or invited to participate in a research study. Each visitor (or subject) undergoes a standardised evaluation which leads to a diagnosis that basically indicates whether they have dementia or not, as well as the type of dementia if so. Subjects are asked to bring along a co-participant, who is also questioned in order to provide supplementary information on the subject. All participants (i.e. subjects)

and co-participants are required to provide written informed consent. The majority of subjects attend follow-up visits, which are conducted on a yearly basis, meaning there can be multiple visits associated with a single subject.

NACC was chosen as the source of the data for this work as it is one of the biggest and most comprehensive of its kind (National Alzheimer's Coordinating Center, 2019). Morris et al. (2006), Beekly et al. (2007) and Weintraub et al. (2009) provide a detailed look at the UDS shortly after its inception. However, the researchers data dictionary (National Alzheimer's Coordinating Center, 2017) is the primary resource that was used throughout the research, and it discusses the variables in depth. Data collection is continuous; the data obtained from NACC was collected between September 2005 and February 2016. It included 112,719 visits to 35 ADCs, concerning 33,415 subjects and 755 variables.

2.2 Forms

The UDS is populated using forms, of which there have been 19 across three versions (1.2, 2.0 and 3.0) of the data set resulting in a total of 374, 407 and 523 variables for each version respectively. This section gives a very brief overview of the information collected by each form, in accordance with the researchers data dictionary (National Alzheimer's Coordinating Center, 2017), except for Form D1 which is considered in section 2.4. The variables derived by NACC and included in the UDS are discussed with the relevant forms. Table 2.1 provides a list of the forms covered and indicates the number of variables each of them gives rise to. 12 variables are associated with every form and make up the form header; these variables should be identical across the forms for a single visit. They provide the subject and center identification numbers; visit date, number and type; version; and statistics pertaining to visits, such as the total number.

Subject and Co-participant Demographics

Forms A1 and A2 obtain information about the subject and co-participant, resulting in 25 and 22 variables for each respectively. Basic data such as date of birth, sex and

Form	Name	Variables
	Form Header	12
A1	Subject Demographics	25
A2	Co-participant Demographics	22
A3	Family History	15
A4	Medications	62
A5	Health History	75
B1	Physical Examination	12
B2	Hachinski Ischemic Score and Cerebrovascular Disease	17
B3	Unified Parkinson's Disease Rating Scale	55
B4	Clinical Dementia Rating	10
B5	Neuropsychiatric Inventory Questionnaire	26
B6	Geriatric Depression Scale	17
B7	Functional Activities Questionnaire	10
B8	Physical/Neurological Exam Findings	47
B9	Clinician Judgment of Symptoms	59
C1	Neuropsychological Battery	48
C2	Neuropsychological Battery (version 3.0)	47
D2	Clinician-assessed Medical Conditions	33
	Milestones	16

Table 2.1: The forms used to populate the NACC UDS, except for Form D1, along with the number of variables associated with each of them. The form header is also included.

race are collected for both the subject and co-participant, along with the number of years they spent in education. The primary reason for the subject visiting the ADC and the principal referral source is recorded. The living situation and level of independence of the subject is also recorded. For the co-participant, the nature of their relationship with the subject, along with the type and frequency of their contact is noted. In addition, the perceived reliability of the co-participant is reported.

Family History, Medications and Health History

Forms A3, A4 and A5 record the subject's family history, medications and health history respectively. A total of 152 variables are produced, of which 15 describe family history, 62 provide details of medications and 75 report health history. Family history focuses on whether the subject has a first-degree family member with cognitive impairment, specifically their mother and/or father. Any evidence of gene mutations is also noted. Medications the subject has taken within the last two weeks are of interest, and the different types are identified from those provided. The data

concerning the subject's health history is quite extensive. It covers history of smoking, alcohol consumption, stroke and various psychiatric disorders to name a few.

Physical Examination

Form B1 is used to record the findings of the subject's physical examination, which is basic and generates just 12 variables. The subject's height and weight are measured, allowing their body mass index (BMI) to be calculated. The subject's blood pressure and resting heart rate are also assessed. Any issues with vision and hearing are reported, along with the subject's use of corrective lenses and hearing aids.

Hachinski Ischemic Score and Cerebrovascular Disease

Form B2 includes the eight clinical features required to calculate the modified Hachinski ischemic score (HIS), proposed by Rosen et al. (1980). It also collects information pertaining to cerebrovascular disease, particularly imaging evidence, resulting in a total of 17 variables. The HIS, specifically, is used to identify people with vascular dementia. The eight features considered include whether the subject experienced an abrupt onset of cognitive decline and stepwise deterioration. They also look at whether the subject has a history of stroke, and any focal neurological signs and symptoms.

Unified Parkinson's Disease Rating Scale

Form B3 records the results of the motor examination which forms part of the Unified Parkinson's Disease Rating Scale (UPDRS) (Fahn, Elton and UPDRS Development Committee, 1987). It produces 55 variables, some of which indicate the presence and severity of certain motor problems, such as tremors and rigidity. During the examination, the subject is asked to perform a number of tasks to enable any impairments to be identified. For example, one task involves the subject opening and closing their hands in rapid succession. The subject's speech, facial expression and posture are also inspected and rated.

Clinical Dementia Rating

Form B4 features the Clinical Dementia Rating (CDR®) Dementia Staging Instrument plus NACC FTLD Behaviour & Language Domains, the latter of which helps to identify cases of frontotemporal dementia (also referred to as frontotemporal lobar degeneration (FTLD)) and/or primary progressive aphasia (ADC Clinical Task Force and National Alzheimer's Coordinating Center, 2019). It gives rise to 10 variables. The CDR, as detailed by Morris (1993), assesses six categories; and a score indicating the perceived level of impairment is recorded for each. The categories are memory, orientation, judgment and problem-solving, community affairs, home and hobbies, and personal care. The scores are summed to yield the CDR *sum of boxes*. An overall score is also derived, namely the global CDR. The second component (NACC FTLD) involves the assessment of two additional constructs. The first encapsulates behaviour, comportment and personality; and the second is language.

Neuropsychiatric Inventory Questionnaire

Form B5 is the Neuropsychiatric Inventory Questionnaire (NPI-Q) (Kaufer et al., 2000). It assesses the presence and severity of a number of behavioural disorders for the month prior to the assessment, resulting in 26 variables. Examples of the disorders considered are appetite and eating problems, agitation or aggression, depression or dysphoria, hallucinations, and motor disturbances.

Geriatric Depression Scale

Form B6 comprises the Short Form of the Geriatric Depression Scale (GDS), which is discussed by Sheikh and Yesavage (1986); and generates 17 variables. It is used to screen the subject for depression and consists of 15 questions for them to answer, such as "Are you in good spirits most of the time?" (ADC Clinical Task Force and National Alzheimer's Coordinating Center, 2017c). Scores are accumulated across the questions and, ultimately, summed to produce the total GDS score.

Functional Activities Questionnaire

Form B7 is an adaptation of the Functional Activities Questionnaire (FAQ) presented by Pfeffer et al. (1982). It evaluates whether the subject has had any difficulty, or needed help, with an instrumental activity of daily living (IADL) in the past four weeks, and to what degree. 10 variables are generated, each corresponding to one of the 10 IADLs which are assessed. They cover tasks such as preparing a balanced meal; shopping alone for clothes, household necessities or groceries; writing checks, paying bills or balancing a checkbook; and remembering appointments, family occasions, holidays or medications.

Physical/Neurological Exam Findings

Form B8 captures the findings of the physical/neurological exam and enables the identification of a syndrome that could be responsible. It produces 47 variables, and does not record information pertaining to cognition or behaviour. There is some overlap with Form B3 but findings consistent with syndromes such as central nervous system disorder, posterior cortical atrophy, progressive supranuclear palsy and amyotrophic lateral sclerosis are also considered.

Clinician Judgment of Symptoms

Form B9 enables the symptoms the subject is experiencing to be identified, and looks into the nature and onset of these symptoms. Any decline in memory reported by the subject or co-participant is noted, along with whether the clinician believes there is a meaningful decline in cognition, behaviour or movement. Subdivisions of these domains are also considered, for example, personality change for behaviour. In total, the form gives rise to 59 variables.

Neuropsychological Battery

Form C1 results in 48 variables. It incorporates the Mini-Mental State Examination (MMSE) (Folstein, Folstein and McHugh, 2001), which is a neuropsychological battery

in itself that produces a score indicating cognitive impairment; and a number of neuropsychological tests, such as digit span forward. This test requires the subject to repeat number sequences of increasing length and is widely used to assess working memory (ADC Clinical Task Force and National Alzheimer's Coordinating Center, 2014a). Once the tests have been completed, the clinician rates the subject's cognitive status based on their performance. Form C2 replaced C1 for version 3.0 of the UDS. There is significant overlap between them but there are 47 variables solely associated with Form C2, most of which correspond to alternative neuropsychological tests.

Clinician-assessed Medical Conditions

Form D2, which produces 33 variables, provides the clinician with the opportunity to report any active medical conditions or procedures performed in the last 12 months. The clinician must have sufficient evidence for any medical conditions, which can be sleep disorders, diabetes, arthritis, cancer and congestive heart failure to name a few (ADC Clinical Task Force and National Alzheimer's Coordinating Center, 2017a). Procedures such as having a pacemaker and/or defibrillator fitted, and a heart valve replacement or repair are recorded.

Milestones

Information such as whether the subject is deceased; no longer visits an ADC; has permanently moved to a nursing home; or has additional data associated with them, specifically neuropathology or FTLD, is documented in Milestones. All 16 variables associated with this form are derived by NACC.

2.3 Data Cleansing

Discrepancies were apparent between the data obtained and the documentation for the UDS, such as missing variables and undocumented variable values. After receiving a number of corrected versions of the data set, it was necessary to extract the data of interest, improve the representation of the missing data present, and

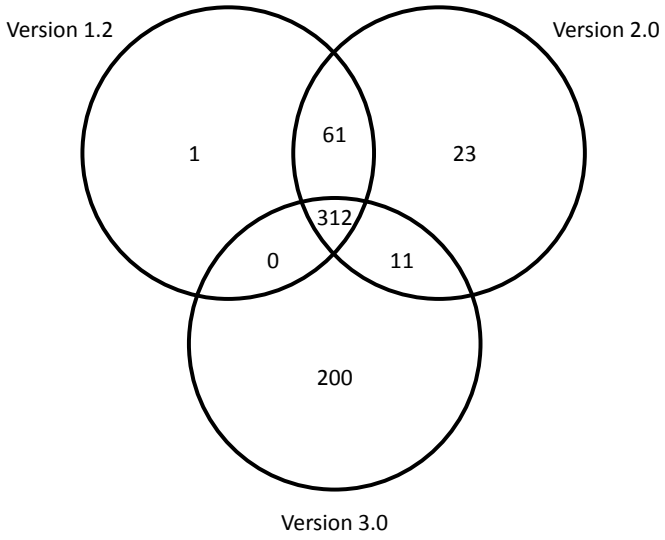


Figure 2.1: A Venn diagram showing the associations between the 608 variables considered and the three versions of the NACC UDS.

identify and verify relationships between variables. The forms were considered in turn to make the process more manageable; this section provides an explanation of each stage. For more information, specifically regarding how each variable was handled, the reader is referred to appendix A.

2.3.1 Visit and Variable Selection

Despite the UDS being longitudinal, the research undertaken was not; therefore, every subject's initial visit was extracted in the first instance. This reduced the number of visits from 112,719 to 33,415 and equalised the number of visits and subjects. 15,804 of these visits used version 1.2 of the UDS, 16,769 used version 2.0 and 842 used version 3.0. 608 of the 755 variables in the data set were scrutinised. The remaining 147 variables, which resulted from Form D1, were considered separately as they provide data pertaining to diagnosis; these variables are discussed in section 2.4.

Figure 2.1 highlights that significant changes were made to the UDS when version 3.0 was introduced. 200 variables were added and 84 were removed. In order to maximise the number of visits and variables, the 842 visits which employed version 3.0 were dropped, and the variables were initially restricted to those included in both versions 1.2 and 2.0. Consequently, 32,573 visits and 373 variables were retained. These visits fell between September 2005 and February 2015 and were to 35 ADCs.

The variables excluded from versions 1.2 and 2.0, of which there were 35 in total, were not considered to provide key information.

Additional variables were excluded from the 373 preserved for a number of reasons, such as the variable contained free-text (e.g. other Hispanic origins), provided irrelevant information (e.g. ADC identification number) or was constant across the subjects (e.g. number of days from initial visit). However, a single constant variable providing the visit number (NACCVNUM) was retained for testing purposes, which are explained in chapter 4. Duplicate information was removed where possible, mainly on a form-by-form basis; and a selection of variables were replaced by some which had been newly derived, in order to consolidate data and provide it in a more suitable format. For example, the two variables indicating whether the subject had experienced hallucinations in the last month, and the severity of the hallucinations, were combined to form a single variable (HALL_SEV). This resulted in a total of 258 variables, which are detailed in table 2.2. Two of these variables, specifying the subject's identification number (NACCID) and visit date (VISIT_DATE), were maintained for administrative reasons; and were excluded when analysis was applied to the data set. Of the other 256 variables, 39 were continuous, 63 were categorical, 60 were ordinal and 94 were binary.

Prior to analysis, four randomly generated synthetic variables were introduced into the data set, increasing the total number of variables to 260. The exact nature of these variables is discussed in chapter 4, as they were added for testing purposes.

2.3.2 Missing Data

The UDS contains two types of missing values. The first arises when data is unexpectedly not recorded, and the second occurs for a number of reasons. The main reason is the associated question was not relevant, either in its own right or because a response to a previous question rendered it so. An example of a question for the former case, from Form B7, is "In the past four weeks, did the subject have difficulty or need help with preparing a balanced meal?" (ADC Clinical Task Force and National Alzheimer's Coordinating Center, 2017c). In short, the

Variable	Description	Data Type	Form
NACCID	Subject identification number	NACC identifier	FH
FORMVER	Form version number	Continuous	FH
VISIT_DATE*	Form date (year, month, day)	Date/Time object	FH
NACCVNUM	UDS visit number (order)	Continuous	FH
NACCREAS	Primary reason for coming to an Alzheimer's Disease Center (ADC)	Categorical	A1
NACCREFR	Principal referral source	Categorical	A1
BIRTH_#MOS*	Months from subject's month/year of birth to month/year of visit	Continuous	A1
SEX	Subject's sex	Binary	A1
HISPANIC	Hispanic/Latino ethnicity	Binary	A1
HISPOR	Hispanic origins	Categorical	A1
PRIMLANG	Primary language	Categorical	A1
EDUC	Years of education	Continuous	A1
NACCLIVS	Living situation	Categorical	A1
INDEPEND	Level of independence	Ordinal	A1
RESIDENC	Type of residence	Categorical	A1
MARISTAT	Marital status	Categorical	A1
HANDED	Is the subject left- or right-handed?	Categorical	A1
NACCAGE	Subject's age at visit	Continuous	A1
NACCNIHR	Derived National Institutes of Health (NIH) race definitions	Categorical	A1
INBIR_#MOS*	Months from co-participant's month/year of birth to month/year of visit	Continuous	A2
INSEX	Co-participant's sex	Binary	A2
INRELTO	Co-participant's relationship to subject	Categorical	A2
INLIVWTH	Does the co-participant live with the subject?	Binary	A2
INVISITS	If no, approximate frequency of in-person visits?	Ordinal	A2
INCALLS	If no, approximate frequency of telephone contact?	Ordinal	A2
INRELY	Is there a question about the co-participant's reliability?	Binary	A2
NACCMOM	Indicator of mother with cognitive impairment	Binary	A3

NACCDAD	Indicator of father with cognitive impairment	Binary	A3
NACCFAM	Indicator of first-degree family member with cognitive impairment	Binary	A3
ANYMEDS	Subject taking any medications	Binary	A4
NACCAMD	Total number of medications reported at each visit	Continuous	A4
NACCHTNC	Reported current use of an antihypertensive combination therapy	Binary	A4
NACCACEI	Reported current use of an angiotensin converting enzyme (ACE) inhibitor	Binary	A4
NACCAAAS	Reported current use of an antiadrenergic agent	Binary	A4
NACCBETA	Reported current use of a beta-adrenergic blocking agent (Beta-Blocker)	Binary	A4
NACCCCBS	Reported current use of a calcium channel blocking agent	Binary	A4
NACCDIUR	Reported current use of a diuretic	Binary	A4
NACCVASD	Reported current use of a vasodilator	Binary	A4
NACCANGI	Reported current use of an angiotensin II inhibitor	Binary	A4
NACCAHTN	Reported current use of any type of antihypertensive or blood pressure medication	Binary	A4
NACCLIPL	Reported current use of lipid lowering medication	Binary	A4
NACCNSD	Reported current use of nonsteroidal anti-inflammatory medication	Binary	A4
NACCAC	Reported current use of an anticoagulant or antiplatelet agent	Binary	A4
NACCADEP	Reported current use of an antidepressant	Binary	A4
NACCAPSY	Reported current use of an antipsychotic agent	Binary	A4
NACCAANX	Reported current use of an anxiolytic, sedative or hypnotic agent	Binary	A4
NACCPDMD	Reported current use of an antiparkinson agent	Binary	A4
NACCEMD	Reported current use of estrogen hormone therapy	Binary	A4
NACCEPMD	Reported current use of estrogen + progestin hormone therapy	Binary	A4
NACCDBMD	Reported current use of a diabetes medication	Binary	A4
CVHATT	Heart attack/cardiac arrest	Categorical	A5
CVAFIB	Atrial fibrillation	Categorical	A5
CVANGIO	Angioplasty/endarterectomy/stent	Categorical	A5
CVBYPASS	Cardiac bypass procedure	Categorical	A5
CVPACE	Pacemaker	Categorical	A5

CVCHF	Congestive heart failure	Categorical	A5
CVOTHR	Other cardiovascular disease	Categorical	A5
CBSTROKE	Stroke	Categorical	A5
NACCSTYR_#YRS*	Years from most recently reported year of stroke as of the initial visit to year of visit	Continuous	A5
CBTIA	Transient ischemic attack (TIA)	Categorical	A5
NACCTIYR_#YRS*	Years from most recently reported year of TIA as of the initial visit to year of visit	Continuous	A5
PD	Parkinson's disease (PD)	Binary	A5
PDYR_#YRS*	Years from year of PD diagnosis to year of visit	Continuous	A5
PDOTHR	Other parkinsonian disorder	Binary	A5
PDOTHRYR_#YRS*	Years from year of parkinsonian disorder diagnosis to year of visit	Continuous	A5
SEIZURES	Seizures	Categorical	A5
TRAUMBRF	Brain trauma - brief unconsciousness	Categorical	A5
TRAUMEXT	Brain trauma - extended unconsciousness	Categorical	A5
TRAUMCHR	Brain trauma - chronic deficit	Categorical	A5
NCOTHR	Other neurological condition	Categorical	A5
HYPERTEN	Hypertension	Categorical	A5
HYPERCHO	Hypercholesterolemia	Categorical	A5
DIABETES	Diabetes	Categorical	A5
B12DEF	Vitamin B12 deficiency	Categorical	A5
THYROID	Thyroid disease	Categorical	A5
INCONTU	Incontinence - urinary	Categorical	A5
INCONTF	Incontinence - bowel	Categorical	A5
DEP2YRS	Active depression in the last two years	Binary	A5
DEPOTHR	Depression episodes more than two years ago	Binary	A5
ALCOHOL	Alcohol abuse - clinically significant impairment occurring over a 12-month period manifested in one of the following areas: work, driving, legal or social	Categorical	A5

TOBAC30	Smoked cigarettes in last 30 days	Binary	A5
TOBAC100	Smoked more than 100 cigarettes in life	Binary	A5
SMOKYRS	Total years smoked cigarettes	Continuous	A5
PACKSPER	Average number of packs smoked per day	Ordinal	A5
QUITSMOK	If the subject quit smoking, age at which he/she last smoked (i.e. quit)	Continuous	A5
ABUSOTHR	Other abused substances - clinically significant impairment occurring over a 12-month period manifested in one of the following areas: work, driving, legal or social	Categorical	A5
PSYCDIS	Other psychiatric disorder	Categorical	A5
NACCTBI	History of traumatic brain injury (TBI)	Binary	A5
HEIGHT	Subject's height (inches)	Continuous	B1
WEIGHT	Subject's weight (lbs)	Continuous	B1
BPSYS	Subject blood pressure (sitting), systolic	Continuous	B1
BPDIAS	Subject blood pressure (sitting), diastolic	Continuous	B1
HRATE	Subject resting heart rate (pulse)	Continuous	B1
VISION	Without corrective lenses, is the subject's vision functionally normal?	Binary	B1
VISCORR	Does the subject usually wear corrective lenses?	Binary	B1
VISWCORR	If the subject usually wears corrective lenses, is the subject's vision functionally normal with corrective lenses?	Binary	B1
HEARING	Without a hearing aid(s), is the subject's hearing functionally normal?	Binary	B1
HEARAID	Does the subject usually wear a hearing aid(s)?	Binary	B1
HEARWAID	If the subject usually wears a hearing aid(s), is the subject's hearing functionally normal with a hearing aid(s)?	Binary	B1
NACCBMI	Body mass index (BMI)	Continuous	B1
ABRUPT	Abrupt onset (re: cognitive status)	Binary	B2
STEPWISE	Stepwise deterioration (re: cognitive status)	Binary	B2
SOMATIC	Somatic complaints	Binary	B2
EMOT	Emotional incontinence	Binary	B2
HXHYPER	History or presence of hypertension	Binary	B2

HXSTROKE	History of stroke	Binary	B2
FOCLSYM	Focal neurological symptoms	Binary	B2
FOCLSIGN	Focal neurological signs	Binary	B2
HACHIN	Hachinski ischemic score	Continuous	B2
SPEECH	Speech	Ordinal	B3
FACEXP	Facial expression	Ordinal	B3
TRESTFAC	Tremor at rest - face, lips, chin	Ordinal	B3
TRESTRHD	Tremor at rest - right hand	Ordinal	B3
TRESTLHD	Tremor at rest - left hand	Ordinal	B3
TRESTRFT	Tremor at rest - right foot	Ordinal	B3
TRESTLFT	Tremor at rest - left foot	Ordinal	B3
TRACTRHD	Action or postural tremor - right hand	Ordinal	B3
TRACTLHD	Action or postural tremor - left hand	Ordinal	B3
RIGDNECK	Rigidity - neck	Ordinal	B3
RIGDUPRT	Rigidity - right upper extremity	Ordinal	B3
RIGDUPLF	Rigidity - left upper extremity	Ordinal	B3
RIGDLORT	Rigidity - right lower extremity	Ordinal	B3
RIGDLOLF	Rigidity - left lower extremity	Ordinal	B3
TAPSRT	Finger taps - right hand	Ordinal	B3
TAPSLF	Finger taps - left hand	Ordinal	B3
HANDMOVR	Hand movements - right hand	Ordinal	B3
HANDMOVL	Hand movements - left hand	Ordinal	B3
HANDALTR	Alternating movement - right hand	Ordinal	B3
HANDALTL	Alternating movement - left hand	Ordinal	B3
LEGRT	Leg agility - right leg	Ordinal	B3
LEGLF	Leg agility - left leg	Ordinal	B3
ARISING	Arising from chair	Ordinal	B3
POSTURE	Posture	Ordinal	B3
GAIT	Gait	Ordinal	B3

POSSTAB	Posture stability	Ordinal	B3
BRADYKIN	Body bradykinesia and hypokinesia	Ordinal	B3
PDNORMAL	Unified Parkinson's Disease Rating Scale (UPDRS) normal	Binary	B3
MEMORY	Memory	Ordinal	B4
ORIENT	Orientation	Ordinal	B4
JUDGMENT	Judgment and problem-solving	Ordinal	B4
COMMUN	Community affairs	Ordinal	B4
HOMEHOBB	Home and hobbies	Ordinal	B4
PERSCARE	Personal care	Ordinal	B4
CDRSUM	Clinical Dementia Rating (CDR) sum of boxes	Continuous	B4
CDRGLOB	Global CDR	Ordinal	B4
NPIQINF	Neuropsychiatric Inventory Questionnaire (NPI-Q) co-participant	Categorical	B5
DEL_SEV*	Delusions and their severity in the last month	Ordinal	B5
HALL_SEV*	Hallucinations and their severity in the last month	Ordinal	B5
AGIT_SEV*	Agitation or aggression, and the severity, in the last month	Ordinal	B5
DEPD_SEV*	Depression or dysphoria, and the severity, in the last month	Ordinal	B5
ANX_SEV*	Anxiety and the severity in the last month	Ordinal	B5
ELAT_SEV*	Elation or euphoria, and the severity, in the last month	Ordinal	B5
APA_SEV*	Apathy or indifference, and the severity, in the last month	Ordinal	B5
DISN_SEV*	Disinhibition and the severity in the last month	Ordinal	B5
IRR_SEV*	Irritability or lability, and the severity, in the last month	Ordinal	B5
MOT_SEV*	Motor disturbance and the severity in the last month	Ordinal	B5
NITE_SEV*	Nighttime behaviours and their severity in the last month	Ordinal	B5
APP_SEV*	Appetite and eating problems, and their severity, in the last month	Ordinal	B5
NOGDS	Is the subject able to complete the Geriatric Depression Scale (GDS), based on the clinician's best judgment?	Binary	B6
SATIS	Are you basically satisfied with your life?	Binary	B6
DROPACT	Have you dropped many of your activities and interests?	Binary	B6
EMPTY	Do you feel that your life is empty?	Binary	B6

BORED	Do you often get bored?	Binary	B6
SPIRITS	Are you in good spirits most of the time?	Binary	B6
AFRAID	Are you afraid that something bad is going to happen to you?	Binary	B6
HAPPY	Do you feel happy most of the time?	Binary	B6
HELPLESS	Do you often feel helpless?	Binary	B6
STAYHOME	Do you prefer to stay at home, rather than going out and doing new things?	Binary	B6
MEMPROB	Do you feel you have more problems with memory than most?	Binary	B6
WONDRFUL	Do you think it is wonderful to be alive now?	Binary	B6
WRTHLESS	Do you feel pretty worthless the way you are now?	Binary	B6
ENERGY	Do you feel full of energy?	Binary	B6
HOPELESS	Do you feel that your situation is hopeless?	Binary	B6
BETTER	Do you think that most people are better off than you are?	Binary	B6
NACCGDS	Total GDS score	Continuous	B6
BILLS	In the past four weeks, did the subject have difficulty or need help with: writing checks, paying bills or balancing a checkbook	Ordinal	B7
TAXES	In the past four weeks, did the subject have difficulty or need help with: assembling tax records, business affairs or other papers	Ordinal	B7
SHOPPING	In the past four weeks, did the subject have difficulty or need help with: shopping alone for clothes, household necessities or groceries	Ordinal	B7
GAMES	In the past four weeks, did the subject have difficulty or need help with: playing a game of skill such as bridge or chess, or working on a hobby	Ordinal	B7
STOVE	In the past four weeks, did the subject have difficulty or need help with: heating water, making a cup of coffee, or turning off the stove	Ordinal	B7
MEALPREP	In the past four weeks, did the subject have difficulty or need help with: preparing a balanced meal	Ordinal	B7
EVENTS	In the past four weeks, did the subject have difficulty or need help with: keeping track of current events	Ordinal	B7

PAYATTN	In the past four weeks, did the subject have difficulty or need help with: paying attention to and understanding a TV programme, book or magazine	Ordinal	B7
REMDATES	In the past four weeks, did the subject have difficulty or need help with: remembering appointments, family occasions, holidays or medications	Ordinal	B7
TRAVEL	In the past four weeks, did the subject have difficulty or need help with: travelling out of the neighbourhood, driving, or arranging to take public transportation	Ordinal	B7
NACCNREX	Were all findings unremarkable?	Binary	B8
FOCLDEF	Are focal deficits present indicative of central nervous system disorder?	Binary	B8
GAITDIS	Is gait disorder present indicative of central nervous system disorder?	Binary	B8
EYEMOVE	Are there eye movement abnormalities present indicative of central nervous system disorder?	Binary	B8
DECSUB	Does the subject report a decline in memory (relative to previously attained abilities)?	Binary	B9
DECIN	Does the co-participant report a decline in subject's memory (relative to previously attained abilities)?	Binary	B9
DECCLIN	Clinician believes there is a meaningful decline in memory, non-memory cognitive abilities, behaviour, ability to manage his/her affairs, or there are motor/movement changes	Binary	B9
DECAGE	Based on clinician's assessment, at what age did the cognitive decline begin?	Continuous	B9
COGMEM	Indicate whether the subject currently is meaningfully impaired, relative to previously attained abilities, in memory	Binary	B9
COGJUDG	Indicate whether the subject currently is meaningfully impaired, relative to previously attained abilities, in executive function - judgment, planning or problem-solving	Binary	B9
COGLANG	Indicate whether the subject currently is meaningfully impaired, relative to previously attained abilities, in language	Binary	B9

COGVIS	Indicate whether the subject currently is meaningfully impaired, relative to previously attained abilities, in visuospatial function	Binary	B9
COGATTN	Indicate whether the subject currently is meaningfully impaired, relative to previously attained abilities, in attention or concentration	Binary	B9
COGOTHR	Indicate whether the subject currently is meaningfully impaired, relative to previously attained abilities, in other cognitive domains	Binary	B9
NACCCOGF	Indicate the predominant symptom that was first recognised as a decline in the subject's cognition	Categorical	B9
COGMODE	Mode of onset of cognitive symptoms	Categorical	B9
BEAPATHY	Subject currently manifests meaningful change in behaviour - Apathy, withdrawal	Binary	B9
BEDEP	Subject currently manifests meaningful change in behaviour - Depressed mood	Binary	B9
BEVHALL	Subject currently manifests meaningful change in behaviour - Psychosis - Visual hallucinations	Binary	B9
BEAHALL	Subject currently manifests meaningful change in behaviour - Psychosis - Auditory hallucinations	Binary	B9
BEDEL	Subject currently manifests meaningful change in behaviour - Psychosis - Abnormal, false or delusional beliefs	Binary	B9
BEDISIN	Subject currently manifests meaningful change in behaviour - Disinhibition	Binary	B9
BEIRRIT	Subject currently manifests meaningful change in behaviour - Irritability	Binary	B9
BEAGIT	Subject currently manifests meaningful change in behaviour - Agitation	Binary	B9
BEPERCH	Subject currently manifests meaningful change in behaviour - Personality change	Binary	B9
BEOTHR	Subject currently manifests meaningful change in behaviour - Other	Binary	B9
NACCBEHF	Indicate the predominant symptom that was first recognised as a decline in the subject's behaviour	Categorical	B9
BEMODE	Mode of onset of behavioural symptoms	Categorical	B9

MOGAIT	Indicate whether the subject currently has meaningful changes in motor function - Gait disorder	Binary	B9
MOFALLS	Indicate whether the subject currently has meaningful changes in motor function - Falls	Binary	B9
MOTREM	Indicate whether the subject currently has meaningful changes in motor function - Tremor	Binary	B9
MOSLOW	Indicate whether the subject currently has meaningful changes in motor function - Slowness	Binary	B9
NACCMOTF	Indicate the predominant symptom that was first recognised as a decline in the subject's motor function	Categorical	B9
MOMODE	Mode of onset of motor symptoms	Categorical	B9
COURSE	Overall course of decline of cognitive/behavioural/motor syndrome	Categorical	B9
FRSTCHG	Indicate the predominant domain that was first recognised as changed in the subject	Categorical	B9
MMSELOC	Administration of the Mini-Mental State Examination (MMSE) was:	Categorical	C1
MMSELAN	Language of MMSE administration	Categorical	C1
MMSEORDA	Orientation subscale score - Time	Continuous	C1
MMSEORDA_PROB*	Reason an answer was not provided for MMSEORDA	Categorical	C1
MMSEORLO	Orientation subscale score - Place	Continuous	C1
MMSEORLO_PROB*	Reason an answer was not provided for MMSEORLO	Categorical	C1
NACCMMSE	Total MMSE score (using D-L-R-O-W)	Continuous	C1
NACCMMSE_PROB*	Reason an answer was not provided for NACCMMSE	Categorical	C1
NPSYCLOC	The remainder of the battery was administered:	Categorical	C1
NPSYLAN	Language of test administration	Categorical	C1
LOGIMEM	Total number of story units recalled from this current test administration	Continuous	C1
LOGIMEM_PROB*	Reason an answer was not provided for LOGIMEM	Categorical	C1
DIGIF	Digit span forward trials correct	Continuous	C1
DIGIFLEN	Digit span forward length	Continuous	C1
DIGIF_PROB*	Reason an answer was not provided for DIGIF	Categorical	C1

DIGIFLEN_PROB*	Reason an answer was not provided for DIGIFLEN	Categorical	C1
DIGIB	Digit span backward trials correct	Continuous	C1
DIGIBLEN	Digit span backward length	Continuous	C1
DIGIB_PROB*	Reason an answer was not provided for DIGIB	Categorical	C1
DIGIBLEN_PROB*	Reason an answer was not provided for DIGIBLEN	Categorical	C1
ANIMALS	Animals - Total number of animals named in 60 seconds	Continuous	C1
ANIMALS_PROB*	Reason an answer was not provided for ANIMALS	Categorical	C1
VEG	Vegetables - Total number of vegetables named in 60 seconds	Continuous	C1
VEG_PROB*	Reason an answer was not provided for VEG	Categorical	C1
TRAILA	Trail Making Test Part A - Total number of seconds to complete	Continuous	C1
TRAILA_PROB*	Reason an answer was not provided for TRAILA	Categorical	C1
TRAILB	Trail Making Test Part B - Total number of seconds to complete	Continuous	C1
TRAILB_PROB*	Reason an answer was not provided for TRAILB	Categorical	C1
WAIS	Wechsler Adult Intelligence Scale (Revised) (WAIS-R) Digit Symbol	Continuous	C1
WAIS_PROB*	Reason an answer was not provided for WAIS	Categorical	C1
MEMUNITS	Logical Memory IIA - Delayed - Total number of story units recalled	Continuous	C1
MEMTIME	Logical Memory IIA - Delayed - Time elapsed since Logical Memory IA - Immediate	Continuous	C1
MEMUNITS_PROB*	Reason an answer was not provided for MEMUNITS	Categorical	C1
BOSTON	Boston Naming Test (30) - Total score	Continuous	C1
BOSTON_PROB*	Reason an answer was not provided for BOSTON	Categorical	C1
COGSTAT	Per clinician, based on the neuropsychological examination, the subject's cognitive status is deemed	Categorical	C1
NACCC1	Form date discrepancy between UDS Form A1 and Form C1	Binary	C1

Table 2.2: Variables utilised from the NACC UDS, excluding those pertaining to diagnosis. The variables marked with an * were newly derived. Each variable's name, description and data type is provided, along with the form it is associated with, where FH denotes the form header. The descriptions are based on those provided in the researchers data dictionary (National Alzheimer's Coordinating Center, 2017).

subject may never have performed this task, making the question irrelevant. An example for the latter case, from Form A1, includes "Does the subject report being of Hispanic/Latino ethnicity (i.e. having origins from a mainly Spanish-speaking Latin American country), regardless of race?" and "If yes, what are the subject's reported origins?" (ADC Clinical Task Force and National Alzheimer's Coordinating Center, 2017c). Fundamentally, a negative response to the first question renders the follow-up question inconsequential. Other reasons are the clinician believed the subject was unable to be assessed in the prescribed way; and the corresponding form did not have to be completed and was not due to the subject having a physical, cognitive/behaviour or other problem, or they verbally refused. This second type of missing data is termed *conditional missingness*. Values indicating that the variable's question did not feature in the version of the UDS used to assess the subject can be associated with both types of missingness, but they were eliminated when variables were excluded from the data set.

Numerical codes are used by NACC to represent non-numerical values for variables, including missing and conditionally missing values. These codes were inconsistent across the data set, and some even represented different types of missingness for a single variable. Each variable was interrogated individually so the meaning of each code pertaining to missingness could be identified. It was important that the two types of missingness were uniformly labelled across all the variables and the distinction between them was clear, so that they could be handled appropriately. Any missingness was recoded accordingly, and the values indicated as genuinely missing were later imputed as described in chapter 3. In contrast, the values which were designated conditionally missing were handled, as their presence was meaningful. As a result, any missing values that could not be sensibly imputed, for example those associated with a variable providing the number of years since the subject's most recent stroke (NACCSTYR_#YRS), were marked as conditionally missing. It was previously stated that conditionally missing values could result due to the relevant form not having to be completed. These values also needed to be identifiable so they were not drawn on during imputation, ultimately to avoid introducing bias towards

conditionally missing values.

The percentage of subjects missing each form is included in appendix A, along with the proportions of the different types of missingness for the relevant variables. NACCBMI, which provides the subject's body mass index (BMI), had the most genuinely missing values, accounting for 10.55%. The average amount of missing values per variable, however, was just 0.66%. The percentage of missing values for the data set as a whole also equalled 0.66%, but the number of subjects with at least one missing value totalled 15,494 (47.57%). The significant proportion of subjects with missing values was the main motivation for performing imputation, rather than simply discarding the subjects affected. By imputing the missing values, the difficult task of handling two types of missingness was also avoided.

2.3.3 Variable Relationships

Section 2.3.2 implied there are relationships between variables within the UDS, due to there being links between questions in the forms. The relations were of interest as a result of missing values being present, and those featured in the UDS were split into two groups. The first comprised the relationships involving only two variables, where one (parent variable) can cause the other (child variable) to have a conditionally missing value if it takes on a specific value itself. These types of relationships are referred to as *dependencies*; and the example concerning Hispanic/Latino ethnicity, provided in section 2.3.2, is representative of this group.

The diagram in figure 2.2 shows the dependencies between the 21 variables used from Form A4, which provide the medications the subject took in the two weeks prior to their visit. There is a node to represent the form itself, which does not have a label and is coloured purple, as well as each of the variables. The ANYMEDS variable, which indicates whether the subject has taken any medications, is linked to the form node and every other variable. The former link signifies that the variable is associated with Form A4, whilst the latter links indicate that the other 20 variables are dependent on it. These variables provide information such as the total number of medications reported, and highlight different types of medications that have been

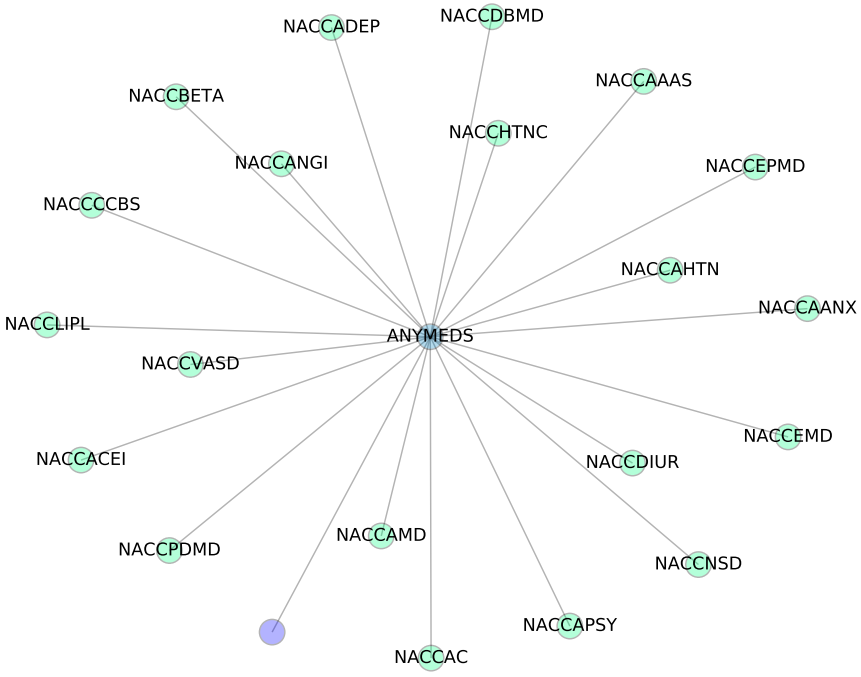


Figure 2.2: A visualisation of the dependencies between the 21 variables used from Form A4, which provide the medications the subject took in the two weeks prior to their visit.

taken (e.g. diabetes medication). It, therefore, makes sense that the values for these variables are dependent on whether or not any medications have been taken. Appendix A features a similar diagram for each of the forms.

The second group of relationships included those for which one or more variables can determine the value of another. The vast majority of these relationships arose due to a number of variables needing to be split in two, specifically to separate out two sets of data contained within them pertaining to different things. An example of a relationship, from Form B1, involving three variables providing the height (HEIGHT), weight (WEIGHT) and BMI (NACCBMI) of the subject can be characterised using an equation. More specifically, BMI $= (w \times 703)/h^2$, where w is weight (in pounds) and h is height (in inches) (National Alzheimer's Coordinating Center, 2017). The interactions between these variables constitute a relationship as the HEIGHT and WEIGHT variables determine the value of NACCBMI. For more examples of relations in the UDS, especially those which were considered during imputation, please refer to section 3.2.5.

The dependencies and relationships were not all clearly indicated in the documentation for the UDS, so it was necessary to identify and verify them. In fact,

some relations that were stated in the documentation were not found to hold in the data set, whilst some were present in the data set but not documented.

A number of the variables which can determine others (e.g. parent variables) had missing values; thus, the values in their associated variables corresponding to those that were missing had to be identified. It was important to set out which of these values would be updated, if the missing values were imputed, to ensure the relations within the UDS were maintained so far as possible. The majority of the values were either missing or conditionally missing, which were suitable to update for the most part. Any measured values were retained, however, even if they could be recalculated. An example of a variable which did not have its values updated is NACCGDS from Form B6. In particular, it provides the total GDS score which is calculated using 15 other variables, all of which had missing values that were imputed. Measured values which could potentially be updated arose as a total could still be calculated even if the values of up to three of the 15 variables were missing.

2.4 Diagnosis Data

Data pertaining to the subject's diagnosis is collected using Form D1. It generates 147 variables which mainly provide information as to the cause of any cognitive impairment. Dementia subtypes such as Alzheimer's disease (AD), vascular dementia (VD), dementia with Lewy bodies (DLB) and frontotemporal dementia (FTD) are considered, along with others which are less common. Neurological, genetic and infectious conditions, including epilepsy, Down's syndrome and human immunodeficiency virus (HIV) respectively; and psychiatric diseases, such as depression and schizophrenia, are also possible diagnoses. Furthermore, cognitive impairment due to substance abuse or medications is considered. For the majority of the diagnoses, whether it is a primary, contributing or non-contributing cause of the cognitive impairment is indicated. It is possible for multiple causes to be reported, but just one must be designated as the primary cause. This form also records whether the subject has any hereditary mutations, biomarker findings or imaging evidence, which could point towards a cause; and indicates if the diagnosis was made by a single

Variable	Description
NACCUDSD	Cognitive status at UDS visit
NACCALZD	Presumptive etiologic diagnosis of the cognitive disorder - Alzheimer's disease
VASC	Presumptive etiologic diagnosis of the cognitive disorder - Probable vascular dementia
VASCPS	Presumptive etiologic diagnosis of the cognitive disorder - Possible vascular dementia
STROKE	Presumptive etiologic diagnosis - Stroke
NACCLBDE	Presumptive etiologic diagnosis - Lewy body disease (dementia with Lewy bodies or Parkinson's disease)
NACCBVFT	Behavioural variant frontotemporal dementia syndrome
NACCPPA	Primary progressive aphasia with cognitive impairment
NACCETPR	Primary etiologic diagnosis of the cognitive disorder

Table 2.3: Diagnosis variables utilised from the NACC UDS. Each variable's name and description is given. The descriptions are based on those provided by the National Alzheimer's Coordinating Center (2017).

clinician or whether it was a consensus diagnosis.

The variables from Form D1 were used to generate a number of labels (or classification targets) for every subject, each concerning an aspect of their diagnosis. The cognitive status of the subjects, and their dementia subtypes where applicable, were of most interest due to the objectives of the research, which were outlined in section 1.1. In particular, the relevant variables in both versions 1.2 and 2.0 of the UDS were identified but, as there was considerable overlap between them, a number of the variables were found to be redundant. Ultimately, only nine variables were required, namely those detailed in table 2.3.

Cognitive status was broken down into dementia, mild cognitive impairment (MCI) and normal cognition. The latter category comprised subjects without any cognitive impairment and those who were cognitively impaired but failed to meet the criteria for MCI, which are outlined by Petersen and Morris (2005). According to the National Alzheimer's Coordinating Center (2017), diagnostic criteria for dementia were not specified. However, it is suggested that most ADCs probably used those in the DSM-IV (American Psychiatric Association, 1994). In particular, 12,136 subjects were found to have been diagnosed with dementia, 6,815 with MCI and 13,622 with normal cognition.

The four main subtypes were focused on for those diagnosed with dementia.

Dementia Subtype	All Cases	Primary Cases	Pure Cases
Alzheimer's Disease	9501	8896	7757
Vascular Dementia	1070	239	134
Dementia with Lewy Bodies	1107	749	567
Frontotemporal Dementia	1658	1439	1203

Table 2.4: The number of subjects diagnosed with the four main dementia subtypes. The primary and 'pure' case frequencies only include subjects with certain diagnoses. Pure cases, in particular, were considered to be those in which a subject had a primary diagnosis of the subtype but no diagnoses of any of the other main subtypes.

McKhann et al. (1984), Román et al. (1993), McKeith et al. (2005) and Neary et al. (1998) provide the diagnostic criteria for each of the subtypes. Reflecting the way in which NACC grouped diagnoses, subjects with a diagnosis of stroke were combined with those diagnosed with vascular dementia. Likewise, Parkinson's disease and dementia with Lewy bodies diagnoses were grouped together, along with primary progressive aphasia and frontotemporal dementia diagnoses. It was ensured that all subjects considered with alternative diagnoses had dementia. Aho et al. (1980), Litvan et al. (2003), Mesulam (2001) and Mesulam (2003) provide the additional diagnostic criteria to distinguish between the diagnoses for the three pairings.

Table 2.4 provides a breakdown of those diagnosed with dementia, indicating the number of cases of each key subtype. Subjects with any diagnosis (primary, contributing or non-contributing) of the subtype were considered for all cases, whilst only those with a primary diagnosis of the subtype were included in primary cases. Moreover, subjects with a primary diagnosis of the subtype, but no diagnoses of any of the other main subtypes, were considered for pure cases. These pure cases were as pure as they reasonably could be without close inspection of rarer subtypes of dementia. Due to the way in which subjects are recruited and referred to ADCs, the frequencies are not representative of the prevalence of the subtypes in the general population (National Alzheimer's Coordinating Center, 2020).

2.5 Training and Test Sets

The data set, containing 32,573 subjects, was randomly partitioned into training and test sets. 22,801 subjects (70%) were included in the training set and the remaining

9,772 (30%) formed the test set. The proportion of missing values in the training and test sets was 0.65% and 0.68% respectively. The percentage of conditionally missing values, however, was 13.74% for the training set and 13.71% for the test set. All the missing values were imputed, and a number of those conditionally missing were updated in order to maintain relations between variables.

Tables 2.5, 2.6, 2.7 and 2.8 provide the basic characteristics for the two sets of subjects, broken down according to cognitive status and dementia subtype, post-imputation. N indicates the number of subjects or visits considered for each cognitive status and dementia subtype. The statistics concerning continuous variables are presented using medians and interquartile ranges (IQR), as the distribution for each of the relevant variables was found to be skewed. Those relating to binary or ordinal variables are given in terms of numbers and percentages. The last row of each table, providing the number of subjects without an MMSE score, corresponds to the conditionally missing values in the variable pertaining to the score (NACCMMSE). In particular, the score ranges from 0 to 30, with higher values indicating better cognition. The circumstances under which a score was not recorded were when the subject had a physical, cognitive/behaviour or other problem, or there was a verbal refusal. Tables 2.7 and 2.8 provide the characteristics for the pure cases of each dementia subtype; the last column (Other) of each table corresponds to all the remaining cases of dementia.

No considerable differences were found between the training and test sets when comparing them as a whole, and by cognitive status, but there were minor differences for the dementia subtypes. These were, however, mainly attributable to the small number of VD cases considered.

2.6 Variable Analysis

An investigation of each variable's predictive capacity was conducted to gain an understanding of how they correlate with their fellow variables, as well as targets distinguishing between dementia and no dementia (i.e. normal cognition or MCI). Of the 258 variables (in table 2.2), 255 were considered. Notably, NACCID (subject's

Characteristic	All Subjects ($N = 22801$)		Dementia ($N = 8500$)		MCI ($N = 4737$)		Normal Cognition ($N = 9564$)	
Age in years, median (IQR)	73	(14)	75	(15)	74	(13)	71	(13)
Sex, no. (%)								
Male	9836	(43.14)	4057	(47.73)	2363	(49.88)	3416	(35.72)
Female	12965	(56.86)	4443	(52.27)	2374	(50.12)	6148	(64.28)
Education in years, median (IQR)	16	(6)	14	(4)	16	(6)	16	(5)
Global CDR, no. (%)								
No impairment	8563	(37.56)	69	(0.81)	478	(10.09)	8016	(83.81)
Questionable impairment	8228	(36.09)	2541	(29.89)	4167	(87.97)	1520	(15.89)
Mild impairment	3798	(16.66)	3680	(43.29)	92	(1.94)	26	(0.27)
Moderate impairment	1443	(6.33)	1441	(16.95)	0	(0.00)	2	(0.02)
Severe impairment	769	(3.37)	769	(9.05)	0	(0.00)	0	(0.00)
MMSE score, median (IQR)	27	(6)	22	(9)	27	(3)	29	(2)
No MMSE score, no. (%)	797	(3.50)	470	(5.53)	77	(1.63)	250	(2.61)

Table 2.5: Training set characteristics by cognitive status.

Characteristic	All Subjects ($N = 9772$)		Dementia ($N = 3636$)		MCI ($N = 2078$)		Normal Cognition ($N = 4058$)	
Age in years, median (IQR)	73	(14)	75	(15)	73	(13)	72	(13)
Sex, no. (%)								
Male	4216	(43.14)	1728	(47.52)	1012	(48.70)	1476	(36.37)
Female	5556	(56.86)	1908	(52.48)	1066	(51.30)	2582	(63.63)
Education in years, median (IQR)	16	(6)	14	(4)	16	(6)	16	(5)
Global CDR, no. (%)								
No impairment	3615	(36.99)	23	(0.63)	229	(11.02)	3363	(82.87)
Questionable impairment	3555	(36.38)	1064	(29.26)	1807	(86.96)	684	(16.86)
Mild impairment	1661	(17.00)	1608	(44.22)	42	(2.02)	11	(0.27)
Moderate impairment	594	(6.08)	594	(16.34)	0	(0.00)	0	(0.00)
Severe impairment	347	(3.55)	347	(9.54)	0	(0.00)	0	(0.00)
MMSE score, median (IQR)	27	(6)	21	(9)	28	(3)	29	(2)
No MMSE score, no. (%)	326	(3.34)	186	(5.12)	41	(1.97)	99	(2.44)

Table 2.6: Test set characteristics by cognitive status.

Characteristic	Dementia ($N = 8500$)		AD ($N = 5414$)		VD ($N = 96$)		DLB ($N = 381$)		FTD ($N = 856$)		Other ($N = 1753$)	
Age in years, median (IQR)	75	(15)	76	(13)	78	(10.25)	73	(10)	64	(12)	74	(16)
Sex, no. (%)												
Male	4057	(47.73)	2354	(43.48)	46	(47.92)	285	(74.80)	476	(55.61)	896	(51.11)
Female	4443	(52.27)	3060	(56.52)	50	(52.08)	96	(25.20)	380	(44.39)	857	(48.89)
Education in years, median (IQR)	14	(4)	14	(4)	13	(4.25)	15	(6)	16	(6)	14	(4)
Global CDR, no. (%)												
No impairment	69	(0.81)	1	(0.02)	1	(1.04)	0	(0.00)	43	(5.02)	24	(1.37)
Questionable impairment	2541	(29.89)	1638	(30.25)	29	(30.21)	110	(28.87)	264	(30.84)	500	(28.52)
Mild impairment	3680	(43.29)	2477	(45.75)	39	(40.62)	174	(45.67)	292	(34.11)	698	(39.82)
Moderate impairment	1441	(16.95)	884	(16.33)	18	(18.75)	71	(18.64)	160	(18.69)	308	(17.57)
Severe impairment	769	(9.05)	414	(7.65)	9	(9.38)	26	(6.82)	97	(11.33)	223	(12.72)
MMSE score, median (IQR)	22	(9)	21	(8)	22	(9.50)	23	(7)	23	(10)	22	(9)
No MMSE score, no. (%)	470	(5.53)	233	(4.30)	2	(2.08)	19	(4.99)	75	(8.76)	141	(8.04)

Table 2.7: Training set characteristics by dementia subtype.

Characteristic	Dementia ($N = 3636$)	AD ($N = 2343$)	VD ($N = 38$)	DLB ($N = 186$)	FTD ($N = 347$)	Other ($N = 722$)
Age in years, median (IQR)	75 (15)	76 (14)	76.5 (13.00)	73 (10.75)	64 (11)	75 (17)
Sex, no. (%)						
Male	1728 (47.52)	959 (40.93)	19 (50.00)	147 (79.03)	223 (64.27)	380 (52.63)
Female	1908 (52.48)	1384 (59.07)	19 (50.00)	39 (20.97)	124 (35.73)	342 (47.37)
Education in years, median (IQR)	14 (4)	14 (4)	12 (2.75)	16 (6.00)	16 (4)	14 (5)
Global CDR, no. (%)						
No impairment	23 (0.63)	1 (0.04)	0 (0.00)	0 (0.00)	14 (4.03)	8 (1.11)
Questionable impairment	1064 (29.26)	676 (28.85)	19 (50.00)	58 (31.18)	124 (35.73)	187 (25.90)
Mild impairment	1608 (44.22)	1093 (46.65)	12 (31.58)	84 (45.16)	129 (37.18)	290 (40.17)
Moderate impairment	594 (16.34)	362 (15.45)	5 (13.16)	31 (16.67)	49 (14.12)	147 (20.36)
Severe impairment	347 (9.54)	211 (9.01)	2 (5.26)	13 (6.99)	31 (8.93)	90 (12.47)
MMSE score, median (IQR)	21 (9)	21 (8)	24 (6.00)	24 (7.00)	24 (10)	21 (11)
No MMSE score, no. (%)	186 (5.12)	86 (3.67)	3 (7.89)	6 (3.23)	32 (9.22)	59 (8.17)

Table 2.8: Test set characteristics by dementia subtype.

identification number), VISIT_DATE (visit date) and NACCVNUM (visit number) were excluded as they were originally maintained for administrative reasons and testing purposes.

Firstly, a naïve Bayes classifier was trained to predict dementia or no dementia using each variable one by one. As Hand and Yu (2001) explain, a naïve Bayes classifier, which is simple but effective, utilises Bayes' theorem whilst assuming conditional independence between variables given the target value (or class). A variety of naïve Bayes classifiers can be produced which are suited to different types of variables, thus this was taken into consideration. In fact, Gaussian, multinomial and Bernoulli naïve Bayes classifiers were produced for continuous/ordinal, categorical and binary variables respectively (see scikit-learn documentation (scikit-learn developers, 2020c) for more details). Naturally, data from the training set was used to train each classifier and data from the test set was used to determine its accuracy, but any missing or conditionally missing values were disregarded for the sake of simplicity.

Table 2.9 provides the top 10 variables predictive of dementia, namely the variables whose classifiers had the highest accuracies. The accuracy of each variable's classifier is given, which is simply the percentage of subjects correctly classified, along with abridged descriptions of the variables themselves. It is clear from the table that some of the variables are highly predictive, particularly those pertaining to the Clinical Dementia Rating (CDR). However, this does not seem to be the case for the majority of the 255 variables. In fact, it appears their classifiers simply predicted no dementia (the predominant class) for every subject, based on their accuracies. Crucially, the variables which are predictive of dementia focus mainly on cognitive impairment and the subject's ability to engage in activities of daily living, corresponding with the fundamental aspects clinicians consider when diagnosing dementia.

Subsequently, naïve Bayes classifiers and (simple) linear regression models, the latter of which employed the ordinary least squares method, were trained to predict the values of a variable from another. Every possible pairing (and permutation) of the 255 variables was considered, except for those where a variable was paired

with itself; and whether a classifier or regression model was produced depended on the type of the variable acting as the target. Notably, a naïve Bayes classifier was produced if the target variable was binary or categorical and a linear regression model was produced if it was continuous or ordinal. The type of the predictor variable also had to be taken into consideration. As a matter of fact, Gaussian, multinomial and Bernoulli naïve Bayes classifiers were generated, and one-hot encoding was performed for binary and categorical predictor variables prior to the generation of a linear regression model. Once again, data from the training and test sets, minus any missing or conditionally missing values, was used. However, the normalised root mean squared error (NRMSE), which can be defined using equation 2.1, was calculated for each linear regression model as opposed to accuracy.

$$\text{NRMSE} = \sqrt{\frac{mean((X^f_{true} - X^f_{pred})^2)}{var(X^f_{true})}} \tag{2.1}$$

To clarify, X^f_{true} and X^f_{pred} are the true and predicted values for a variable, respectively, whilst $mean(\cdot)$ and $var(\cdot)$ represent the empirical mean and variance. It should be noted that lower NRMSE values indicate better performance and a very small number of variable pairings were ultimately excluded from this analysis as a result of the training and/or test set being empty once any missingness had been eliminated.

Table 2.10 provides the top 10 predictive variable pairings, according to the NRMSE, where the target variable is either continuous or ordinal. An abridged description of every variable is given, along with the NRMSE for each variable pairing. The role of each variable (predictor or target) is not specified for any of the pairings as the NRMSE was equivalent (to two decimal places) for both permutations. Interestingly, half of the pairings in table 2.10 comprise CDR variables and all five of these pairings feature CDRSUM, which provides the sum of the scores for six categories (home and hobbies, community affairs, etc.) that essentially assess the subject's cognitive impairment and their ability to engage in activities of daily living. It could be inferred that the CDRSUM variable is somewhat correlated with the other CDR variables. For these top 10 pairings, age is also a recurring theme. In fact, the pairing with by far the lowest NRMSE includes variables pertaining to the

Variable	Description	Accuracy (%)
CDRSUM	CDR sum of boxes	92.25
MEMORY	CDR - Memory	90.29
JUDGMENT	CDR - Judgment and problem-solving	89.26
TAXES	Recent difficulty with taxes	88.65
COGJUDG	Impaired in judgment, planning or problem-solving	88.42
CDRGLOB	Global CDR	88.33
BILLS	Recent difficulty with bills	87.78
COMMUN	CDR - Community affairs	87.68
HOMEHOBB	CDR - Home and hobbies	87.65
ORIENT	CDR - Orientation	86.39

Table 2.9: Top 10 variables predictive of dementia.

Variables	Descriptions	NRMSE
NACCAGE BIRTH_#MOS	Subject's age at visit Months from subject's birth	0.03
CDRGLOB CDRSUM	Global CDR CDR sum of boxes	0.27
CDRSUM HOMEHOBB	CDR sum of boxes CDR - Home and hobbies	0.29
CDRSUM COMMUN	CDR sum of boxes CDR - Community affairs	0.29
JUDGMENT CDRSUM	CDR - Judgment and problem-solving CDR sum of boxes	0.32
DIGIB DIGIBLEN	Digit span backward trials correct Digit span backward length	0.33
NACCAGE DECAGE	Subject's age at visit Age cognitive decline began	0.33
BIRTH_#MOS DECAGE	Months from subject's birth Age cognitive decline began	0.33
BILLS TAXES	Recent difficulty with bills Recent difficulty with taxes	0.34
ORIENT CDRSUM	CDR - Orientation CDR sum of boxes	0.34

Table 2.10: Top 10 predictive variable pairings, according to the normalised root mean squared error (NRMSE), where the target variable is either continuous or ordinal.

subject's age, whilst two other pairings comprise variables concerning the subject's age and the age at which cognitive decline began. It is unlikely these ages would be drastically different, hence the apparent correlation between the variables.

Only a few examples of predictive variable pairings where the target variable is either binary or categorical are provided presently, as it seems the naïve Bayes classifiers for many of the pairings achieved a high accuracy by simply predicting the predominant target value. Three variables are included in the example pairings collectively, namely DECCLIN, COGMEM and MEMORY. DECCLIN indicates whether the clinician believed there was a meaningful decline in one or more of a variety of domains, such as memory, or there were motor/movement changes. COGMEM and MEMORY, on the other hand, indicate whether the subject's memory was meaningfully impaired by means of yes/no and a (CDR) score, respectively. The pairings, in the form of (predictor, target), are as follows: (DECCLIN, COGMEM), (COGMEM, DECCLIN) and (MEMORY, COGMEM). Notably, the classifier produced for the third pairing had an accuracy of 94%, whereas those produced for the first and second pairings had an accuracy of 96.15%. From the descriptions of these variables, it is clear they primarily concern the subject's memory; therefore, it follows that these pairs of variables seem to be correlated.

To summarise, each variable's predictive capacity was investigated by training various naïve Bayes classifiers and (simple) linear regression models. It was ascertained that variables, such as CDRSUM, which essentially provide information on the subject's cognitive impairment and their ability to engage in activities of daily living are highly predictive of dementia. It was also demonstrated that variables covering the same or similar topics are largely predictive of each other.

2.7 Summary

The Uniform Data Set was obtained from the National Alzheimer's Coordinating Center. It includes data pertaining to Alzheimer's Disease Center (ADC) visits at which a number of forms are completed. These provide demographic information for the subject and co-participant, insights into the subject's health, results of

standardised tests and evaluations for the subject, and an assessment of the subject's symptoms. Two types of missingness are present within the data set. Missing values occur due to data unexpectedly not being recorded; these values were imputed, where possible. Conditionally missing values arise as a result of information being irrelevant or unobtainable for a known reason; these values were handled rather than imputed.

Data cleansing was necessary, and the first step was to extract the data of interest. This resulted in a data set comprising 32,573 visits/subjects and 258 variables, two of which were not utilised during analysis. In fact, 260 variables were included in the data set subjected to analysis, as four randomly generated synthetic variables were added for testing purposes. The codes corresponding to missingness were subsequently examined and replaced, ensuring uniformity throughout the data set and enabling the two types of missing values to be easily identified. Crucially, any missing values which could not be sensibly imputed were marked as conditionally missing. The conditionally missing values which resulted due to an omitted form were also noted, so they were not drawn on during imputation. The proportion of missing values for the data set was just 0.66%, but the percentage of subjects with at least one missing value was 47.57%. The significant number of subjects with missing values motivated the use of imputation. Finally, relations between variables were identified and verified. They were separated into two groups, namely dependencies and relationships. The former group included relations in which a single variable can cause another to have a conditionally missing value if it takes on a specific value itself. The latter, however, encompassed the relationships in which one or more variables can dictate the value of another. Missing values were present for some of the variables which can determine others, so it was important to deduce what the associated values were in the determined variables and set out whether they should be updated post-imputation.

The cognitive status of every subject was deduced, and dementia subtypes associated with them were identified, in order to assign each of them labels (or classification targets). Cognitive status was broken down into normal cognition, mild cognitive impairment (MCI) and dementia. Those designated as having normal

cognition were free from cognitive impairment or failed to meet the criteria for MCI. The four main dementia subtypes, which are Alzheimer's disease, vascular dementia, dementia with Lewy bodies and frontotemporal dementia, were focused on for those with dementia. The prevalence of the subtypes in the data set does not reflect the true prevalence for the general population due to the way in which subjects are enrolled at ADCs.

In preparation for classification, the data set was split 70:30 into training and test sets. There were found to be no considerable differences between the training and test sets when the basic characteristics were compared for the subjects as a whole and based on cognitive status. Prior to imputation, these training and test sets were used to aid the investigation of each variable's predictive capacity. Ultimately, this investigation revealed that some of the variables are highly predictive of dementia and variables covering the same or similar topics are largely predictive of each other. Incidentally, the imputation of the missing values in the training set was closely coupled with the construction of a classifier; the next chapter discusses the imputation approach.

Chapter 3

Imputation and Learning with Missing Data using Random Forests

This chapter discusses the imputation approach developed, which simultaneously builds a classifier whilst handling conditionally missing values. It begins with a brief overview of related literature to put the work into context. This is followed by an explanation of the approach, for which each step is considered in turn. Finally, experimental work is recounted which informed the number of imputation iterations performed, and the number of trees used. It also ascertained the effects of additional missingness on the imputation and classification performance.

Throughout the chapter the focus is on the NACC data, for which results are presented in chapter 4, along with the clinical implications. Nevertheless, the method could be applied to alternative data sets if it was tailored appropriately. In fact, at the time of writing, it is being adapted for clinical data from the Sentinel Stroke National Audit Programme (King's College London, 2020) by another researcher. It could also prove particularly useful for survey data generally, as conditionally missing values are likely and there does not appear to be a standard approach that takes them into consideration.

3.1 Background

Literature concerning methods for handling missing data is discussed in this section. Firstly, a brief explanation of the mechanisms of missingness is provided, which is followed by an overview of imputation techniques. Single and multiple imputation methods are covered, and the imputation of derived variables is considered. In this context, derived variables are those which can be determined by known relations between variables. Finally, a selection of decision-tree-based approaches that deal with missing data, via imputation or otherwise, are discussed.

3.1.1 Mechanisms of Missingness

Determining the mechanism behind any missingness is considered important, specifically for ensuring it is dealt with appropriately; there are three widely accepted mechanisms which are outlined by Little and Rubin (2002). The first is missing completely at random (MCAR), for which it is assumed that the pattern of missingness is independent of the data. Missing values that are MCAR could have arisen as a result of accidental omission and can be handled with relative ease. The second mechanism is missing at random (MAR), where the missingness is considered to be dependent on observed values in the data set as opposed to those that are missing. The third mechanism is not missing at random (NMAR), or missing not at random (MNAR) as it is more commonly known. For this mechanism, the missingness is deemed to be dependent on the missing values themselves. Dealing with missing values which are MNAR is much more difficult.

In practice, it is rarely possible to confidently identify the mechanism. Croninger and Douglas (2005) discuss this, along with the fact that more than one mechanism could be at work for data sets with large numbers of variables. Due to the relatively low degree of missingness in the NACC data set, and the complexity of modelling mechanisms of missingness, no formal investigation was undertaken to determine the mechanisms of the missing values; and they were treated as if they were MAR.

3.1.2 Imputation

Imputation, which replaces missing values with suitable substitutions, is one way of handling missing data. It promotes the preservation of data and enables the application of standard methods of analysis. Little and Rubin (2002), Schafer and Graham (2002) and Enders (2010) provide detailed reviews of imputation approaches, along with alternative techniques for dealing with missing data. There are numerous methods of imputation and some examples are discussed presently.

3.1.2.1 Single Imputation Methods

Single imputation methods generate a single value for every missing value. They have been very popular, but their use is now generally discouraged as they fail to account for imputation uncertainty. Four different approaches are considered below to illustrate this type of imputation.

Mean Imputation In its simplest form, mean imputation replaces a missing value with the mean of the observed values for the variable. It can be adapted for categorical variables by substituting the mode for the mean. The technique is simple to implement but naive in its approach. It alters the distribution of the variable, although the mean is unchanged, as well as its correlation with other variables (Schafer and Graham, 2002; Twala, 2005).

Hot Deck Imputation Andridge and Little (2010) describe the multiple forms of hot deck imputation but, in short, every missing value is replaced with an observed value for the same variable from a similar observation. This method is a favourite of those working with survey data. It is not based on a parametric model and preserves the distributions of variables (Schafer and Graham, 2002; Twala, 2005). However, a similarity metric must be chosen; and it relies heavily on identifying well-matched observations, which could prove difficult if there are very few observations to start with (Andridge and Little, 2010).

Regression Imputation In order to impute the missing values of a variable using regression imputation, a regression model based on the other variables in the data

set is built. It is trained using the observations for which the variable to be imputed has observed values, and can be used to generate predictions for the values that are missing. Little and Rubin (2002) explain that the approach can be extended by adding a residual to each of the predicted values to account for uncertainty; this is stochastic regression imputation and it has the potential to preserve the correlations between variables (van Buuren, 2018).

Expectation Maximisation Algorithm The Expectation Maximisation (EM) algorithm, formalised by Dempster, Laird and Rubin (1977), can be employed to find maximum likelihood estimates, or more specifically parameter estimates which maximise a likelihood function, for parametric models using data with missing values. As the parameters of a model are optimised, the missing values are inferred (or imputed). Schafer (1997) outlines the complete procedure, which Little and Rubin (2002) highlight is conceptually simple, even though it can be difficult to implement. In addition, considerable missingness can adversely affect the speed at which the parameters converge (Little and Rubin, 2002).

3.1.2.2 Multiple Imputation Methods

Multiple imputation methods, as the name suggests, produce multiple values for each missing value. They have gained prominence more recently, and are highly recommended as they deal with the issue of imputation uncertainty. Nevertheless, they are more labour-intensive than single imputation approaches.

Multiple imputation is performed in three steps. Initially, several versions of the data set are generated which incorporate different imputed values. Each data set is subjected to the same analysis, and the results are then combined. Little and Rubin (2002) discuss how certain single imputation methods can be used to complete the first step, but multiple imputation can naturally be motivated from the Bayesian perspective (Schafer and Graham, 2002). Data augmentation (Tanner and Wong, 1987) is a well-known iterative Bayesian approach to multiple imputation which has been explored by Schafer (1997) for use with categorical and mixed data.

3.1.2.3 Imputation of Derived Variables

Some data sets, such as the NACC UDS, include variables in relations with one another (e.g. height, weight and body mass index). The variables which can be determined by these relations are generally referred to as derived variables when discussed in relation to imputation. Nonetheless, literature on the imputation of derived variables is relatively recent and limited.

Desai et al. (2016) provide a brief overview of the literature and state that the approaches can broadly be categorised as either active or passive. It appears the main distinction between active and passive methods is the former allow implausible values and the latter do not. In particular, a value is implausible if it introduces inconsistencies into the data set, with regards to known relations, whilst a value is plausible if it does not. There is no consensus as to which type of approach is best, as plausible values are desirable but any bias that may be induced in obtaining them needs to be considered. Desai et al. (2016) note that the type of derived variable also warrants consideration when choosing a technique.

Alternatively, van Buuren (2018) discusses the imputation of derived variables in terms of the types of these variables. The review places considerable importance on generating plausible values, whilst highlighting that specifically tailored approaches are needed to deal with some types of variables in order to achieve the desired result.

One method championed by van Buuren (2018) builds on fully conditional specification (FCS), namely substantive model compatible FCS (SMC-FCS) (Bartlett et al., 2015). FCS is a popular multiple imputation approach which generates a number of imputed data sets using a collection of univariate models. Each model pertains to one of the variables with missing values, and is conditional on all the other variables in the data set. The substantive model, relating an outcome to the complete set of variables, is fitted to every imputed data set; and the results are combined. Fundamentally, SMC-FCS ensures all the univariate (imputation) models are compatible with the substantive model (i.e. analysis undertaken). Bartlett et al. (2015) point out that, in practice, an imputation model is unlikely to be perfect, but suggest that if the aspects of the data which are of interest in the analysis are

preserved, then any bias introduced by the imputation model may be small.

3.1.3 Missing Data and Decision Trees

As explained in chapter 1, decision trees were chosen for classification for a number of reasons, such as their ability to handle continuous and categorical variables with relative ease, their interpretability, and their performance when employed as members of an ensemble. Decision trees, however, can also be used to perform imputation; and Stekhoven and Bühlmann (2012) and van Buuren (2018) discuss the benefits. The most discernible advantage of decision-tree-based approaches is they are able to deal with mixed data, which the vast majority of imputation methods are unequipped for.

Research into imputation with decision trees is gaining momentum, but there is already an array of literature discussing techniques for handling missing data with decision trees. Twala (2005, 2009) and Ding and Simonoff (2010) provide overviews of the various approaches, whilst van Buuren (2018) focuses on imputation. Tang and Ishwaran (2017), more specifically, review imputation methods using random forests, which were introduced in chapter 1. Essentially, a random forest is an ensemble of different decision trees, each of which have been generated using a process with an element of randomness. A selection of techniques are discussed in the remainder of this section. The first two are well-known imputation approaches, which utilise random forests and ordinarily employ the Random Forests algorithm (section 1.3.2), but the final one is an alternative method which is suitable for dealing with the conditionally missing values in the NACC data (section 2.3.2).

missForest Stekhoven and Bühlmann (2012) proposed missForest: an iterative imputation method which uses random forests. It begins by substituting initial guesses for the missing values; and determining the order in which the variables should be considered, based on their number of missing values (smallest to largest). The approach proceeds by constructing an ensemble of regression trees for each variable with missing values in the designated order, using only the observations for which the variable has observed values. Once each random forest is formed, it is used to make predictions for the missing values in the variable, which are subsequently

substituted for the initial guesses. After all the missing values have been imputed, the first variable is revisited; and the whole procedure is repeated using the newly imputed values, until a stopping criterion is met.

Stekhoven and Bühlmann (2012) claim that each ensemble naturally executes multiple imputation, but this is not recognised by van Buuren (2018). Regardless, Tang and Ishwaran (2017) recommend missForest when correlation between variables is high. The authors do point out, however, that the approach can be slow; this can be mitigated to some degree by using Extra-Trees (section 1.3.3) rather than the Random Forests algorithm. The computational efficiency of the Extra-Trees algorithm, as well as its accuracy, is why it was chosen for classification on the NACC data, as discussed in chapter 1.

Proximity Imputation The approach described by Breiman and Cutler (2004) and Cutler, Cutler and Stevens (2012) is the original imputation method for random forests. The first step is to roughly impute the missing values, which can be achieved by substituting the median (continuous data) or mode (categorical data) of the observed values on a variable-by-variable basis. A random forest is generated using the imputed data set; and an $N-\text{by}-N$ matrix is populated, where N is the number of observations. This proximity (or similarity) matrix captures how similar the observations are to one another, by providing the proportion of times each pair ended up in the same terminal node across the ensemble. The missing values are then imputed again using the proximity matrix: the proximity-weighted average and proximity-weighted mode are used for continuous and categorical variables respectively. A new ensemble is subsequently constructed, and the process is repeated. Breiman and Cutler (2004) note that four to six iterations are typically enough to give stable imputed values.

The technique actively uses the classification targets to inform the imputation, as each proximity matrix is populated using a random forest; this is generally recommended, but it means that there can be no targets missing (Josse et al., 2019; Stekhoven and Bühlmann, 2012). The imputation is also closely coupled with the construction of the random forest which is later used for classification. It can be

advantageous to integrate the two stages, as it is easier to ensure they are compatible with one another, but imputing test cases can become more difficult (Bartlett et al., 2015). Within the literature, there does not appear to be a clear explanation of how to impute test data when this approach is used. However, Breiman and Cutler (2004) point out that it is possible to identify the proximity for each training and test observation pairing.

Missingness Incorporated in Attributes A conceptually simple method which handles rather than imputes missing values is missingness incorporated in attributes (MIA) (Twala, Jones and Hand, 2008). As explained in chapter 1, a split S on a variable X^f, which partitions the data set X and the set of classification targets Y, is chosen to be associated with an internal splitting node during the construction of a decision tree. MIA can increase the number of splits generated for each variable considered for splitting to $2\kappa + 1$, where κ is the number produced when no missing values are present. Essentially, observations which have a missing value for the variable are collectively incorporated into either side of a split, enabling two splits to be formed ($S_{MIA_{1-2}}$); they are also split from the observations which have an observed value for the variable, to generate one additional split (S_{MIA_3}). When the Extra-Trees algorithm is employed, K randomly selected variables are considered, and a single random split is generated for each of them ($\kappa = 1$). By using MIA in conjunction with Extra-Trees, the number of splits produced for each variable can be increased from one to three.

In the pseudocode for Extra-Trees (algorithm 1), which was provided in chapter 1, S was defined as

$$S \triangleq \{(X_L, Y_L), (X_R, Y_R)\},$$

where X_L and X_R are the subsets of observations sent to the left and right child nodes respectively, and Y_L and Y_R are the corresponding subsets of classification targets. Alternatively, the three MIA splits $S_{MIA_{1-3}}$ can be defined as

$$S_{MIA_1} \triangleq \{(X_L, Y_L), (X_R \cup X_{mis}, Y_R \cup Y_{mis})\},$$
$$S_{MIA_2} \triangleq \{(X_L \cup X_{mis}, Y_L \cup Y_{mis}), (X_R, Y_R)\},$$

$$S_{MIA_3} \triangleq \{(X_{mis}, Y_{mis}), (X_{obs}, Y_{obs})\}.$$

In order to produce these three splits, X must be partitioned into X_{mis} and X_{obs}, and Y must be divided into Y_{mis} and Y_{obs}. X_{mis} includes the observations with a missing value for the variable X^f on which to split, whilst X_{obs} comprises those which have an observed value. $S_{MIA_{1-2}}$ use S as a basis but, crucially, X_L and X_R only include observations with an observed value for X^f ($X_L \cup X_R = X_{obs}$).

MIA is recommended by many, such as Josse et al. (2019) and Kapelner and Bleich (2015), as it can successfully deal with missing values without imputing them. Twala, Jones and Hand (2008) also highlight that it can be utilised in conjunction with any method of building decision trees.

3.2 Proximity Imputation with MIA

There is no optimal imputation strategy. The suitability of an approach is dependent on the characteristics of the data set itself, including the proportion of missing data, mechanisms of missingness and number of observations. It is also contingent on whether the data set includes variables of mixed type, like the NACC UDS, as most imputation strategies are unable to deal with mixed data. Decision-tree-based approaches, however, are capable of handling mixed data, as highlighted in section 3.1.3. The proximity imputation method was the natural choice for the NACC data and this research, particularly as it enables the imputation to be closely coupled with the construction of a random forest classifier. It was also possible to utilise missingness incorporated in attributes (MIA) in conjunction with the proximity imputation approach to handle the conditionally missing values. In fact, MIA was integrated into Extra-Trees: the algorithm chosen to construct random forests.

The approach developed, simply termed *proximity imputation with MIA*, begins by crudely imputing the missing values in the data set (or training set) to enable a random forest to be constructed. Extra-Trees and MIA are subsequently employed to build the ensemble of decision trees, using the imputed data set. By inspecting the paths of the observations through every tree, the similarity of each pair of

observations can be ascertained. These similarities (or proximities) are used to populate a proximity matrix, which is then utilised to impute the missing values for a second time. It was necessary to specifically tailor this step of the proximity imputation method to maintain the known relations between variables in the NACC data set (section 2.3.3), so far as possible. Nonetheless, all the variables are still used to inform the imputation. Once a newly imputed data set has been formed, another random forest is built and the process is repeated for a number of iterations. The remainder of this section provides a detailed explanation of the approach, which is discussed step-by-step in an attempt to aid understanding.

3.2.1 Initial Imputation

The first step of the approach eliminates any genuinely missing values in the data set, to leave only those which are conditionally missing; and does so by roughly imputing them. The missing values must be filled in (i.e. imputed) to permit a random forest to be built, as MIA is only employed to handle conditionally missing values. It was suggested in section 3.1.3 that missing values could be crudely imputed by substituting the median or mode of the observed values on a variable-by-variable basis. However, due to the presence of conditionally missing values, which are deemed observed for this step alone, the median could not be calculated for a number of the continuous variables in the NACC UDS. Consequently, a simple implementation of hot deck imputation is used which, in short, replaces each missing value with an observed value for the variable from an observation associated with the same class. It is able to take any conditionally missing values into consideration for each of the variables, and imputes values that are somewhat informed.

As indicated above, the classification targets (or classes) Y are required to initially impute the data set X. Each observation X_n has a corresponding class label Y_n (0 or 1), and the class labels can be used to ascertain whether observations could be deemed similar. The procedure considers every variable X^f with missing values in turn (algorithm 2 line 4). For each variable, it begins by identifying the missing values, along with the observed values from observations associated with

Algorithm 2 Initial imputation of missing values

1: **function initial_imputation**(X,Y)
2: $\tilde{X} \leftarrow X$
3: $\Lambda \leftarrow \{1, \ldots, |X|\}$
4: **for each** X^f with missing values **do**
5: $\Lambda_{mis} \leftarrow \{i \in \Lambda \,|\, \boldsymbol{\phi}(X_i^f)\}$ ▷ $\boldsymbol{\phi}(X_i^f) = true$ if X_i^f missing
6: $\Lambda_{obs} \leftarrow \Lambda \setminus \Lambda_{mis}$
7: $\Lambda_0 \leftarrow \{i \in \Lambda_{obs} \,|\, Y_i = 0\}$; $\Lambda_1 \leftarrow \Lambda_{obs} \setminus \Lambda_0$
8: $X^f_{obs\,|\,0} \leftarrow \{X_i^f \; \forall i \in \Lambda_0\}$; $X^f_{obs\,|\,1} \leftarrow \{X_i^f \; \forall i \in \Lambda_1\}$
9: **for** $i \in \Lambda_{mis}$ **do**
10: **if** $Y_i = 0$ **then**
11: $\tilde{X}_i^f \leftarrow$ random sample from $X^f_{obs\,|\,0}$
12: **else if** $Y_i = 1$ **then**
13: $\tilde{X}_i^f \leftarrow$ random sample from $X^f_{obs\,|\,1}$
14: **end if**
15: **end for**
16: **end for**
17: **return** $\tilde{X}$
18: **end function**

class 0 and class 1, denoted by $X^f_{obs\,|\,0}$ and $X^f_{obs\,|\,1}$ respectively (algorithm 2 lines 5–8). In algorithm 2, a predicate (Boolean-valued function) $\boldsymbol{\phi}(\cdot)$ is employed to assess whether a value X_n^f is missing. Every missing value is subsequently imputed with a random sample from either $X^f_{obs\,|\,0}$ or $X^f_{obs\,|\,1}$, according to the class of the observation with the missing value (algorithm 2 lines 9–15). There is no guarantee that known relations between variables are maintained, but any inconsistencies introduced are eliminated in the next stage of imputation. Once all the missing values in X have been imputed, the imputed data set $\tilde{X}$ can be used to build a random forest.

It was stated in chapter 2 that conditionally missing values could arise in the NACC data due to a form not having to be completed. It was also explained that these values had the potential to introduce bias into the imputation if used to inform it; thus, any conditionally missing values which occured for this reason were excluded from the sets of observed values, namely $X^f_{obs\,|\,0}$ and $X^f_{obs\,|\,1}$.

3.2.2 Extra-Trees with MIA

With the imputed data set, which is simply referred to as X, a random forest can be constructed. The random forest is fundamental to the approach as it enables the similarity of each pair of observations to be ascertained, all of which are subsequently used to inform the imputation of the missing values. As previously explained, the Extra-Trees algorithm was chosen to build random forests, due to its accuracy and computational efficiency; and it is employed in conjunction with MIA, which handles the conditionally missing values that remain in the data set.

Extra-Trees, which was discussed in detail in section 1.3.3, builds an ensemble of decision trees (or random forest) using random feature selection and random split selection. Every tree is built using all the observations in X (or a training set); and a split S on a variable X^f is chosen for each internal splitting node from K randomly generated splits, each of which corresponds to one of the inconstant variables randomly selected at the node. In particular, S is chosen so as to maximise the information gain $\mathcal{I}_S(X)$ resulting from the split.

MIA, also known as missingness incorporated in attributes, can be used together with any method of building decision trees (e.g. Extra-Trees) to deal with missing values, or conditionally missing values in this context, without imputing them. The approach ultimately generates more splits for each variable considered for splitting that has conditionally missing values, specifically $2\kappa + 1$ as opposed to κ; the three possible MIA splits $S_{MIA_{1-3}}$ were defined in section 3.1.3. For Extra-Trees, in particular, the number of splits is increased from one to three.

Modifications can be made to the pseudocode for Extra-Trees (algorithm 1), provided in chapter 1, to reflect the changes introduced by MIA. In fact, lines 14–15 can be amended as follows:

13: Randomly select K inconstant variables $\{X^{\varphi_1}, \dots, X^{\varphi_K}\}$ without replacement
14: Generate splits $\{S_1, \dots\}$ using **generate_splits**(X,X^{φ_i},Y) $\forall i \leftarrow 1, \dots, K$
15: Choose a split $S \triangleq \{(X_L, Y_L), (X_R, Y_R)\}$
such that $\mathcal{I}_S(X) \leftarrow \max_{i \leftarrow 1,\dots,|\{S_1,\dots\}|} \mathcal{I}_{S_i}(X)$ using equation 1.1

Line 13 has been included, regardless of the fact it is unchanged, as together these

three lines detail how a split is chosen. The amendments made to lines 14–15 are minor, and simply reflect that it is no longer known exactly how many splits will be generated. In addition, lines 21–35 can be replaced with those constituting algorithm 3, which outlines how splits are generated.

The new procedure for generating splits extends the original, as two of the three MIA splits ($S_{MIA_{1-2}}$) use a standard split as a basis. The first step is to identify any conditionally missing values for the variable X^f on which to split, as well as the observed values X^f_{obs} (algorithm 3 lines 2–5). In algorithm 3, a predicate $\boldsymbol{\psi}(\cdot)$ is employed to assess whether a value X^f_n is conditionally missing, similar to $\boldsymbol{\phi}(\cdot)$ in algorithm 2. X can subsequently be partitioned into X_{mis} and X_{obs}, and Y can correspondingly be divided into Y_{mis} and Y_{obs} (algorithm 3 lines 6–7). X_{mis} is the subset of observations which have a conditionally missing value for X^f, which could be empty, whilst X_{obs} is the subset of observations which have an observed value. The next step is to examine the set of observed values X^f_{obs}. If there is only one unique observed value, a single split separating X_{mis} from X_{obs}, as well as Y_{mis} from Y_{obs}, is generated (algorithm 3 lines 8–10). These are the only conditions under which a single split, specifically S_{MIA_3}, is produced for a variable with conditionally missing values, but exceptional circumstances that Twala, Jones and Hand (2008) seemingly overlooked. Alternatively, X_{obs} is partitioned into X_L and X_R, and Y_{obs} is divided into Y_L and Y_R, using a cut-point or subset depending on the type of X^f (algorithm 3 lines 11–21). It must then be ascertained whether one or three splits need to be generated, by essentially checking if X^f has conditionally missing values (algorithm 3 line 22). If not, a (standard) split is formed which separates X_L from X_R, as well as Y_L from Y_R (algorithm 3 line 23). Otherwise, the three MIA splits $S_{MIA_{1-3}}$ are produced (algorithm 3 lines 25–28). The split generated if there are no conditionally missing values forms the foundations of S_{MIA_1} and S_{MIA_2}. For S_{MIA_1}, X_{mis} and Y_{mis} are combined with X_R and Y_R respectively. For S_{MIA_2}, however, they are combined with X_L and Y_L.

In section 1.3.1, it was explained how a new (unseen) observation (or test case) is classified using a decision tree: it is simply passed through the tree, once the latter

Algorithm 3 Generating splits for a variable

function generate_splits(X,X^f,Y)
  $\Lambda \leftarrow \{1, \ldots, |X|\}$
  $\Lambda_{mis} \leftarrow \{i \in \Lambda \,|\, \boldsymbol{\psi}(X_i^f)\}$ ▷ $\boldsymbol{\psi}(X_i^f) = true$ if X_i^f conditionally missing
  $\Lambda_{obs} \leftarrow \Lambda \setminus \Lambda_{mis}$
  $X_{obs}^f \leftarrow \{X_i^f \; \forall i \in \Lambda_{obs}\}$
  $X_{mis} \leftarrow \{X_i \; \forall i \in \Lambda_{mis}\}$; $Y_{mis} \leftarrow \{Y_i \; \forall i \in \Lambda_{mis}\}$
  $X_{obs} \leftarrow \{X_i \; \forall i \in \Lambda_{obs}\}$; $Y_{obs} \leftarrow \{Y_i \; \forall i \in \Lambda_{obs}\}$
  if single unique value $\in X_{obs}^f$ **then**
    return $S_{MIA_3} \leftarrow \{(X_{mis}, Y_{mis}), (X_{obs}, Y_{obs})\}$
  end if
  if X^f continuous, ordinal or binary **then**
    Uniformly draw a cut-point f_{cp} in $(\min X_{obs}^f, \max X_{obs}^f)$
    $\Lambda_L \leftarrow \{i \in \Lambda_{obs} \,|\, X_i^f < f_{cp}\}$
  else if X^f categorical **then**
    Identify all possible values present in X_{obs}^f ($f^{(in)} \subseteq f$)
    Randomly draw $f_1 \subset f^{(in)}$ where $f_1 \neq \varnothing$
    Randomly draw $f_2 \subseteq f \setminus f^{(in)}$
    $\Lambda_L \leftarrow \{i \in \Lambda_{obs} \,|\, X_i^f \in f_1 \cup f_2\}$
  end if
  $X_L \leftarrow \{X_i \; \forall i \in \Lambda_L\}$; $X_R \leftarrow \{X_i \; \forall i \in \Lambda_{obs} \setminus \Lambda_L\}$
  $Y_L \leftarrow \{Y_i \; \forall i \in \Lambda_L\}$; $Y_R \leftarrow \{Y_i \; \forall i \in \Lambda_{obs} \setminus \Lambda_L\}$
  if $X_{mis} = \varnothing$ **then**
    return $S \leftarrow \{(X_L, Y_L), (X_R, Y_R)\}$
  end if
  $S_{MIA_1} \leftarrow \{(X_L, Y_L), (X_R \cup X_{mis}, Y_R \cup Y_{mis})\}$
  $S_{MIA_2} \leftarrow \{(X_L \cup X_{mis}, Y_L \cup Y_{mis}), (X_R, Y_R)\}$
  $S_{MIA_3} \leftarrow \{(X_{mis}, Y_{mis}), (X_{obs}, Y_{obs})\}$
  return $\{S_{MIA_1}, S_{MIA_2}, S_{MIA_3}\}$
end function

is fully formed. As a test case traverses through a tree built using Extra-Trees with MIA, an internal splitting node may be encountered which employs a standard split on a variable for which the test case has a conditionally missing value. In the event that this occurs, the test case is sent to the left or right child node at random.

3.2.3 Proximity Matrix

Using the random forest, it is possible to determine how similar the observations in X are to one another. In fact, the similarity of two observations is calculated based on their paths through the trees in the ensemble. The similarities (or proximities) are used to impute the missing values more rigorously by means of an $N-\text{by}-N$ proximity matrix, where N is the number of observations. Fundamentally, a missing value is imputed with the average, or alternatively the mode, of the observed values for the variable, weighted by proximity. The proximities could be calculated as the ensemble is constructed but, for the sake of simplicity, they are computed once it is fully formed.

Breiman and Cutler (2004) define the proximity of two observations for a single tree t as follows:

$$P_t(X_i, X_j) = \mathbb{I}(\rho_i = \rho_j)\,. \tag{3.2}$$

ρ_n is the path of an observation through t, and $\mathbb{I}(\cdot)$ is an indicator function which equals one when the paths of the two observations are identical. In order to utilise more of the information provided by t, the proximity is defined as

$$P_t(X_i, X_j) = \sum_{\forall \eta, \ell \in \rho_i \cap \rho_j} \begin{cases} \mathcal{I}_S(X) \dfrac{N_S}{N} & \text{if } \eta\,, \\ \left(1 + \displaystyle\sum_{c \in \{0,1\}} \frac{|Y_\ell^{(c)}|}{|Y_\ell|} \log_2 \frac{|Y_\ell^{(c)}|}{|Y_\ell|}\right) \dfrac{|Y_\ell|}{N} & \text{if } \ell\,. \end{cases} \tag{3.3}$$

More generally, it is the total information gain $\mathcal{I}_S(X)$ across the common nodes of the paths, weighted by the proportion of observations at each node $\frac{N_S}{N}$ when t was built. As a split is not associated with a terminal node ℓ, unlike an internal node η, the weighted inverse entropy of ℓ is also added if required. Y_ℓ and $Y_\ell^{(c)}$ are needed to

Algorithm 4 Determining the similarities (or proximities) of the observations

1: **function calculate_proximities**(X,N,T)
2: **for each** $t \in T$ **do**
3: $P_t \leftarrow 0_{N,N}$ ▷ $0_{N,N}$ is a zero matrix of size $N-$by$-N$
4: **for** $i \leftarrow 1, \ldots, N$ **do**
5: $\rho_i \leftarrow$ path of X_i through t
6: **for** $j \leftarrow 1, \ldots, N$ **do**
7: $\rho_j \leftarrow$ path of X_j through t
8: Calculate $P_t(X_i, X_j)$ using ρ_i, ρ_j and equation 3.3
9: **end for**
10: **end for**
11: $\bar{P}_t \leftarrow$ normalise P_t using equation 3.4
12: **end for**
13: $P \leftarrow$ average across $\bar{P}_t \; \forall t \in T$ using equation 3.5
14: **return** P
15: **end function**

calculate the latter, namely the classification targets for ℓ and the instances of each class ($c \in \{0, 1\}$). Zhu, Loy and Gong (2014) consider alternative ways of defining proximity for the purposes of spectral clustering.

As the proximity is a similarity, it must be normalised to ensure $P_t(X_i, X_i) = 1 \; \forall i = 1, \ldots, N$. This is achieved as follows:

$$\bar{P}_t(X_i, X_j) = \frac{P_t(X_i, X_j)}{\sqrt{P_t(X_i, X_i) P_t(X_j, X_j)}} \,. \tag{3.4}$$

Once (normalised) proximities have been obtained for every tree in the ensemble T, proximities for T itself can be ascertained. In fact, the proximity of two observations is calculated by simply averaging across the trees using

$$P(X_i, X_j) = \frac{1}{M} \sum_{t \in T} \bar{P}_t(X_i, X_j) \,, \tag{3.5}$$

where M is the size of T (i.e. the number of trees).

To clarify, proximities are computed for every tree in T by essentially comparing the paths of the observations through the tree (algorithm 4 lines 2–10). For each tree, the proximities form an $N-$by$-N$ matrix P_t, which is ultimately normalised to yield $\bar{P}_t$ (algorithm 4 line 11). The normalised proximities are averaged across

Algorithm 5 Imputation of missing values

1: **function impute**(X,P)
2: $\quad\tilde{X} \leftarrow X$
3: $\quad\Lambda \leftarrow \{1, \ldots, |X|\}$
4: $\quad$**for each** X^f with missing values **do**
5: $\quad\quad\Lambda_{mis} \leftarrow \{i \in \Lambda \mid \boldsymbol{\phi}(X_i^f)\}$ $\quad\triangleright\ \boldsymbol{\phi}(X_i^f) = true$ if X_i^f missing
6: $\quad\quad\Lambda_{obs} \leftarrow \Lambda \setminus \Lambda_{mis}$
7: $\quad\quad$**for** $i \in \Lambda_{mis}$ **do**
8: $\quad\quad\quad$**if** X^f continuous or ordinal **then**
9: $\quad\quad\quad\quad\tilde{X}_i^f \leftarrow round(\sum_{j \in \Lambda_{obs}} X_j^f P(X_i, X_j) \,/\, \sum_{j \in \Lambda_{obs}} P(X_i, X_j))$
10: $\quad\quad\quad$**else if** X^f categorical or binary **then**
11: $\quad\quad\quad\quad X_{obs}^f \leftarrow \{X_j^f \ \forall j \in \Lambda_{obs}\}$; $X_{obs} \leftarrow \{X_j \ \forall j \in \Lambda_{obs}\}$
12: $\quad\quad\quad\quad\tilde{X}_i^f \leftarrow mode(X_{obs}^f)$ weighted by $P(X_i, X_{obs})$
13: $\quad\quad\quad$**end if**
14: $\quad\quad$**end for**
15: $\quad$**end for**
16: $\quad$**return** $\tilde{X}$
17: **end function**

the trees to give the proximities for T, which constitute the proximity matrix P (algorithm 4 lines 13–14).

3.2.4 Imputation

With the proximities in P, the missing values in the data set X can be imputed in a more rigorous manner, ultimately enabling more meaningful conclusions to be drawn from the subsequent analysis. In practice, the crudely imputed values in $\tilde{X}$ are updated. However, the necessary steps are outlined in a way which suggests that the missing values have yet to be imputed, to simplify the explanation of the process. Conditionally missing values are also initially disregarded for the same reason, along with derived variables which are considered in section 3.2.5.

As for the initial imputation, each variable X^f with missing values is considered in turn (algorithm 5 line 4). The missing and observed values of X^f are identified; and every missing value is subsequently imputed according to the type of X^f, to ultimately form a newly imputed data set $\tilde{X}$ (algorithm 5 lines 5–14). If X^f is continuous or ordinal, the missing value is imputed with the proximity-weighted average of the

observed values, rounded to conform with the observed values (algorithm 5 lines 8–9). If X^f is categorical or binary, however, the proximity-weighted mode is substituted (algorithm 5 lines 10–12). In the event of multiple modes, one is chosen at random.

Without due consideration of the conditionally missing values or derived variables, there are only minor differences between the imputation stages of the newly developed approach and the original, specifically pertaining to how different types of variables are handled. Similarly to Geurts, Ernst and Wehenkel (2006), who proposed Extra-Trees (section 1.3.3), Breiman and Cutler (2004) fail to explicitly specify how ordinal and binary variables should be handled. As previously indicated, ordinal variables are treated as if they are continuous, and binary variables are regarded as if they are categorical; this is due to their fundamental nature. The proximity-weighted averages are also rounded, where appropriate, to ensure each imputed value is legitimate.

As explained in chapter 2, conditionally missing values can arise of their own accord in the NACC data set; the primary reason being a question is irrelevant in its own right. In section 2.3.2, the question associated with the MEALPREP variable, namely "In the past four weeks, did the subject have difficulty or need help with preparing a balanced meal?" (ADC Clinical Task Force and National Alzheimer's Coordinating Center, 2017c), was provided as an example. For this question in particular, a conditionally missing value would ensue if the subject had never performed the task. As a result, a conditionally missing value could be a legitimate fill value (i.e. imputed value) for certain variables in the NACC UDS.

The process described above had to be adapted to allow for conditionally missing fill values. In particular, the conditionally missing values have to be identified for each variable considered, along with those which are missing and observed. The relevant lines in algorithm 5 are 5–6, and they can be replaced with the following:

$\Lambda_{mis} \leftarrow \{i \in \Lambda \mid \boldsymbol{\phi}(X_i^f)\}$ ▷ $\boldsymbol{\phi}(X_i^f) = true$ if X_i^f missing

$\Lambda_{c-mis} \leftarrow \{i \in \Lambda \mid \boldsymbol{\psi}(X_i^f)\}$ ▷ $\boldsymbol{\psi}(X_i^f) = true$ if X_i^f conditionally missing

$\Lambda_{obs} \leftarrow \Lambda \setminus (\Lambda_{mis} \cup \Lambda_{c-mis})$

In section 3.2.1, it was stated that certain conditionally missing values, specifically

those which arose due to a form not having to be completed, were not used to inform the initial imputation; this was in fact the case for all imputation steps. Thus, these values were not considered for Λ_{c-mis}, or Λ_{mis} and Λ_{obs} for that matter.

In addition to identifying the conditionally missing values, it must be ascertained whether a conditionally missing value is a legitimate fill value for the variable in question, prior to calculating the proximity-weighted average or mode for a missing value. The relevant calculation must then allow for a conditionally missing fill value, if appropriate. For continuous or ordinal variables, line 9 of algorithm 5 can be substituted with those that follow.

---

if $\boldsymbol{\gamma}(X^f)$ and $\sum_{j \in \Lambda_{c-mis}} P(X_i, X_j) > \sum_{j \in \Lambda_{obs}} P(X_i, X_j)$ **then**
▷ $\boldsymbol{\gamma}(X^f) = true$ if conditionally missing fill value legitimate
▷ ties broken randomly for condition two
    $\tilde{X}_i^f \leftarrow$ conditionally missing
else
    $\tilde{X}_i^f \leftarrow round(\sum_{j \in \Lambda_{obs}} X_j^f P(X_i, X_j) \,/\, \sum_{j \in \Lambda_{obs}} P(X_i, X_j))$
end if

Initially, the legitimacy of a conditionally missing fill value is determined, using a predicate $\boldsymbol{\gamma}(\cdot)$. The proximity-weighted mode of the conditionally missing values and the observed values, collectively, for the variable X^f is also calculated. If a conditionally missing value is appropriate, and they are found to be most frequent value in X^f, then the missing value is imputed as conditionally missing. Otherwise, the proximity-weighted average of the observed values is simply substituted. For categorical or binary variables, lines 11–12 can be replaced with the following:

$X_{obs}^f \leftarrow \{X_j^f \; \forall j \in \Lambda_{obs}\}$; $X_{obs} \leftarrow \{X_j \; \forall j \in \Lambda_{obs}\}$
if $\boldsymbol{\gamma}(X^f)$ **then**
    $X_{c-mis}^f \leftarrow \{X_j^f \; \forall j \in \Lambda_{c-mis}\}$; $X_{c-mis} \leftarrow \{X_j \; \forall j \in \Lambda_{c-mis}\}$
    $\tilde{X}_i^f \leftarrow mode(X_{obs}^f \cup X_{c-mis}^f)$ weighted by $P(X_i, X_{obs} \cup X_{c-mis})$
else
    $\tilde{X}_i^f \leftarrow mode(X_{obs}^f)$ weighted by $P(X_i, X_{obs})$
end if

The proximity-weighted mode is calculated regardless of whether a conditionally missing fill value is appropriate or not, but the conditionally missing values are only

Algorithm 6 Proximity imputation with MIA procedure

Input:
X : data set
Y : classification targets
N : number of observations

Output:
$\tilde{X}$: imputed data set
T : random forest
P : proximity matrix

1: $\tilde{X} \leftarrow$ **initial_imputation**(X, Y)
2: **while** imputed values unstable **do**
3: $\quad T \leftarrow$ **build_ensemble**$(\tilde{X}, Y)$
4: $\quad P \leftarrow$ **calculate_proximities**$(\tilde{X}, N, T)$
5: $\quad \tilde{X} \leftarrow$ **impute**$(\tilde{X}, P)$
6: **end while**
7: $T \leftarrow$ **build_ensemble**$(\tilde{X}, Y)$
8: $P \leftarrow$ **calculate_proximities**$(\tilde{X}, N, T)$

considered if so. In order to calculate the mode of the observed and conditionally missing values, the set of conditionally missing values X^f_{c-mis} is combined with the set of observed values X^f_{obs}. As the frequencies of the values are weighted, the proximities are also required for the observations with conditionally missing and observed values for X^f, denoted by X_{c-mis} and X_{obs} respectively. Crucially, the observed values are considered individually not collectively, unlike for continuous or ordinal variables.

As explained in section 3.2, the imputed values are iteratively updated (algorithm 6 line 2). Cutler, Cutler and Stevens (2012) highlight that the intention is for the imputed values to stabilise, and Breiman and Cutler (2004) state four to six iterations are typically sufficient for the original proximity imputation approach. Experimental work was undertaken to inform the number of iterations for proximity imputation with MIA on the NACC data, and this is discussed in section 3.3.1.

During a single iteration, a random forest is constructed, proximities are calculated, and the missing values in the data set are imputed (algorithm 6 lines 3–5). The first two steps are also repeated for proximity imputation with MIA once the imputed values have stabilised and the imputation iterations have ceased (algorithm 6 lines 7–8). This ensures the random forest and proximity matrix used for analysis are based on the final imputed data set, as it cannot be guaranteed that the imputed values generated during the last and penultimate iterations are identical;

something which Breiman and Cutler (2004) do not acknowledge.

3.2.5 Derived Variables

As explained in chapter 2, a number of the variables included in the NACC data set are involved in either dependencies or relationships with each other. The variables which can be determined by these relations are referred to as derived variables within this chapter, in accordance with the imputation literature. The imputation step of the approach was specifically tailored to maintain the known relations between variables in the NACC UDS so far as possible, as stated in section 3.2; this was due to the general consensus in the literature being that it is ultimately desirable to do so. In particular, it was necessary to update certain derived variables with plausible values, specifically values which do not introduce inconsistencies into the data set with regards to the known relations, where the variables which can determine them had missing values. This idea was first discussed in section 2.3.3, which stated the values that were suitable to be updated were identified at the outset.

For the NACC data, a dependency arises when a certain value for one variable (parent) can trigger a conditionally missing value for another (child), whilst a relationship involves one or more variables which can determine the value of another. The HISPANIC and HISPOR variables form a dependency. The HISPANIC variable indicates whether the subject is of Hispanic/Latino ethnicity, and the HISPOR variable provides their origins if so. As a result, the former is the parent variable and the latter is the child. In the event that the subject is not of Hispanic/Latino ethnicity, the value of the HISPANIC variable is 0 and the HISPOR variable is assigned a conditionally missing value. The NACCMOM, NACCDAD and NACCFAM variables are involved in a relationship. They specify whether the subject's mother or father, or any of their first-degree family members, have or had cognitive impairment respectively. As a parent is a first-degree family member, the value of the NACCFAM variable is 1, indicating cognitive impairment was reported, if the value of at least one of the NACCMOM and NACCDAD variables is 1.

Tables 3.1 and 3.2 include the six dependencies and 20 relationships handled

during imputation. In the latter, the relationships are detailed under the assumption that the variables involved are complete (i.e. do not have missing values). Each of the variables which are part of a validated dependency or relationship in the NACC UDS are also highlighted in appendix A, along with exactly how they interact with one another. Not all of these interactions (or relations) needed to be considered during imputation for several reasons, including one or more of the variables involved could not be sensibly imputed; the derived variable's values should not be updated; and missing values were absent from the variable(s) which can determine another, not only in the data set but also in the documentation provided by NACC.

Only two of the relationships in table 3.2, specifically those including NACCTBI and NACCBMI, invariably determine the value of their derived variables. For the derived variables of the remaining relationships and the dependencies in table 3.1, imputation can be required to settle on a value. In fact, 15 of the relationships specify a range of values for the derived variable if the determining variable has a conditionally missing (CM) value, one of which needs to be chosen via imputation.

A number of variables feature in a dependency and a relationship; two variables which do so, namely DIGIFLEN and DIGIBLEN, are important. Both variables pertain to the digit span tests, for which subjects are asked to repeat number sequences of increasing length in order (forward) or in reverse order (backward) (ADC Clinical Task Force and National Alzheimer's Coordinating Center, 2014a). The DIGIFLEN and DIGIBLEN variables provide the length of the longest sequence correctly repeated forwards and backwards respectively. Crucially, these variables act as the derived variable in their respective dependencies, and can determine the value of another in their relationships. As a result, it is vital that the derived variables of the dependencies are updated prior to those of the relationships, to ensure any updates required for the DIGIFLEN and DIGIBLEN variables can be appropriately dealt with for the variables they can determine.

The imputation is staggered to maintain relations in the data set. In fact, the derived variables of the dependencies and relationships are updated immediately after the missing values of the variables which can determine them have been imputed,

Variables	Descriptions	Dependency Trigger
HISPANIC	Hispanic/Latino ethnicity	0 (no)
HISPOR	Hispanic origins	
DECCLIN	Clinician believes there is a meaningful decline in memory, non-memory cognitive abilities, behaviour, ability to manage his/her affairs, or there are motor/movement changes	0 (no)
DECAGE	Age cognitive decline began, based on clinician's assessment	
NACCCOGF	Predominant symptom first recognised as a decline in cognition	CM
COGMODE	Mode of onset of cognitive symptoms	
DIGIF	Digit span forward trials correct	CM
DIGIFLEN	Digit span forward length	
DIGIB	Digit span backward trials correct	CM
DIGIBLEN	Digit span backward length	
MEMUNITS	Logical Memory IIA - Delayed - Total number of story units recalled	CM
MEMTIME	Logical Memory IIA - Delayed - Time elapsed since Logical Memory IA (Immediate)	

Table 3.1: Six NACC UDS dependencies handled during imputation. For each dependency, the parent (top) and child (bottom) variables are provided, along with their descriptions and the dependency trigger. The dependency trigger is the value of the parent which causes the child to be conditionally missing; it is possible for this value to also be conditionally missing, denoted by CM.

Variables	Descriptions	Relationship
NACCMOM	Mother with cognitive impairment	NACCFAM = 1 (yes) if
NACCDAD	Father with cognitive impairment	NACCMOM and/or NACCDAD =
NACCFAM*	First-degree family member with cognitive impairment	1 (yes)
TRAUMBRF	Brain trauma - brief unconsciousness	NACCTBI = 1 (yes) if
TRAUMEXT	Brain trauma - extended unconsciousness	TRAUMBRF, TRAUMEXT
TRAUMCHR	Brain trauma - chronic deficit	and/or TRAUMCHR = 1 or 2
NACCTBI*	History of traumatic brain injury (TBI)	(recent or remote) else 0 (no)
HEIGHT	Subject's height (inches)	
WEIGHT	Subject's weight (lbs)	NACCBMI = $\frac{\text{WEIGHT}\times 703}{\text{HEIGHT}^2}$
NACCBMI*	Body mass index (BMI)	
NACCCOGF	Predominant symptom first recognised as a decline in cognition	COURSE = CM if NACCCOGF,
NACCBEHF	Predominant symptom first recognised as a decline in behaviour	NACCBEHF and NACCMOTF =
NACCMOTF	Predominant symptom first recognised as a decline in motor function	CM
COURSE*	Overall course of decline of cognitive/behavioural/motor syndrome	
NACCCOGF	Predominant symptom first recognised as a decline in cognition	FRSTCHG = CM if NACCCOGF,
NACCBEHF	Predominant symptom first recognised as a decline in behaviour	NACCBEHF and NACCMOTF =
NACCMOTF	Predominant symptom first recognised as a decline in motor function	CM
FRSTCHG*	Predominant domain first recognised as changed	
MMSEORDA	Mini-Mental State Examination - Orientation subscale score - Time	MMSEORDA_PROB = 95-98 if
MMSEORDA_PROB*	Reason orientation subscale score (time) not recorded	MMSEORDA = CM else CM
MMSEORLO	Mini-Mental State Examination - Orientation subscale score - Place	MMSEORLO_PROB = 95-98 if
MMSEORLO_PROB*	Reason orientation subscale score (place) not recorded	MMSEORLO = CM else CM

NACCMMSE NACCMMSE_PROB*	Total Mini-Mental State Examination (MMSE) score (using D-L-R-O-W) Reason MMSE not completed	NACCMMSE_PROB = 95-98 if NACCMMSE = CM else CM
LOGIMEM LOGIMEM_PROB*	Logical Memory IA - Immediate - Total number of story units recalled Reason Logical Memory IA (Immediate) not completed	LOGIMEM_PROB = 95-98 if LOGIMEM = CM else CM
DIGIF DIGIF_PROB*	Digit span forward trials correct Reason digit span forward trials not completed	DIGIF_PROB = 95-98 if DIGIF = CM else CM
DIGIFLEN DIGIFLEN_PROB*	Digit span forward length Reason digit span forward length not recorded	DIGIFLEN_PROB = 95-98 if DIGIFLEN = CM else CM
DIGIB DIGIB_PROB*	Digit span backward trials correct Reason digit span backward trials not completed	DIGIB_PROB = 95-98 if DIGIB = CM else CM
DIGIBLEN DIGIBLEN_PROB*	Digit span backward length Reason digit span backward length not recorded	DIGIBLEN_PROB = 95-98 if DIGIBLEN = CM else CM
ANIMALS ANIMALS_PROB*	Animals - Total number of animals named in 60 seconds Reason animal naming not completed	ANIMALS_PROB = 95-98 if ANIMALS = CM else CM
VEG VEG_PROB*	Vegetables - Total number of vegetables named in 60 seconds Reason vegetable naming not completed	VEG_PROB = 95-98 if VEG = CM else CM
TRAILA TRAILA_PROB*	Trail Making Test Part A - Total number of seconds to complete Reason Trail Making Test Part A not completed	TRAILA_PROB = 995-998 if TRAILA = CM else CM
TRAILB TRAILB_PROB*	Trail Making Test Part B - Total number of seconds to complete Reason Trail Making Test Part B not completed	TRAILB_PROB = 995-998 if TRAILB = CM else CM
WAIS	Wechsler Adult Intelligence Scale (Revised) (WAIS-R) Digit Symbol	WAIS_PROB = 95-98 if

WAIS_PROB*	Reason WAIS-R Digit Symbol not completed	WAIS = CM else CM
MEMUNITS	Logical Memory IIA - Delayed - Total number of story units recalled	MEMUNITS_PROB = 95-98 if
MEMUNITS_PROB*	Reason Logical Memory IIA (Delayed) not completed	MEMUNITS = CM else CM
BOSTON	Boston Naming Test (30) - Total score	BOSTON_PROB = 95-98 if
BOSTON_PROB*	Reason Boston Naming Test not completed	BOSTON = CM else CM

Table 3.2: 20 NACC UDS relationships handled during imputation. For each relationship, the relevant variables and their descriptions are provided, along with an explanation of the relationship between them. The derived variables, namely those which can be determined, are marked with an *; and conditionally missing values are denoted by CM.

and the missing values of the remaining variables are imputed subsequently. The derived variables are included for the last step as a number of them can have missing values of their own, although some will have already been dealt with by the updates.

Algorithm 6 outlines the proximity imputation with MIA procedure, and line 5 pertains to the imputation. This particular line can be replaced with those that follow, which describe the required modifications.

$\tilde{X} \leftarrow$ impute the variables which can determine others using **impute**$(\tilde{X}, P)$
$\tilde{X} \leftarrow$ update the derived variables for the dependencies
    using **impute**$(\tilde{X}, P)$ where not conditionally missing
$\tilde{X} \leftarrow$ update the derived variables for the relationships
    using **impute**$(\tilde{X}, P)$ where value not predetermined
$\tilde{X} \leftarrow$ impute all the remaining variables using **impute**$(\tilde{X}, P)$

As previously explained, imputation can be required to settle on a value for the derived variables of the dependencies and certain relationships. For dependencies it is needed when the value of the parent variable is not the dependency trigger, namely the value which causes the child to be conditionally missing. For the majority of the relationships, it is necessary when the value is simply not predetermined based on the values of the other variables involved in the relationship.

The imputation carried out at each step may act on different variables, and also different types of values where updates are concerned (i.e. not just missing values), but the fundamental process is the same. Consequently, the pseudocode outlining how missing values are imputed, presented in section 3.2.4 and pieced together in algorithm 7 (lines 94–120), is not altered to reflect this. In short, the relevant variables are imputed rather than every one with missing values; and specific values are replaced, not simply missing values, when the derived variables are updated.

In addition to the conditionally missing values which arose due to a form not having to be completed, certain values were disregarded during imputation for three of the four steps. In particular, the values to be imputed in the fourth step were dismissed for the two update steps, and the updated values were ignored for the last (i.e. fourth) step. No values were disregarded for the first step, as the DIGIFLEN and DIGIBLEN variables did not have any missing values of their own.

Each aspect of the approach developed has now been discussed. As a result, the algorithms and pseudocode snippets presented thus far, both in this chapter and chapter 1, can be assembled to form the pseudocode for proximity imputation with MIA; this is provided in algorithm 7. How the approach is applied to test cases is explained in section 3.2.6.

3.2.6 Imputation of Test Cases

Test cases could also have missing values. As stated in section 3.1.3, no clear explanation of how to impute test cases using the proximity imputation approach was found in the literature, but Breiman and Cutler (2004) do highlight it is possible to determine the similarity (or proximity) of training and test observations. In fact, proximity imputation with MIA can be used to impute test cases with a few alterations. Crucially, the imputed values are generated based on the imputed training cases alone, as each test case would be considered independently in practice.

Test cases do not typically have classification targets, so none are required. As a result, the initial imputation step simply substitutes a random value, which is associated with the same variable as the missing value, from one of the imputed training observations; the difference is any training observation is considered regardless of their class. In order to impute the missing values more rigorously, the random forest constructed using the final imputed training set is required, along with the proximities pertaining to the pairs of training and test cases which are obtained using the preconstructed forest. The proximity of each pair of test observations is also calculated to enable the latter to be normalised, and as they are likely to be of interest during analysis. The four stages of imputation are subsequently completed, for which the values of the imputed training observations are used to inform the imputed values, bar the conditionally missing values which arose due to a form not having to be completed. These imputed values are then updated over the course of a number of iterations, and the specified proximities are calculated one final time.

Algorithm 7 Pseudocode for proximity imputation with MIA

Input:
X : data set
Y : classification targets
N : number of observations

Output:
$\tilde{X}$: imputed data set
T : random forest
P : proximity matrix

$\tilde{X} \leftarrow$ **initial_imputation**(X, Y)
while imputed values unstable **do**
    $T \leftarrow$ **build_ensemble**$(\tilde{X}, Y)$
    $P \leftarrow$ **calculate_proximities**$(\tilde{X}, N, T)$
    $\tilde{X} \leftarrow$ impute the variables which can determine others using **impute**$(\tilde{X}, P)$
    $\tilde{X} \leftarrow$ update the derived variables for the dependencies
        using **impute**$(\tilde{X}, P)$ where not conditionally missing
    $\tilde{X} \leftarrow$ update the derived variables for the relationships
        using **impute**$(\tilde{X}, P)$ where value not predetermined
    $\tilde{X} \leftarrow$ impute all the remaining variables using **impute**$(\tilde{X}, P)$
end while
$T \leftarrow$ **build_ensemble**$(\tilde{X}, Y)$
$P \leftarrow$ **calculate_proximities**$(\tilde{X}, N, T)$

function initial_imputation(X,Y)
    $\tilde{X} \leftarrow X$
    $\Lambda \leftarrow \{1, \ldots, |X|\}$
    for each X^f with missing values **do**
        $\Lambda_{mis} \leftarrow \{i \in \Lambda \,|\, \boldsymbol{\phi}(X_i^f)\}$ ▷ $\boldsymbol{\phi}(X_i^f) = true$ if X_i^f missing
        $\Lambda_{obs} \leftarrow \Lambda \setminus \Lambda_{mis}$
        $\Lambda_0 \leftarrow \{i \in \Lambda_{obs} \,|\, Y_i = 0\}$; $\Lambda_1 \leftarrow \Lambda_{obs} \setminus \Lambda_0$
        $X^f_{obs\,|\,0} \leftarrow \{X_i^f \; \forall i \in \Lambda_0\}$; $X^f_{obs\,|\,1} \leftarrow \{X_i^f \; \forall i \in \Lambda_1\}$
        for $i \in \Lambda_{mis}$ **do**
            if $Y_i = 0$ **then**
                $\tilde{X}_i^f \leftarrow$ random sample from $X^f_{obs\,|\,0}$
            else if $Y_i = 1$ **then**
                $\tilde{X}_i^f \leftarrow$ random sample from $X^f_{obs\,|\,1}$
            end if
        end for
    end for
    return $\tilde{X}$
end function

30: **function build_ensemble**(X,Y)
31: **for** $i \leftarrow 1, \ldots, M$ **do**
32: $t_i \leftarrow$ **build_tree**(X,Y)
33: **end for**
34: **return** $T \leftarrow \{t_1, \ldots, t_M\}$
35: **end function**

36: **function build_tree**(X,Y)
37: **if** $X^i \; \forall i \leftarrow 1, \ldots, F$ constant or Y constant or $|X| < n_{min}$ **then**
38: $Y^{(0)} \leftarrow \{Y_n \in Y \,|\, Y_n = 0\}$
39: $Y^{(1)} \leftarrow \{Y_n \in Y \,|\, Y_n = 1\}$
40: **return** $\ell \leftarrow \{|Y^{(0)}|, |Y^{(1)}|\}$
41: **end if**
42: Randomly select K inconstant variables $\{X^{\varphi_1}, \ldots, X^{\varphi_K}\}$ without replacement
43: Generate splits $\{S_1, \ldots\}$ using **generate_splits**(X,X^{φ_i},Y) $\forall i \leftarrow 1, \ldots, K$
44: Choose a split $S \triangleq \{(X_L, Y_L), (X_R, Y_R)\}$
such that $\mathcal{I}_S(X) \leftarrow \max_{i \leftarrow 1,\ldots,|\{S_1,\ldots\}|} \mathcal{I}_{S_i}(X)$ using equation 1.1
45: $t_L \leftarrow$ **build_tree**(X_L,Y_L)
46: $t_R \leftarrow$ **build_tree**(X_R,Y_R)
47: Create η for S and attach t_L and t_R to form t
48: **return** t
49: **end function**

50: **function generate_splits**(X,X^f,Y)
51: $\Lambda \leftarrow \{1, \ldots, |X|\}$
52: $\Lambda_{mis} \leftarrow \{i \in \Lambda \,|\, \boldsymbol{\psi}(X_i^f)\}$ ▷ $\boldsymbol{\psi}(X_i^f) = true$ if X_i^f conditionally missing
53: $\Lambda_{obs} \leftarrow \Lambda \setminus \Lambda_{mis}$
54: $X_{obs}^f \leftarrow \{X_i^f \; \forall i \in \Lambda_{obs}\}$
55: $X_{mis} \leftarrow \{X_i \; \forall i \in \Lambda_{mis}\}$; $Y_{mis} \leftarrow \{Y_i \; \forall i \in \Lambda_{mis}\}$
56: $X_{obs} \leftarrow \{X_i \; \forall i \in \Lambda_{obs}\}$; $Y_{obs} \leftarrow \{Y_i \; \forall i \in \Lambda_{obs}\}$
57: **if** single unique value $\in X_{obs}^f$ **then**
58: **return** $S_{MIA_3} \leftarrow \{(X_{mis}, Y_{mis}), (X_{obs}, Y_{obs})\}$
59: **end if**
60: **if** X^f continuous, ordinal or binary **then**
61: Uniformly draw a cut-point f_{cp} in $(\min X_{obs}^f, \max X_{obs}^f)$
62: $\Lambda_L \leftarrow \{i \in \Lambda_{obs} \,|\, X_i^f < f_{cp}\}$

63: **else if** X^f categorical **then**
64: Identify all possible values present in X^f_{obs} ($f^{(in)} \subseteq f$)
65: Randomly draw $f_1 \subset f^{(in)}$ where $f_1 \neq \varnothing$
66: Randomly draw $f_2 \subseteq f \setminus f^{(in)}$
67: $\Lambda_L \leftarrow \{i \in \Lambda_{obs} \,|\, X^f_i \in f_1 \cup f_2\}$
68: **end if**
69: $X_L \leftarrow \{X_i \; \forall i \in \Lambda_L\}$; $X_R \leftarrow \{X_i \; \forall i \in \Lambda_{obs} \setminus \Lambda_L\}$
70: $Y_L \leftarrow \{Y_i \; \forall i \in \Lambda_L\}$; $Y_R \leftarrow \{Y_i \; \forall i \in \Lambda_{obs} \setminus \Lambda_L\}$
71: **if** $X_{mis} = \varnothing$ **then**
72: **return** $S \leftarrow \{(X_L, Y_L), (X_R, Y_R)\}$
73: **end if**
74: $S_{MIA_1} \leftarrow \{(X_L, Y_L), (X_R \cup X_{mis}, Y_R \cup Y_{mis})\}$
75: $S_{MIA_2} \leftarrow \{(X_L \cup X_{mis}, Y_L \cup Y_{mis}), (X_R, Y_R)\}$
76: $S_{MIA_3} \leftarrow \{(X_{mis}, Y_{mis}), (X_{obs}, Y_{obs})\}$
77: **return** $\{S_{MIA_1}, S_{MIA_2}, S_{MIA_3}\}$
78: **end function**

79: **function calculate_proximities**(X,N,T)
80: **for each** $t \in T$ **do**
81: $P_t \leftarrow 0_{N,N}$ ▷ $0_{N,N}$ is a zero matrix of size $N-$by$-N$
82: **for** $i \leftarrow 1, \ldots, N$ **do**
83: $\rho_i \leftarrow$ path of X_i through t
84: **for** $j \leftarrow 1, \ldots, N$ **do**
85: $\rho_j \leftarrow$ path of X_j through t
86: Calculate $P_t(X_i, X_j)$ using ρ_i, ρ_j and equation 3.3
87: **end for**
88: **end for**
89: $\bar{P}_t \leftarrow$ normalise P_t using equation 3.4
90: **end for**
91: $P \leftarrow$ average across $\bar{P}_t \; \forall t \in T$ using equation 3.5
92: **return** P
93: **end function**

94: **function impute**(X,P)
95: $\tilde{X} \leftarrow X$
96: $\Lambda \leftarrow \{1, \ldots, |X|\}$
97: **for each** X^f with missing values **do**
98: $\Lambda_{mis} \leftarrow \{i \in \Lambda \,|\, \boldsymbol{\phi}(X^f_i)\}$ ▷ $\boldsymbol{\phi}(X^f_i) = true$ if X^f_i missing
99: $\Lambda_{c-mis} \leftarrow \{i \in \Lambda \,|\, \boldsymbol{\psi}(X^f_i)\}$ ▷ $\boldsymbol{\psi}(X^f_i) = true$ if X^f_i conditionally missing
100: $\Lambda_{obs} \leftarrow \Lambda \setminus (\Lambda_{mis} \cup \Lambda_{c-mis})$

101: **for** $i \in \Lambda_{mis}$ **do**
102: **if** X^f continuous or ordinal **then**
103: **if** $\boldsymbol{\gamma}(X^f)$ and $\sum_{j \in \Lambda_{c-mis}} P(X_i, X_j) > \sum_{j \in \Lambda_{obs}} P(X_i, X_j)$ **then**
▷ $\boldsymbol{\gamma}(X^f) = true$ if conditionally missing fill value legitimate
▷ ties broken randomly for condition two
104: $\tilde{X}_i^f \leftarrow$ conditionally missing
105: **else**
106: $\tilde{X}_i^f \leftarrow round(\sum_{j \in \Lambda_{obs}} X_j^f P(X_i, X_j) \,/\, \sum_{j \in \Lambda_{obs}} P(X_i, X_j))$
107: **end if**
108: **else if** X^f categorical or binary **then**
109: $X_{obs}^f \leftarrow \{X_j^f \; \forall j \in \Lambda_{obs}\}$; $X_{obs} \leftarrow \{X_j \; \forall j \in \Lambda_{obs}\}$
110: **if** $\boldsymbol{\gamma}(X^f)$ **then**
111: $X_{c-mis}^f \leftarrow \{X_j^f \; \forall j \in \Lambda_{c-mis}\}$; $X_{c-mis} \leftarrow \{X_j \; \forall j \in \Lambda_{c-mis}\}$
112: $\tilde{X}_i^f \leftarrow mode(X_{obs}^f \cup X_{c-mis}^f)$ weighted by $P(X_i, X_{obs} \cup X_{c-mis})$
113: **else**
114: $\tilde{X}_i^f \leftarrow mode(X_{obs}^f)$ weighted by $P(X_i, X_{obs})$
115: **end if**
116: **end if**
117: **end for**
118: **end for**
119: **return** $\tilde{X}$
120: **end function**

3.3 Experiments

A number of experiments were conducted using proximity imputation with MIA on the NACC data, in order to determine an appropriate number of imputation iterations and trees. Moreover, the effects of additional missingness were ascertained with regards to the imputation and classification performance, to assess the capability of the approach developed.

Proximity imputation with MIA was applied to the training and test sets detailed in section 2.5; classification targets indicating whether each subject had received a diagnosis of dementia at their initial visit were utilised in conjunction with the training set. 1,000 missing values were also introduced into the training set before any experimental work was carried out, to enable the imputation performance

to be assessed. Values were replaced at random but a number of restrictions were adhered to. In particular, missing values already present, and conditionally missing values which arose due to a form not having to be completed, were not replaced. When a non-essential form is skipped, every variable related to the form is assigned a conditionally missing value; thus, to replace any less than all of these values would be inappropriate. Furthermore, the additional missing values were restricted to 111 of the 260 variables. These 111 variables were not involved in the dependencies and relationships handled during imputation, and could have missing values according to the National Alzheimer's Coordinating Center (2017).

As stated in section 3.2.5, not all of the validated dependencies and relationships in the NACC UDS needed to be considered during imputation. 19 of the variables (categorical or binary) for which missing values were added were involved in relations that were not handled. By incorporating missing values into these variables, it was confirmed that inconsistencies will almost certainly be introduced into the data if the relation with which a variable is associated is not taken into consideration, when appropriate; this highlights the importance of considering the various relations in the data set. A number of implausible values were imputed for the training set as a result, but they ultimately accounted for just 2.6% of the 1,000 values.

The remainder of this section discusses the experiments carried out, along with their results. The work undertaken to determine a suitable number of imputation iterations is explained to begin with, followed by that which verified 100 trees were sufficient. Finally, the work investigating the effects of additional missingness on the imputation and classification performance is discussed.

3.3.1 Number of Imputation Iterations Required?

As explained in section 3.2.4, the approach iteratively updates the imputed values, intending for them to stabilise. For the original method, Breiman and Cutler (2004) state four to six iterations are typically sufficient. It was unclear whether this would also be the case for the approach developed when applied to the NACC data; thus, work was undertaken to find out. In order to determine an appropriate number of

iterations, the size of the random forest (i.e. number of trees) constructed during each iteration needed to be set. In short, this experiment was carried out with 100 trees, namely the common number as indicated by Geurts, Ernst and Wehenkel (2006), which the experiment described in section 3.3.2 confirmed was sufficient.

Neither Breiman and Cutler (2004) or Cutler, Cutler and Stevens (2012) provide a test for stability which would permit a suitable number of iterations to be identified. In the case of this experiment, the imputation performance for each iteration of the approach, as well as the initial imputation step, was assessed and compared, specifically for the training set. It was explained in the previous section that 1,000 missing values were introduced into the training set to enable the imputation performance to be assessed. In fact, this process was repeated a further 20 times, along with the imputation itself, to allow the variability of the performance to also be determined. For all 21 invocations of the approach, 10 iterations were executed so as to err on the side of caution, and imputation performance was evaluated based on the true and imputed values corresponding to the 1,000 additional missing values, denoted by X^{true} and X^{imp} respectively.

Stekhoven and Bühlmann (2012), who proposed missForest (section 3.1.3), assess imputation performance using the normalised root mean squared error (NRMSE) for continuous variables and the proportion of falsely classified entries (PFC) for categorical variables. These measures, which can be defined using the following equations, were utilised during this experiment.

$$\text{NRMSE} = \sqrt{\frac{mean((X^{true} - X^{imp})^2)}{var(X^{true})}} \tag{3.6}$$

$$\text{PFC} = \frac{\sum \mathbb{I}(X^{true} \neq X^{imp})}{n_{val}} \tag{3.7}$$

For the NRMSE, $mean(\cdot)$ and $var(\cdot)$ represent the empirical mean and variance over the continuous values. For the PFC, $\mathbb{I}(\cdot)$ is an indicator function which equals one when the true and imputed values fail to match for any of the categorical, ordinal or binary variables; and n_{val} is the number of true-imputed value pairs, specifically for

these types of variables. Notably, lower values indicate better imputation performance for both the NRMSE and PFC.

As highlighted above, the ordinal and binary variable values contributed to the PFC. This was due to the fundamental nature of the variables, as well as the possibility of conditionally missing values in X^{true} and/or X^{imp} for a number of them; conditionally missing values are problematic for the NRMSE. Only six of the variables for which missing values were added were continuous, and two of them could be imputed with a conditionally missing value. As a result, true-imputed value pairs which included at least one conditionally missing value were also considered for the PFC and the NRMSE was calculated based on a relatively small number of value pairs. In particular, these two variables provide the number of years the subject has smoked cigarettes (SMOKYRS) and the age at which they quit smoking, if applicable (QUITSMOK). The other four continuous variables provide the number of years of education the subject received (EDUC), their systolic and diastolic blood pressure (BPSYS, BPDIAS), and their resting heart rate (HRATE).

Figure 3.1 presents the PFC and NRMSE for each imputation iteration, including the initial imputation step which results in crudely imputed values, for all 21 invocations of the approach (left). To clarify, the set of 1,000 additional missing values was different for every invocation. It also presents the mean PFC and NRMSE for each iteration across these invocations (right), for which error bars are given showing one standard deviation, characterising the variability of the imputation performance. Interestingly, there is little change in the PFC (blue) and NRMSE (orange) after the initial imputation step for all invocations of the approach. It should be noted that the initial decrease, which almost invariably occurred, indicates the first iteration improved on the crudely imputed values.

It was previously explained that the imputation performance was compared across the iterations in order to determine an appropriate number of them. As shown by figure 3.1, the number of iterations suggested by Breiman and Cutler (2004) for the original proximity imputation approach (i.e. four to six), or even less, may well have been sufficient. 10 iterations were utilised, nonetheless, as minor changes in

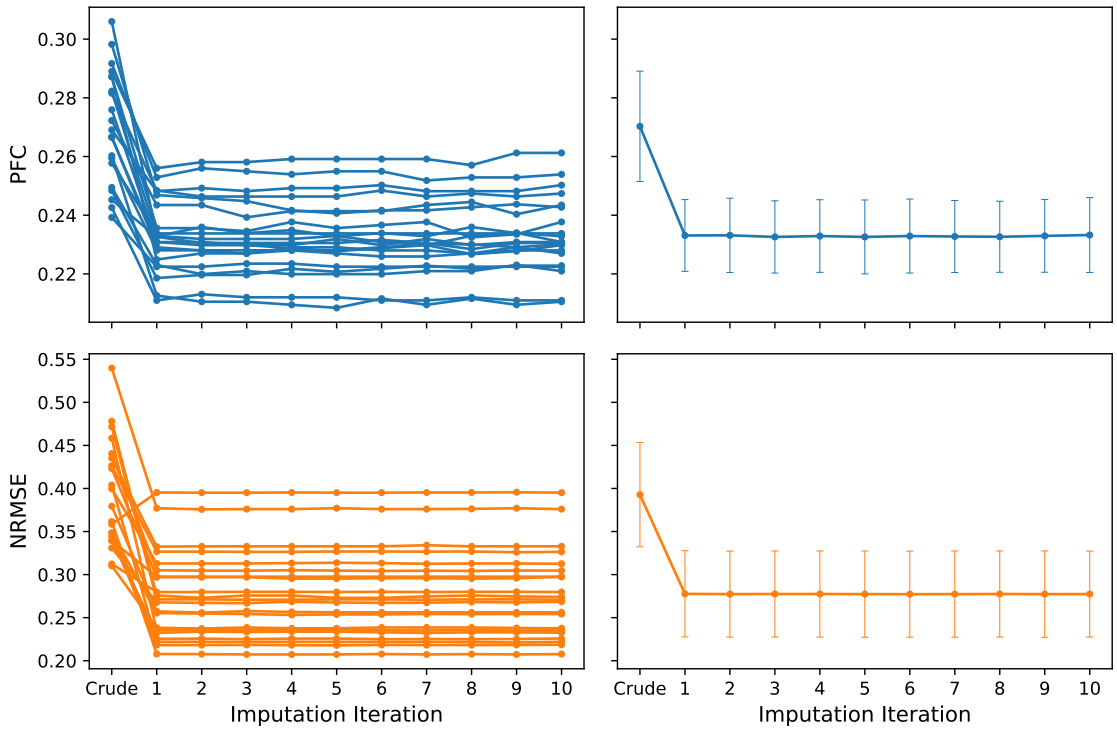


Figure 3.1: Imputation performance, as defined by the proportion of falsely classified entries (PFC) and the normalised root mean squared error (NRMSE), for each imputation iteration, including the initial imputation step which results in crudely imputed values. To be specific, the performance values for all 21 invocations of the approach are indicated (left), along with the mean values across these invocations (right), for which error bars are given showing one standard deviation.

performance were apparent past six iterations for individual invocations.

3.3.2 Number of Trees Required?

In addition to the number of imputation iterations, the number of trees constructed during each iteration needed to be set. As explained in section 3.3.1, 100 trees were used to determine a suitable number of iterations, but it was then necessary to verify that 100 trees were sufficient. Geurts, Ernst and Wehenkel (2006) highlight that enough trees must be utilised to ensure convergence of the ensemble effect.

Proximity imputation with MIA was applied to the training and test sets using a variety of ensemble sizes, ranging from 10 to 100 in steps of 10. 10 classifiers resulted, each of which were associated with their own imputed training and test sets. The performance of every classifier was ascertained by first generating M classifications (or predictions) for each observation in its imputed test set, notably one for each

tree in the ensemble. As discussed in section 1.3.1, a tree makes a prediction based on the class majority of the terminal node reached by the observation. However, as only one class was represented in each terminal node due to the way in which the trees were constructed, the predicted class of the observation (or subject) was simply that which was associated with the terminal node. The arithmetic mean of every set of predictions was then computed to produce a set of ensemble scores, which were essentially estimates of a subject's probability of dementia. Finally, these scores were used, along with the classification targets and a resampling method known as bootstrapping, to determine the (mean) area under the receiver operating characteristic curve (AUC) and 95% confidence interval for the classifier. In particular, the AUC was calculated using the relevant scores and targets for 2,000 bootstrap samples, each of which were generated by randomly sampling subjects from the test set with replacement, and were the same size as the test set. From these values, the mean AUC and 95% confidence interval were deduced.

To put this into context, a receiver operating characteristic (ROC) curve (or graph) shows the true positive rate versus the false positive rate as the classification threshold is varied, providing insight into the classifier's performance. Here, the positive class is 'dementia' and the negative class is 'no dementia'. The AUC, on the other hand, indicates the average performance of the classifier over the range of classification thresholds but, more specifically, is a measure of how well the probability distributions for the two classes are separated (Fawcett, 2006; Hand and Till, 2001). It is also considered to be equivalent to the probability of the classifier scoring a randomly chosen instance of the positive class more highly than that of the negative class (Fawcett, 2006).

Figure 3.2 shows the AUC for each classifier, along with the 95% confidence intervals by means of the shaded region. Notably, the range of AUCs is very small. Regardless, there is a relatively large increase in AUC from 10 to 20 trees, but then it becomes much more gradual. The AUC appears to level off at around 50 trees, and can be considered reasonably stable for larger ensembles despite very small increases. As previously stated, it was ultimately confirmed that 100 trees were sufficient.

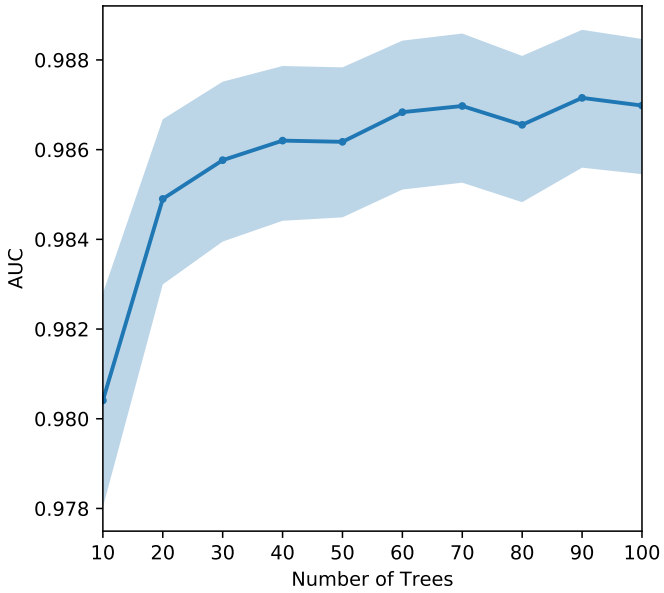


Figure 3.2: Classification performance, as defined by the area under the receiver operating characteristic curve (AUC) and 95% confidence interval, for a range of ensemble sizes.

3.3.3 Effects of Additional Missingness

As explained in section 3.3, the capability of the approach developed was assessed by investigating the effects of additional missingness, specifically in the training set, on the imputation and classification performance. Originally, the proportion of missing values in the training set was 0.65%, which was increased to 0.67% when the 1,000 extra missing values were added. For this experiment, the proportion of missing values was increased further. In fact, an additional 5, 10, 15 and 20% of the values were converted to missing in the manner described in section 3.3. Proximity imputation with MIA was applied to the four new training sets, as well as the test set several times, using 10 imputation iterations and 100 trees; this resulted in four new classifiers with their own imputed training and test sets. A number of implausible values were imputed for each training set, fundamentally due to the way in which missing values were added, but they replaced just 2.8% of the new missing values.

Imputation performance was assessed for each of the training sets by calculating the NRMSE and PFC as described in section 3.3.1. By using the true and (final) imputed values corresponding to the 1,000 missing values originally added, the performance could be compared. Figure 3.3 presents the NRMSE and PFC for

each training set, showing that the imputation performance was fairly consistent as missingness was increased. However, the PFC (blue) does suggest a slight decrease in performance as missingness was increased from 10.67% to 15.67%.

Classification performance was ascertained for each of the classifiers as outlined in section 3.3.2. Figure 3.4 shows the AUC for each classifier, along with the 95% confidence intervals by means of the shaded region. Similarly to figure 3.2, the range of AUCs is very small. In short, classification performance decreased as missingness was increased, but the change in AUC was very gradual. It is highly unlikely the classification performance decreased due to the introduction of implausible values into the various training sets, as the variables involved in the applicable relations were found to be of relatively low importance for differentiating between subjects with and without a diagnosis of dementia. For more information on variable importance, the reader is directed to chapter 4.

It was expected that the imputation and classification performance would decrease as the proportion of missing values was increased, and this did appear to be the case. Nonetheless, the performance seemed to be only marginally affected for the proportions considered.

3.4 Summary

There is a vast amount of literature on handling missing data. Prior to choosing an approach, it is considered important to determine the mechanism behind any missingness, specifically for the purposes of ensuring it is dealt with appropriately. There are three widely accepted mechanisms: missing completely at random (MCAR), missing at random (MAR) and missing not at random (MNAR). Due to the relatively low degree of missingness in the NACC data set, and the complexity of modelling mechanisms of missingness, no formal investigation was undertaken to determine the mechanisms of the missing values; and they were treated as if they were MAR.

Imputation replaces missing values with suitable substitutions, and is one way of dealing with missing data. Single imputation methods, such as mean imputation, hot deck imputation and regression imputation, generate a single value for every

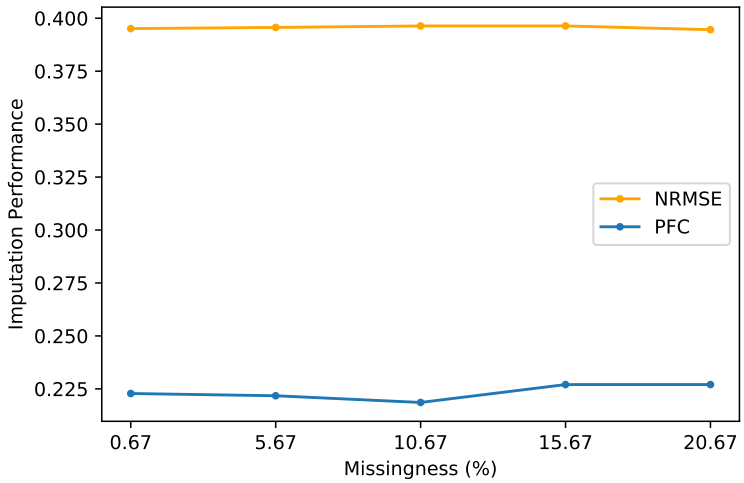


Figure 3.3: Imputation performance, as defined by the normalised root mean squared error (NRMSE) and the proportion of falsely classified entries (PFC), for data with additional missingness.

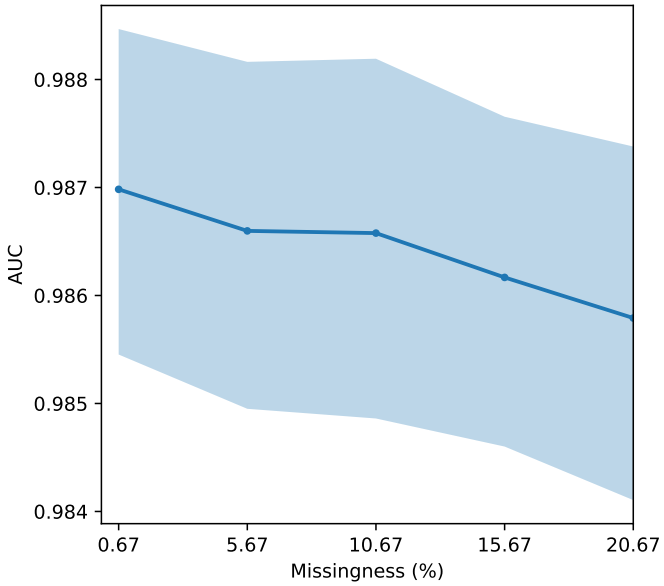


Figure 3.4: Classification performance, as defined by the area under the receiver operating characteristic curve (AUC) and 95% confidence interval, using data with additional missingness.

missing value; their use is generally discouraged. Multiple imputation methods, for example data augmentation, produce multiple values for each missing value; these types of techniques are highly recommended but are more labour-intensive.

The NACC UDS is an example of a data set which includes variables in relations with one another; the variables which can be determined by these relations are referred to as derived variables. The literature discussing the imputation of derived variables is relatively recent and limited, but the general consensus seems to be that plausible values are desirable, namely values which do not introduce inconsistencies into the data. For some types of derived variables, specifically tailored approaches are required to obtain plausible values.

Decision trees were chosen for classification, but they can also be used to perform imputation. Decision-tree-based approaches are advantageous as they are able to handle mixed data, and two prominent examples are missForest and proximity imputation. In addition to missing values, the NACC data set contains conditionally missing values. Missingness incorporated in attributes (MIA) is an acclaimed technique for handling missing data in decision trees, which is suitable for dealing with the conditionally missing values.

The proximity imputation method was the natural choice for the NACC data and this research, as it is able to deal with mixed data and it enables the imputation to be closely coupled with the construction of a random forest classifier. Notably, a random forest is an ensemble of different decision trees, each of which have been generated using a process with an element of randomness. In addition, it was possible to utilise MIA in conjunction with the proximity imputation approach to handle the conditionally missing values.

The approach developed is proximity imputation with MIA. It begins by eliminating any missing values in the data set (or training set), leaving only those which are conditionally missing. In particular, a simple implementation of hot deck imputation is used which replaces each missing value with an observed value for the variable from an observation associated with the same class. Once all the missing values have been crudely imputed, the imputed data set can be used to construct a

random forest. The random forest is built using the Extra-Trees algorithm, along with MIA, the latter of which handles the conditionally missing values without imputing them. Essentially, more splits are generated for each variable considered for splitting that has conditionally missing values.

By inspecting the paths of the observations through every tree in the random forest, the similarity (or proximity) of each pair of observations can be ascertained. In fact, the proximity of two observations for a single tree is the total information gain across the common nodes of their paths, weighted by the proportion of observations at each node when the tree was built. If the two paths have the terminal node in common, an extra quantity must be added as terminal nodes do not have an associated information gain. Alternatively, the weighted information gain could be accumulated according to whether the subsequent nodes in the paths are also common. This would render the additional value corresponding to the terminal node obsolete, and result in observations having zero proximity if their paths diverge after the root node. Ultimately, the proximities for each tree are normalised, and those pertaining to the ensemble are calculated by simply averaging across the trees.

The proximities for the ensemble are used to impute the missing values more rigorously by means of an $N-\text{by}-N$ proximity matrix, where N is the number of observations. Fundamentally, a missing value is imputed with the proximity-weighted average or proximity-weighted mode of the observed values for the variable. As conditionally missing values can arise of their own accord in the NACC data set, a conditionally missing value could legitimately be imputed for certain variables; this is considered when calculating the proximity-weighted average or mode. The imputation is also staggered in order to maintain the known relations between variables in the NACC UDS, so far as possible. At first, the missing values of the variables which can determine others are imputed. The derived variables of the dependencies and relationships are then updated, if and where appropriate; and, finally, all the remaining missing values are imputed. Certain values are prevented from informing the imputation at the various stages, and those predetermined by the relations could be also.

The approach iteratively updates the imputed values, repeatedly generating a new random forest and proximities in the process, intending for the values to stabilise. For the purposes of analysis, a random forest is also built using the final imputed data set, and proximities are calculated. Incidentally, the approach can also be used to impute test cases with a few alterations. Crucially, the imputed values are generated based on the imputed training cases alone.

Four to six iterations are typically sufficient for the proximity imputation method. However, 10 iterations were ultimately utilised for the approach developed, despite there being little change in the imputation performance after the initial imputation step when it was applied to the training set. In particular, imputation performance was assessed using the normalised root mean squared error (NRMSE) and the proportion of falsely classified entries (PFC), which were calculated for 1,000 missing values added to the training set.

100 trees were used to determine a suitable number of imputation iterations, but it was imperative to verify 100 trees were sufficient. In order to do so, the approach was applied to the training and test sets using a range of ensemble sizes; this resulted in a number of random forest classifiers. The performance of every classifier was then assessed on their respective imputed test sets by means of the area under the receiver operating characteristic curve (AUC). In short, it was confirmed that 100 trees were adequate.

The effects of additional missingness on the imputation and classification performance were investigated by increasing the proportion of missing values in the training set from 0.67%. In particular, the approach was applied to the new training sets, as well as the test set several times, using 10 imputation iterations and 100 trees. Classification performance was assessed on the test set using the AUC, whilst the NRMSE and PFC were utilised to evaluate imputation performance on the 1,000 missing values originally added to the training set. In summary, the imputation and classification performance decreased as the proportion of missing values was increased, but the performance seemed to be only marginally affected for the proportions considered.

Chapter 4

Diagnosing Dementia and Differentiating between Subtypes

As explained in chapter 1, one of the primary aims of the research was to investigate the use of machine learning for distinguishing between people, specifically Alzheimer's Disease Center (ADC) subjects, with and without dementia, as well as differentiating between key dementia subtypes where appropriate. In particular, the four main dementia subtypes were considered, namely Alzheimer's disease (AD), vascular dementia (VD), dementia with Lewy bodies (DLB) and frontotemporal dementia (FTD). The approach outlined in chapter 3 was developed for this purpose, and results from the imputation and classification of the NACC data are presented in this chapter. In fact, results pertaining to the dementia classifier, which was constructed during the experiments described in section 3.3, are discussed, along with results concerning pairwise dementia subtype classifiers and a stacking classifier. This chapter also explains how the additional classifiers were built and puts the work into context by discussing related literature and the clinical implications of the findings.

4.1 Dementia Classifier

As previously stated, the dementia classifier was constructed during the experiments described in section 3.3. In particular, the approach outlined in chapter 3, namely proximity imputation with MIA (missingness incorporated in attributes), was em-

ployed with 10 imputation iterations and 100 trees to simultaneously impute the training set and build a random forest classifier. The training set comprised 22,801 subjects selected at random, whilst its classification targets indicated whether each subject had a received a diagnosis of dementia at their initial visit. It was explained in section 2.4 that cognitive status was broken down into dementia, mild cognitive impairment (MCI) and normal cognition; thus, the targets separated the subjects diagnosed with dementia from those diagnosed with MCI or normal cognition. Of the 22,801 subjects, 8,500 had been diagnosed with dementia and 14,301 had been diagnosed with either MCI (4,737) or normal cognition (9,564); these totals suggest there was a slight imbalance of the classes.

This section details the performance of the classifier in terms of the ensemble and the trees comprising it. It also provides the results of the seriation analysis, which give some insight into the similarity of subjects; and the variable importance investigation that aimed to identify the most important variables for diagnosing dementia. In addition, it recounts the work undertaken to determine the number of variables required to match the performance of the classifier, which makes use of all 260 variables, whilst taking the importance of each variable into account. This work indicated a number of assessments should be prioritised, and the research conducted to investigate their importance further is also discussed.

4.1.1 Classification Performance

The test set, which included the remaining 9,772 subjects, was also imputed using proximity imputation with MIA, enabling the performance of the classifier to be assessed for both the training set and the test set. In order to assess the performance, M classifications (or predictions) were generated for each observation in the training or test set, notably one for each tree in the ensemble. As discussed in section 1.3.1, a tree makes a prediction based on the class majority of the terminal node reached by the observation. However, as only one class was represented in each terminal node due to the way in which the trees were constructed, the predicted class of the observation was simply that which was associated with the terminal node. A

		True Class		
		Dementia	No Dementia	Total
Predicted Class	Dementia	3373	303	3676
	No Dementia	263	5833	6096
	Total	3636	6136	9772

Table 4.1: Confusion matrix for the dementia classifier.

single prediction was also generated for each observation, using the appropriate set of classifications, by identifying the class with the most predictions. In fact, the trees effectively voted on the most suitable class (or diagnosis) for every subject. If the trees predicted each class equally, the tie was broken randomly.

One measure of performance is the accuracy of the classifier, which was simply calculated by comparing the set of votes (or predicted classes) to the classification targets (or true classes), and determining the percentage of subjects correctly classified. As previously highlighted, each terminal node in every tree was only associated with a single class; thus, the accuracy for the training set was 100%. Consequently, only the performance of the classifier for the test set is discussed. Prior to this, it is worth noting that the true classes, namely the diagnoses provided by NACC, may be subject to error, despite being based on the results of extensive examinations; this is, fundamentally, due to the fact that it is currently difficult to diagnose dementia reliably. As a result, it should be kept in mind that any inconsistencies between the true and predicted classes could have been due to an incorrect diagnosis by a clinician or otherwise, rather than an erroneous prediction by the classifier.

Of the 9,772 subjects in the test set, 3,636 had dementia and the remaining 6,136 did not. The accuracy of the classifier for the test set was 94.21%; and the confusion matrix, presented in the form of table 4.1, details how its subjects were classified with respect to their true classes. From the information provided by the confusion matrix, the sensitivity (true positive rate) and specificity (true negative rate) of the dementia classifier can be deduced. The sensitivity, in this context, can be defined as the proportion of subjects with dementia that were correctly classified. The specificity, on the other hand, is the proportion of subjects without dementia who

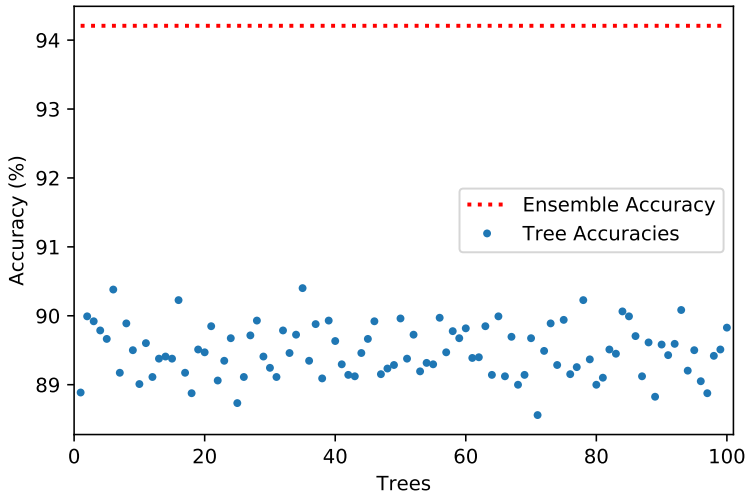


Figure 4.1: Classification accuracies for the trees comprising the ensemble which constitutes the dementia classifier, along with the accuracy for the ensemble itself.

were correctly classified. Consequently, the sensitivity was 0.93 and the specificity was 0.95. These values are high, and indicate that there were relatively few instances of false negatives and false positives in which those with and without dementia were incorrectly classified. Interestingly, 86% of the 303 subjects without dementia that were incorrectly classified had MCI. This suggests that MCI is more difficult to differentiate from dementia than normal cognition, as to be expected.

In addition to the accuracy of the ensemble (i.e. classifier), the accuracy of each tree was determined by comparing their predictions to the set of targets. Figure 4.1 provides the tree accuracies in blue, along with the overall accuracy indicated by the red dotted line. It clearly demonstrates that forming an ensemble is beneficial, as its accuracy is considerably higher than that of each tree on its own.

Ensemble scores, which were generated as detailed in section 3.3.2, were used in conjunction with the targets to produce a receiver operating characteristic (ROC) curve (Fawcett, 2006), in order to gain a better understanding of the classifier's performance. The area under the ROC curve (AUC) was also calculated. The blue ROC curve, in figure 4.2, shows the true positive rate (TPR or sensitivity) versus the false positive rate (FPR or 1 - specificity) as the classification threshold is varied. The curve passes particularly close to point (0, 1) in the top left corner of the graph, which signifies perfect classification; thus, it can be concluded that the classifier is

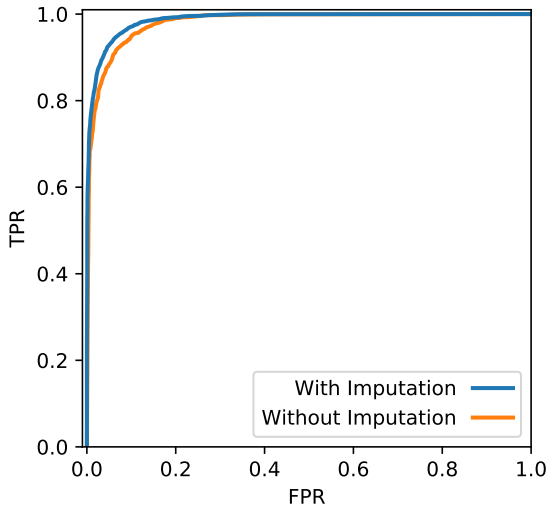


Figure 4.2: Receiver operating characteristic (ROC) curves for dementia classifiers with and without imputation.

able to perform very well. The AUC, which indicates the average performance of the classifier over the range of classification thresholds (Fawcett, 2006), was 0.99. The best possible value is 1.00, reinforcing that there is very little room for improvement.

To provide further justification for the imputation of the missing values present in the data set, another dementia classifier was trained and tested using only the subjects and variables that were free from missing values. As highlighted in section 2.3.2, 47.57% (15,494) of the 32,573 subjects had at least one missing value. Of the 260 variables, 48.46% (126) had at least one missing value. Ultimately, these subjects and variables were removed from the original training and test sets, and a new dementia classifier was constructed. Notably, the new training set included 11,947 subjects and 134 variables, whilst the new test set included 5,132 subjects and 134 variables. The classifier's performance was subsequently assessed using the test set, for which its accuracy, sensitivity and specificity was determined. An ROC curve was also produced, enabling the AUC to be calculated.

The original dementia classifier (with imputation) achieved an accuracy of 94.21%, a sensitivity of 0.93, a specificity of 0.95 and an AUC of 0.99. The new dementia classifier (without imputation) performed marginally worse as it had an accuracy of 92.83%, a sensitivity of 0.89, a specificity of 0.95 and an AUC of 0.98; its (orange) ROC curve, in figure 4.2, also corroborates this. It should be kept in

mind, however, that the test set used to assess the performance of the new dementia classifier was a subset of that which was used for the original. Despite the fact that there was no significant degradation in performance, the imputation step can still be considered valuable. Not only did the original dementia classifier perform better on a larger test set, but also dropping subjects could have biased the data set as it is possible those with severe cognitive impairment were more prone to missing values. Additionally, it was important to preserve subjects to ensure the four main dementia subtypes could be investigated and, although not also necessary in practice, dropping variables would have greatly reduced the scope of the data set.

4.1.2 Seriation Analysis

Seriation can be employed to reveal some of the underlying structure of a data set. It does so by arranging the observations in a sequence along a one-dimensional continuum, specifically placing similar observations close to one another (Liiv, 2010). Spectral seriation (Atkins, Boman and Hendrickson, 1998), in particular, was used to gain an understanding of the training and test set subjects in terms of their similarity; it is similar to spectral clustering, which is covered in chapter 5.

The proximity (similarity) matrix P, which was populated using the random forest as explained in section 3.2.3, was required. More specifically, the proximities calculated using equations 3.3 to 3.5, namely the similarities between observations determined based on their paths through the trees in the random forest, were used to populate P. By using the decision trees to ascertain similarities, both the continuous and categorical variables, as well as the other types of variables in the data set (ordinal and binary), were naturally drawn on in a way which would be very hard to achieve by hand. In addition, the similarities were specialised to the task of distinguishing between subjects with and without a diagnosis of dementia.

For computational simplicity and to ease visualisation, 500 subjects were randomly sampled from each of the training and test sets and the relevant fragments of P were analysed. In fact, seriation was employed for each of these smaller similarity matrices, which are also denoted by P for the sake of simplicity. Initially,

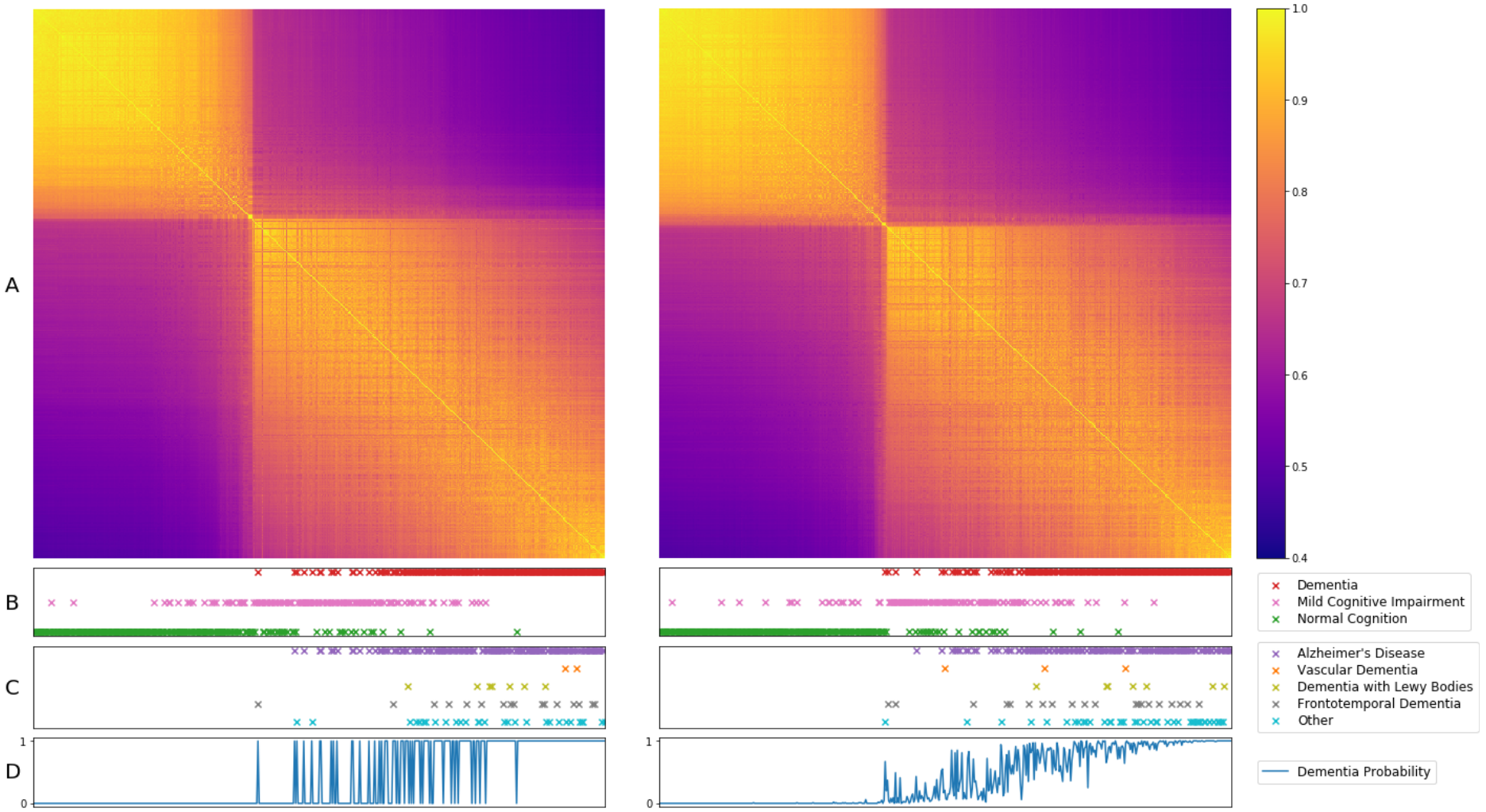
the (unnormalised) Laplacian matrix $\mathcal{L}$ of P was determined using

$$\mathcal{L} = D - P\,, \tag{4.1}$$

where D is the diagonal (or degree) matrix which has the sum of the similarities for every subject along the diagonal. The eigenvalues and eigenvectors of $\mathcal{L}$ were then calculated, and the eigenvector corresponding to the smallest non-zero eigenvalue was identified. This eigenvector is known as the Fiedler vector, and it was ordered to form the permutation vector. Finally, P was symmetrically permuted by reordering its rows and columns using said permutation vector.

Part A of figures 4.3(a) and 4.3(b) shows P reordered according to the permutation vector for the training and test set subjects respectively. The pairs of subjects which are very similar are coloured yellow/orange, whilst those that are less similar are coloured blue/purple. As all the subject pairs have at least a similarity of 0.47, the colourmap is not extended below 0.4. In each visualisation, the pairs situated along the diagonal have a similarity of one, and it is these pairs which compare a subject with itself. As a result of the seriation, the similarity broadly decreases moving away from the diagonal but, most importantly, the subjects clearly fall into one of two distinct groups. The subjects within each of these groups are similar to one another and dissimilar to those in the other group; thus, two clusters have effectively been discovered.

Part B indicates the cognitive status of each subject, ordered according to the permutation vector. Normal cognition is shown in green, mild cognitive impairment (MCI) in pink, and dementia in red. When it is considered in conjunction with part A, it is possible to conclude that the classifier can differentiate between subjects with normal cognition and those with dementia well. It is also shown that, on the whole, the subjects with dementia are more similar to each other than they are to those that are cognitively normal, and vice versa. In contrast, the MCI subjects do not form their own distinct group. As a matter of fact, the vast majority are included in the group of subjects with dementia. Nevertheless, they tend to be situated between the cognitively normal subjects on the left and those with dementia on the right, forming



(a) Seriated training set sample.

(b) Seriated test set sample.

Figure 4.3: Seriated training and test set samples (A) annotated with each subject's cognitive status (B), dementia subtype diagnosis where applicable (C), and dementia probability according to the classifier (D).

a spectrum of cognitive impairment. There is much debate as to whether MCI is a clinical entity (i.e. a condition in its own right), and how useful a diagnosis of MCI is in itself, as highlighted by Zanetti, Geroldi and Frisoni (2009), NeurologyToday (2004), Pinto and Subramanyam (2009) and Beard and Neary (2013). These results suggest that MCI may not be a clinical entity but rather a mild form of dementia.

Part D provides (an estimate of) the dementia probability for every subject which, as highlighted in section 3.3.2, is the ensemble score. A probability of one corresponds to a diagnosis of dementia, whilst a probability of zero indicates the inverse (i.e. no dementia). The range of probabilities differs for the training and test set subjects, due to the way in which the classifier was constructed. In fact, the probability of dementia for those in the training set is either zero or one, coinciding with the targets used to train the classifier, as only one class was represented in each terminal node of every tree in the ensemble. In figure 4.3(b), the probability of dementia for the subjects comprising the test set sample increases, on average, from left to right, appearing to be indicative of cognitive impairment severity. From left to right, 82.76% of those with normal cognition have zero probability of dementia. The probability increases for the MCI subjects, reflecting their mild impairment; and this increase continues for those with dementia. Of the dementia subjects sampled, 23.44% have a probability of one, suggesting severe impairment. The variation in probability for this group of subjects is likely to reflect the differing levels of impairment experienced by individuals, despite receiving the same diagnosis.

Figure 4.4 provides a more detailed view of the dementia probabilities, in terms of cognitive status, for the subjects sampled from the test set. In fact, it is effectively an amalgamation of parts B and D of figure 4.3(b), showing the dementia probability for each subject coloured according to their cognitive status. It corroborates that the probability of dementia appears to be indicative of cognitive impairment severity, whilst highlighting the transition from MCI to dementia roughly coincides with the classification threshold at 0.5.

Reverting back to figure 4.3, part C indicates the dementia subtype diagnoses for those with a dementia diagnosis. Only the ‘pure’ cases of the four main subtypes

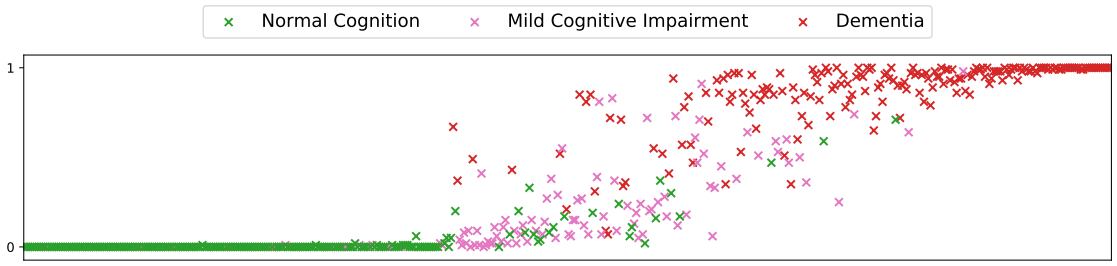


Figure 4.4: Dementia probability coloured according to cognitive status for each of the subjects sampled from the test set, ordered with regards to the permutation vector.

(AD, VD, DLB and FTD) are highlighted, meaning the subject received a primary diagnosis of the subtype and no supplementary diagnoses of any of the other main subtypes. All the remaining cases of dementia are indicated as 'other'. In contrast to part B, in which the subjects were grouped by cognitive status as a result of seriation, the subjects are not arranged by dementia subtype. As a result, it can be concluded that the dementia classifier is unable to differentiate between the key dementia subtypes.

In section 2.1 it was highlighted that the majority of subjects attend follow-up visits. At every follow-up visit the subject is provided with a diagnosis which could be different from before. For example, a subject could have been diagnosed with MCI at their initial visit, but dementia at a subsequent visit. There is likely to have been one of two reasons for this, either the subject's cognitive impairment had worsened or the subject was previously incorrectly diagnosed. Within the NACC data there are instances in which subjects appear to revert from dementia back to normal cognition, but progressive dementias, which include the four main subtypes, are not reversible (Mayo Clinic Staff, 2019).

For the test set, in particular, 2,777 of the 9,772 subjects had visited their respective Alzheimer's Disease Centers (ADCs) only once. Of the 6,995 subjects that made multiple visits, 4,925 had the same cognitive status for every one; therefore, 2,070 subjects received at least one alternative diagnosis. In order to determine whether these changes in diagnosis (i.e. cognitive status) would reveal anything in relation to the spectrum of cognitive impairment formed, the nature of the changes were ascertained in terms of progression and reversion for the sampled subjects.

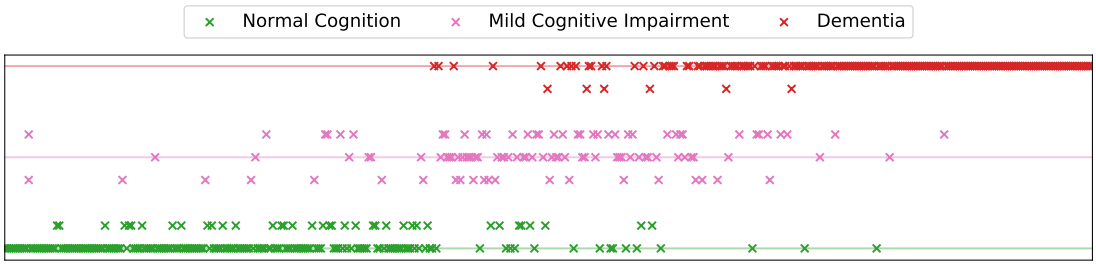


Figure 4.5: Changes in cognitive status over time, in terms of progression and reversion, for the subjects sampled from the test set. The subjects are arranged along the x-axis according to the original permutation vector. Moreover, the cognitive status of each subject at their initial visit is indicated using colour; and progression and/or reversion is shown by the displacement of points, in the y-direction, from the baseline for each cognitive status.

Figure 4.5 was produced by modifying part B of figure 4.3(b), so it indicated the direction of change with regards to each subject's initial diagnosis. In order to indicate progression, the relevant points were displaced upwards, whilst points were shifted downwards to signify reversion. For those subjects with an initial diagnosis of MCI, there was the possibility of both progression and reversion; thus, it is possible for a single subject to have two points. Although changes are indicated across the majority of the spectrum, the subjects situated in the centre appear to be more prone to change. Progression as opposed to reversion also seems to be more likely, which is to be expected. Interestingly, several of the subjects initially diagnosed with MCI which were later diagnosed with dementia, specifically towards the left of the spectrum, were predicted to have approximately zero probability of dementia. Inspecting the variables across the visits for these subjects in particular may uncover information which could prove useful in predicting change in cognitive impairment, which is a potential avenue for future research.

In summary, spectral seriation separated the subjects into two distinct groups, comprising, for the most part, subjects with dementia or MCI and subjects with normal cognition. As a result, it was shown that the subjects with dementia were more similar to each other than they were to those with normal cognition, and vice versa. In particular, the MCI subjects tended to be situated between those with normal cognition and dementia, forming a spectrum of cognitive impairment; and those subjects situated in the centre of the spectrum appeared to be more

prone to change in their cognitive status. Based on these results, it may be more suitable to consider MCI as a mild form of dementia rather than a condition in its own right. In addition, this analysis highlighted that the classifier can differentiate between subjects with normal cognition and those with dementia well. However, it is unable to distinguish between the key dementia subtypes. It also suggested that the dementia probabilities generated by the classifier are indicative of cognitive impairment severity.

4.1.3 Variable Importances

A benefit of using decision trees for classification is it is possible to determine the importance of each variable for said classification, as highlighted in chapter 1. At the outset this was an attractive feature, as it would enable the most important variables (or features) for diagnosing dementia to be identified. In this section, exactly how the variable importances were calculated is detailed, along with the variables which could be considered to be the key diagnostic features for dementia. Prior to this, the five variables that were included in the data set to test the validity of the importances are discussed; these variables were unrelated to the classification targets.

The inclusion of variables in the data set for testing purposes was first considered in section 2.3.1. One of the variables was already present, and it provided the visit number (NACCVNUM); the variable was constant due to only initial visits being utilised. The other four variables were generated and added to the data set, specifically for testing. One of these variables (RAND_VAR) was created by randomly sampling from a normal distribution. The remaining three were produced by randomly permuting variables from the data set, and each of them was of a different type. The variables were generated in this manner so they each had a realistic distribution but lacked correlation with the classification targets. There was a synthetic binary variable (RAND_BVAR) which was based on INSEX (co-participant's sex), a synthetic categorical variable (RAND_CVAR) that was a permutation of TRAILB_PROB (reason Trail Making Test Part B not completed), and a synthetic ordinal/continuous variable (RAND_DOCVAR) generated using CDRSUM (Clinical

Dementia Rating (CDR) sum of boxes). These variables, in particular, were chosen as they were free from missing values and, in the case of TRAILB_PROB and CDRSUM, were reasonably representative of their type in terms of number of unique values. For example, TRAILB_PROB has four categories, which was the average across all the categorical variables. Conditionally missing values were present in two of the three variables, but the underlying reasons for them were ignored. All five of the variables were expected to be of negligible importance.

In the case of the dementia classifier, the variable importances quantify the significance of the variables in the prediction of dementia or no dementia. Initially, the importances were determined on a tree-by-tree basis, fundamentally by counting the number of instances in which a variable X^f was used to perform a split S. The count was weighted by the information gain resulting from the split $\mathcal{I}_S(X)$ and the proportion of observations split on $\frac{N_S}{N}$ when the tree t was constructed. Once the tree-based importances had been determined for all the variables, they were normalised to transform the absolute importances into relative scores that summed to one for each tree, as in scikit-learn (Pedregosa et al., 2011). Essentially, this step was performed to ensure none of the trees disproportionately influenced the importances. The variable importances were then averaged across the ensemble T. The following equation summarises how the (unnormalised) variable importance (VIMP) of a single variable X^f can be calculated, which is, in fact, the mean decrease impurity (MDI) importance (Louppe et al., 2013).

$$\text{VIMP}(X^f) = \frac{1}{|T|} \sum_{t \in T} \sum_{\eta \in t} \mathcal{I}_S(X) \frac{N_S}{N} \mathbb{I}(S \text{ split on } X^f) \tag{4.2}$$

$\mathbb{I}(\cdot)$ is an indicator function, representing the count, which equals one when the variable of interest X^f matches the variable associated with the split S for the node η. Before the importances were inspected, they were scaled to set the score of the most important variable to 100%.

If there are similar variables in the data set, the importance of the information they provide may be split between them. However, it can be assumed that no two variables are identical, meaning valuable insight could still be revealed on an

elemental basis. It may also be possible to identify variables which are similar to one another, enabling their respective importances to be considered in combination. Within the NACC data set, duplicate information was minimised as much as possible; thus, reducing the likelihood of importance being shared by variables.

Figure 4.6 presents the variable importances ascertained from the dementia classifier. Part B, in particular, provides the importances for all 260 variables in decreasing order, and indicates that most variables are of very little importance. As a matter of fact, 194 of the variables have an importance score between 0% and 4%, including the one constant and four synthetic variables discussed earlier. In particular, the constant variable (NACCVNUM) has a score of 0%, whilst the four synthetic variables (RAND_VAR, RAND_DOCVAR, RAND_CVAR and RAND_BVAR) have scores of 2.26%, 2.15%, 3.51% and 3.55%. NACCVNUM, which provides the visit number, is the only variable with exactly 0% importance, due to the fact it could not be chosen as one of the K variables considered for splitting at an internal node as it was constant. Interestingly, the synthetic binary variable (RAND_BVAR) appears to be of marginally more importance than the variable it was based on, namely that which provides the co-participant's sex (INSEX), by 1.1%. However, it is likely this occurred due to INSEX being of very little importance itself, as the other two synthetic variables that were generated by random permutation seem to be of much less importance than their highly ranked (i.e. important) counterparts.

Part A of figure 4.6 gives a closer look at the 60 most important variables for diagnosing dementia. Abridged descriptions of the variables are provided based on those from the researchers data dictionary (National Alzheimer's Coordinating Center, 2017), and the bars of variables pertaining to the Clinical Dementia Rating (CDR), Functional Activities Questionnaire (FAQ) and Mini-Mental State Examination (MMSE) are coloured accordingly; these assessments are discussed in more detail in the following sections. The top two variables, which indicate whether the subject was impaired in judgment, planning or problem-solving (COGJUDG), and pertain to the home and hobbies category of the CDR (HOMEHOBB), appear to be considerably more important than all the others. In particular, the home and hobbies category

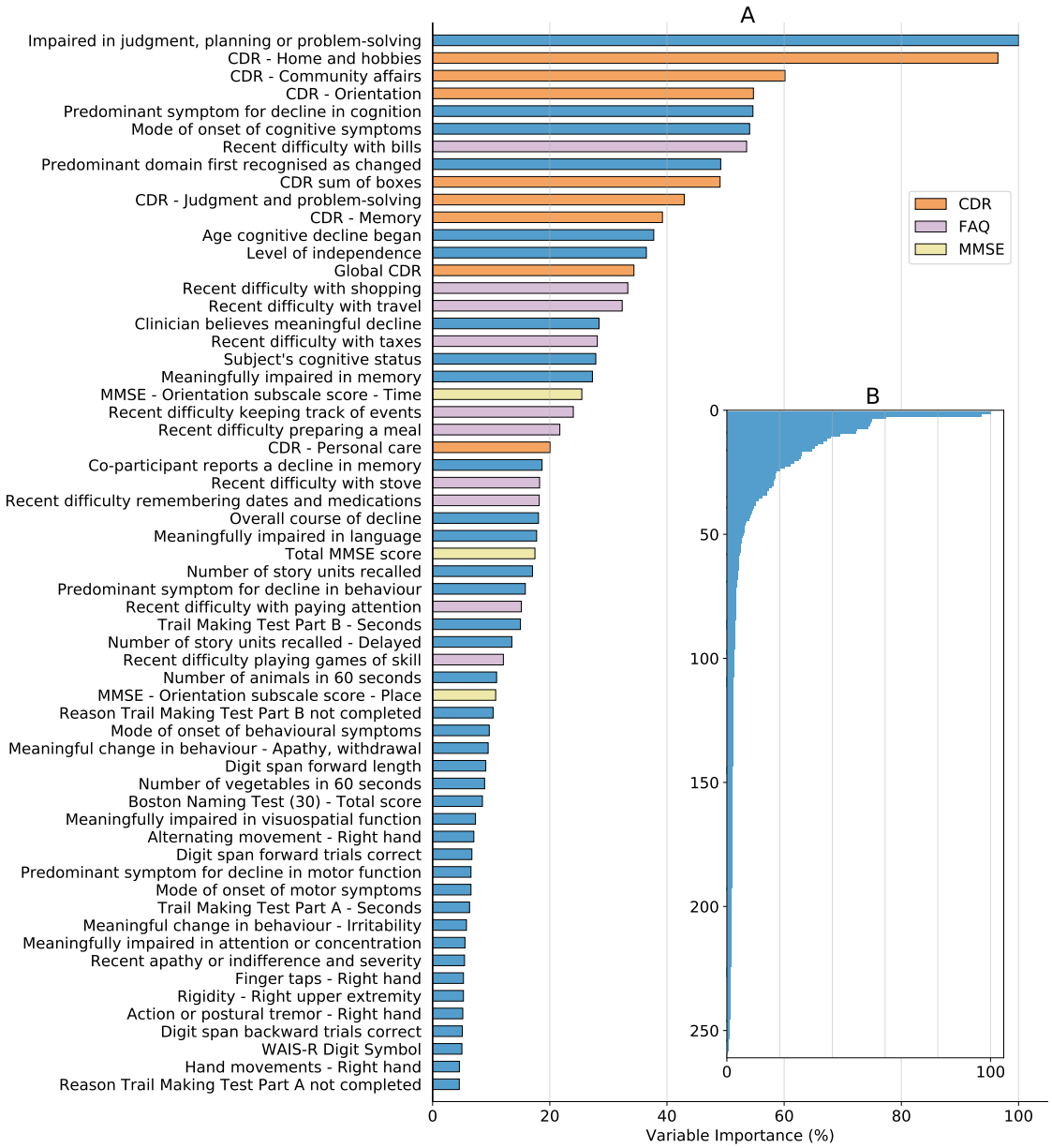


Figure 4.6: The 60 most important variables for diagnosing dementia according to the dementia classifier, along with their importances (A). The importances for all 260 variables (B) are also indicated.

of the CDR broadly assesses whether the subject has been able to carry out chores around the home, and if they are still engaging in hobbies and intellectual interests (ADC Clinical Task Force and National Alzheimer's Coordinating Center, 2014a). These variables relate directly to the general definition of dementia, provided in chapter 1. The remaining variables constituting the top 10 important features concern the community affairs (COMMUN), orientation (ORIENT), and judgment and problem-solving (JUDGMENT) categories of the CDR; the CDR sum of boxes (CDRSUM); the predominant symptom for decline in cognition (NACCCOGF); the mode of onset of cognitive symptoms (COGMODE); any recent difficulty with bills (BILLS); and the predominant domain first recognised as changed (FRSTCHG). The common themes of these variables are cognitive impairment and the subject's ability to engage in activities of daily living, which correspond with the fundamental aspects clinicians consider when diagnosing dementia.

4.1.4 Performance Matching

The ADCs which contribute to the NACC data set are specialised centres. It is unlikely that subjects could be assessed to the same degree outside of specialised centres, such as in primary care, due to constraints on time and resources. In these settings, prioritising data collection in accordance with the variable importances could be beneficial. However, there is no clear indication as to what might constitute a sufficient amount of data for a diagnosis of some substance to be made. The aim of this additional investigative work was to identify the number of variables required to match the performance of the dementia classifier, which utilised all 260 variables; and deduce the subset of features which could be considered fundamental for the diagnosis of dementia. Despite the fact the NACC UDS encompasses a wide range of clinical and neuropsychological data, it does not include all types of data used to diagnose dementia, for instance neuroimaging data is absent. As a result, further research investigating the importance of these absent features is required before a complete list of fundamental features can be proposed.

60 new classifiers were constructed in total; the first used only the most

important variable, namely that which indicated whether the subject was impaired in judgment, planning or problem-solving (COGJUDG). Each subsequent classifier utilised one more variable than its predecessor, and the variables were added in order of importance; thus, the last classifier exploited the top 60 variables, as shown in part A of figure 4.6. No changes were made to the way in which each classifier was built, but the training set was not imputed again to significantly reduce the computational time. As a matter of fact, the imputed training set corresponding to the original dementia classifier was utilised, along with the imputed test set which was needed to ascertain classification performance. In particular, the performance of each classifier was assessed based on a condensed version of the test set, including only the variables used to construct the former. Also, the procedure described in section 3.3.2 was followed. Essentially, a set of predictions was generated for every subject, each of which were converted into an ensemble score. These scores were then used in conjunction with the classification targets to determine the (mean) AUC and 95% confidence interval for the classifier, by employing bootstrapping with 2,000 bootstrap samples.

In figure 4.7, the AUC for each classifier is indicated by the blue line, and the associated confidence intervals are shown by the blue shading. The AUC and 95% confidence interval for the original classifier are also provided in red for reference. The figure shows a clear increase in performance, which is almost monotonic, as the number of variables increases from one to 19. The increase then becomes more gradual from 19 to 32 variables, and even more so from 32 to 42 variables. At 42 variables the performance of the (original) classifier which uses all 260 variables is reached; and, from this point on, the performance is approximately constant. In order to confirm that the performance of the classifier which utilised just 42 variables was statistically indistinguishable from that of the original classifier, one-way analysis of variance (ANOVA) was employed with a significance level of 0.05.

As a result of these findings it can be concluded that the top 42 variables for diagnosing dementia, according to the original dementia classifier, are able to match the performance of all 260 variables in the data set; and it is these features

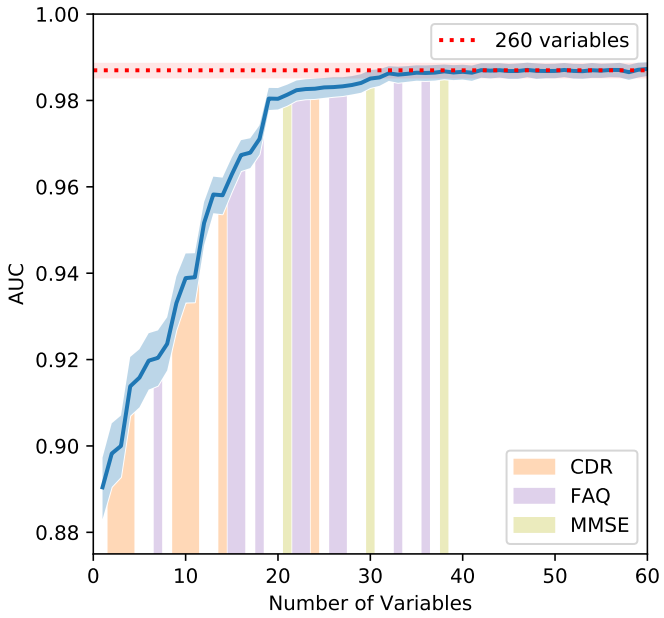


Figure 4.7: Classification performance, as defined by the area under the receiver operating characteristic curve (AUC) and 95% confidence interval, for an increasing number of variables. The performance of the dementia classifier, which uses all 260 variables, is also indicated.

that could be considered fundamental for dementia diagnosis. Despite the fact that 42 is still a reasonably large number of variables, a closer look at them suggests there are three main assessments to prioritise, accounting for exactly half of the variables. These assessments are the Clinical Dementia Rating (CDR), Functional Activities Questionnaire (FAQ) and Mini-Mental State Examination (MMSE). In figure 4.7, the point at which each of the 21 variables pertaining to either the CDR, FAQ or MMSE was introduced is highlighted by means of coloured bars indicating the relevant assessment. Incidentally, more variables may have been significant if more training observations had been available, as it is possible the performance of the original classifier would have been greater.

Figure 4.7 shows that increasing the number of variables from 19 to 42 results in only a small gain in performance. In fact, the AUC for the classifier which utilises just 19 variables is 0.980, whilst the AUC for the classifier that uses 42 variables is 0.987. Consequently, it could be worth considering the effort and resources required to collect the data encapsulated in the additional 23 variables, and whether the small gain in performance outweighs this. If just 19 variables were used, the MMSE, for

example, would not need to be completed.

4.1.5 Assessment Exclusion

In the previous section, it was suggested that there are three main assessments to prioritise, specifically the CDR, FAQ and MMSE. The importance of these assessments, along with Form B9, was investigated further by systematically excluding the sets of features brought about by the assessments (including Form B9) and creating new dementia classifiers whose performance could be compared to that of the original. Form B9, which explores the symptoms the subject was experiencing, was considered as 13 of the 42 variables that could be deemed fundamental for diagnosing dementia were associated with it. Incidentally, there were eight CDR, 10 FAQ, eight MMSE and 32 Form B9 variables.

15 new classifiers were trained and tested using the imputed training and test sets with various combinations of features excluded. As previously, the imputed training and test sets were used to significantly reduce the computational time. Table 4.2 shows that the accuracy, sensitivity (true positive rate or TPR) and specificity (true negative rate or TNR) of each classifier was ascertained, as well as the false negative rate (FNR or 1 - sensitivity) and false positive rate (FPR or 1 - specificity). ROC curves, which are provided in figure 4.8, were also produced and the AUCs were calculated (table 4.2).

For ease of comparison, the statistics pertaining to the original dementia classifier are included in table 4.2 and its ROC curve is present in figure 4.8. Overall, there was relatively little change in performance, suggesting the vast majority of the information encapsulated within the variables excluded could be garnered from those that remained. In other words, the four assessments (CDR, FAQ, MMSE and Form B9) do not seem to be irreplaceable. Nonetheless, the CDR could be considered marginally more important than the other assessments. It may also be unwise to overlook the CDR as the naïve Bayes classifier (from section 2.6) which predicted dementia with an accuracy of 92.25% utilised a CDR variable (CDRSUM).

Classifier	Accuracy % (no.)	Sensitivity TPR (no.)	Specificity TNR (no.)	FNR (no.)	FPR (no.)	AUC
260 variables	94.21 (9206)	0.93 (3373)	0.95 (5833)	0.07 (263)	0.05 (303)	0.99
No CDR	93.11 (9099)	0.92 (3347)	0.94 (5752)	0.08 (289)	0.06 (384)	0.98
No FAQ	94.37 (9222)	0.93 (3373)	0.95 (5849)	0.07 (263)	0.05 (287)	0.99
No MMSE	94.09 (9194)	0.93 (3368)	0.95 (5826)	0.07 (268)	0.05 (310)	0.99
No Form B9	93.95 (9181)	0.92 (3334)	0.95 (5847)	0.08 (302)	0.05 (289)	0.99
No CDR & FAQ	93.07 (9095)	0.93 (3367)	0.93 (5728)	0.07 (269)	0.07 (408)	0.98
No CDR & MMSE	93.00 (9088)	0.92 (3341)	0.94 (5747)	0.08 (295)	0.06 (389)	0.98
No CDR & Form B9	92.47 (9036)	0.90 (3283)	0.94 (5753)	0.10 (353)	0.06 (383)	0.98
No FAQ & MMSE	94.39 (9224)	0.93 (3368)	0.95 (5856)	0.07 (268)	0.05 (280)	0.99
No FAQ & Form B9	94.15 (9200)	0.91 (3324)	0.96 (5876)	0.09 (312)	0.04 (260)	0.99
No MMSE & Form B9	94.12 (9197)	0.92 (3335)	0.96 (5862)	0.08 (301)	0.04 (274)	0.99
No CDR & FAQ & MMSE	92.96 (9084)	0.92 (3358)	0.93 (5726)	0.08 (278)	0.07 (410)	0.98
No CDR & FAQ & Form B9	91.48 (8939)	0.88 (3211)	0.93 (5728)	0.12 (425)	0.07 (408)	0.97
No CDR & MMSE & Form B9	92.01 (8991)	0.90 (3256)	0.93 (5735)	0.10 (380)	0.07 (401)	0.98
No FAQ & MMSE & Form B9	93.96 (9182)	0.91 (3316)	0.96 (5866)	0.09 (320)	0.04 (270)	0.99
No CDR & FAQ & MMSE & Form B9	91.03 (8895)	0.88 (3186)	0.93 (5709)	0.12 (450)	0.07 (427)	0.97

Table 4.2: Classification performance of dementia classifiers with assessments excluded, namely the Clinical Dementia Rating (CDR), Functional Activities Questionnaire (FAQ), Mini-Mental State Examination (MMSE) and Form B9. The performance of the original dementia classifier, which utilised all 260 variables, is also detailed. Performance was measured in terms of the accuracy, sensitivity (true positive rate or TPR), specificity (true negative rate or TNR), false negative rate (FNR) and false positive rate (FPR) of the classifier, and the area under the receiver operating characteristic curve (AUC).

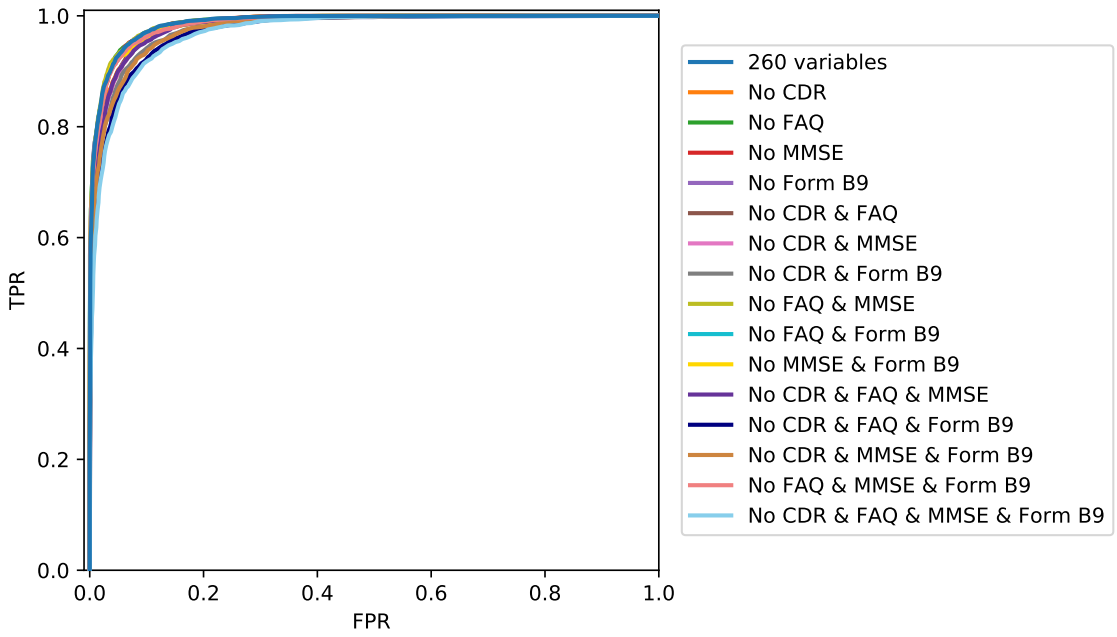


Figure 4.8: Receiver operating characteristic (ROC) curves for dementia classifiers with assessments excluded, specifically the Clinical Dementia Rating (CDR), Functional Activities Questionnaire (FAQ), Mini-Mental State Examination (MMSE) and Form B9. The ROC curve for the original dementia classifier, which utilised all 260 variables, is also shown.

4.2 Pairwise Dementia Subtype Classifiers

In order to determine whether machine learning could discern the main subtypes of dementia, pairwise dementia subtype classifiers were constructed. These classifiers were trained to differentiate between two of the four key subtypes, or a key subtype and alternative dementia diagnoses (referred to as 'other'). Notably, the main subtypes of dementia are Alzheimer's disease (AD), vascular dementia (VD), dementia with Lewy bodies (DLB) and frontotemporal dementia (FTD). This section explains how the classifiers were constructed, along with how each of them performed. It also indicates which features were found to be most important for the differential diagnosis of dementia on a pair-by-pair basis and in general.

4.2.1 Construction

In order to construct the 10 pairwise dementia subtype classifiers, and assess their performance, multiple training and test sets had to be created which comprised

Classifier		Training Set			Test Set		
S1	S2	S1	S2	Total	S1	S2	Total
AD	DLB	3595	258	3853	1819	123	1942
AD	FTD	3595	568	4163	1819	288	2107
AD	Other	3595	266	3861	1819	154	1973
AD	VD	3595	75	3670	1819	21	1840
DLB	FTD	258	568	826	123	288	411
DLB	Other	258	266	524	123	154	277
FTD	Other	568	266	834	288	154	442
VD	DLB	75	258	333	21	123	144
VD	FTD	75	568	643	21	288	309
VD	Other	75	266	341	21	154	175

Table 4.3: Composition of the training and test sets for the pairwise dementia subtype classifiers. S1 and S2 are shorthand for subtype one and subtype two, which were either Alzheimer's disease (AD), vascular dementia (VD), dementia with Lewy bodies (DLB), frontotemporal dementia (FTD) or 'other'.

subjects diagnosed with only the relevant subtypes (including 'other'). These new training and test sets were formed using the subjects in the original training set ($N = 22,801$), and the imputed data was used to significantly reduce the computational time required to build and test each classifier. In fact, the 10 new conditioned training sets were produced from a random sample that contained two-thirds ($N = 15,201$) of the subjects in the imputed training set, whilst the 10 corresponding test sets were created using the remaining one-third ($N = 7,600$). Only 'pure' cases of the subtypes were drawn upon, to ensure those representing each of the subtypes were as distinct as possible. A subject was considered to have a pure diagnosis of a main subtype if that subtype was provided as the primary diagnosis, and they had not received a diagnosis of any of the other main subtypes. Within the original training set there were 5,414 pure cases of AD, 96 of VD, 381 of DLB and 856 of FTD. A subject was considered to have a pure alternative dementia diagnosis, on the other hand, if they had been diagnosed with dementia but none of the four key subtypes; there were 420 of these ('other') cases in the original training set. As the number of cases for each subtype varied, the training and test sets created were all different sizes. Table 4.3 provides the number of subjects in each training and test set, along with a breakdown by subtype.

The subtype classifiers were constructed, in the aforementioned manner, using the 10 new conditioned training sets, along with classification targets which differentiated between the subtypes. The performance of each classifier in terms of accuracy, sensitivity and specificity, among other things, is discussed in the next section.

4.2.2 Classification Performance

The performance of each of the pairwise dementia subtype classifiers was assessed using its conditioned training and test sets, but the former was only used to confirm that the classifier was 100% accurate for its training set. Predictions were generated for every subject in the set, and used to determine ensemble votes (or predicted classes) and scores, as was customary. The latter, in this case, were estimates of the probability that the subject had dementia subtype one. The accuracy, sensitivity (true positive rate or TPR) and specificity (true negative rate or TNR) were then calculated, by comparing the predicted and true classes; and an ROC curve was produced, using the ensemble scores and true classes, which enabled the AUC to be computed. The sensitivity, in this context, was the proportion of subjects with subtype one that were correctly classified, so the specificity was the proportion of subjects with subtype two who were classified as such. As noted in section 4.1.1, the true classes may be subject to error; thus, this should be kept in mind when considering the results.

Table 4.4 presents the accuracy, sensitivity and specificity of each pairwise dementia subtype classifier for its test set, along with the AUC. The number of subjects correctly classified in total, as well as for subtypes one and two (S1 & S2), are also provided. All the accuracies are above 90%, except for the FTD v Other classifier which has an accuracy of 80.77%. The sensitivities are also all relatively high. In fact, the minimum is 0.81, and three of the classifiers have a sensitivity of 1.0 (rounded to two decimal places). The specificities, on the other hand, are much more varied, ranging from 0.38 to 1.0 (again, rounded to two decimal places). When the sensitivities and specificities are considered together, there appears to be some bias towards subtype one for the four AD classifiers, as well as for the FTD v Other

Classifier		Accuracy	Sensitivity	Specificity	AUC
S1	S2	% (no.)	TPR (no.)	TNR (no.)	
AD	DLB	97.37 (1891)	1.00 (1814)	0.63 (77)	0.98
AD	FTD	95.25 (2007)	0.99 (1805)	0.70 (202)	0.98
AD	Other	94.98 (1874)	1.00 (1815)	0.38 (59)	0.88
AD	VD	99.40 (1829)	1.00 (1816)	0.62 (13)	0.94
DLB	FTD	94.89 (390)	0.88 (108)	0.98 (282)	0.98
DLB	Other	90.25 (250)	0.88 (108)	0.92 (142)	0.95
FTD	Other	80.77 (357)	0.93 (269)	0.57 (88)	0.88
VD	DLB	95.83 (138)	0.86 (18)	0.98 (120)	0.99
VD	FTD	98.38 (304)	0.81 (17)	1.00 (287)	1.00
VD	Other	96.00 (168)	0.81 (17)	0.98 (151)	0.98

Table 4.4: Classification performance of the pairwise dementia subtype classifiers. S1 and S2 are shorthand for subtype one and subtype two, which were either Alzheimer's disease (AD), vascular dementia (VD), dementia with Lewy bodies (DLB), frontotemporal dementia (FTD) or 'other'. Performance was measured in terms of the accuracy, sensitivity (true positive rate or TPR) and specificity (true negative rate or TNR) of the classifier, and the area under the receiver operating characteristic curve (AUC).

classifier. It is possible that this occurred, for the AD classifiers in particular, due to an imbalance of the classes in the conditioned data sets. The performance of these classifiers for subtype two could potentially be improved, however, by adjusting the classification threshold. The threshold was 0.5, by default, as the predicted classes were decided by majority vote across the ensemble. The AUC is independent of the classification threshold, as it indicates the average performance of the classifier, so it could be considered a more reliable measure of performance. All the AUCs are very high, and only two classifiers have an AUC below 0.94. In particular, these two classifiers have an AUC of 0.88, have the lowest specificities, and compare a key dementia subtype (AD or FTD) to the set of alternative diagnoses (Other). It is not surprising that the classifiers with the worst performance, in terms of the AUC, attempt to distinguish between a main subtype and 'other', as the latter includes subjects with a variety of dementia diagnoses instead of a single subtype, making its subjects harder to classify.

Figure 4.9 shows the ROC curves for the 10 pairwise dementia subtype classifiers. It reiterates that the AD v Other (dark green) and FTD v Other (pink) classifiers are the poorest performers. It also highlights that the VD v FTD classifier (light green)

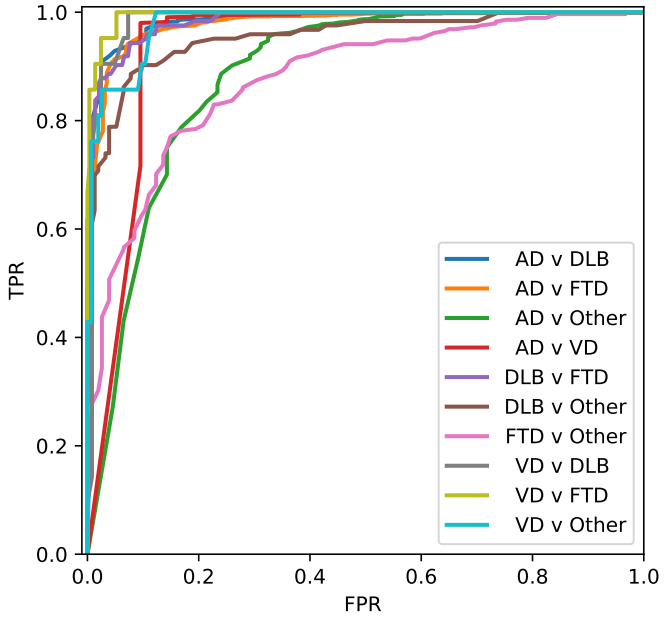


Figure 4.9: Receiver operating characteristic (ROC) curves for the pairwise dementia subtype classifiers.

is the best, although it did not achieve perfect classification as its AUC (rounded to two decimal places) in table 4.4 suggests. Nevertheless, the ROC curves all pass reasonably close to point (0, 1) in the top left corner of the graph, and indicate that the trade-off between sensitivity and specificity could most likely be improved for at least some of the classifiers. Overall, the pairwise dementia subtype classifiers are able to perform well, suggesting that machine learning could be used to differentiate between the key dementia subtypes.

4.2.3 Variable Importances

In order to identify the key features for the differential diagnosis of dementia, the importance of every variable was determined for each of the pairwise dementia subtype classifiers in turn, following the procedure outlined in section 4.1.3. This resulted in 10 sets of rankings, each of which denoted the order of the variables in terms of decreasing importance.

Table 4.5 provides the top five variables, in the form of abridged descriptions, for each pairwise dementia subtype classifier, which could be considered the key features for differentiating between the two specified subtypes (including 'other'). A

AD v DLB	1	Years from Parkinson's disease diagnosis
	2	Parkinson's disease
	3	Years from parkinsonian disorder diagnosis
	4	Other parkinsonian disorder
	5	Meaningful change in behaviour - Visual hallucinations
AD v Other	1	Predominant domain first recognised as changed
	2	Eye movement abnormalities indicative of CNS disorder
	3	Gait disorder indicative of CNS disorder
	4	Focal neurological symptoms
	5	Meaningfully impaired in memory
DLB v FTD	1	Meaningful change in behaviour - Visual hallucinations
	2	Parkinson's disease
	3	Years from Parkinson's disease diagnosis
	4	Meaningful change in behaviour - Disinhibition
	5	Other parkinsonian disorder
FTD v Other	1	Eye movement abnormalities indicative of CNS disorder
	2	Mode of onset of cognitive symptoms
	3	Predominant symptom for decline in cognition
	4	Gait disorder indicative of CNS disorder
	5	Meaningful change in behaviour - Disinhibition
VD v FTD	1	History of stroke
	2	Years from last stroke
	3	Stroke
	4	Stepwise deterioration (cognitive status)
	5	Abrupt onset (cognitive status)
AD v FTD	1	Predominant symptom for decline in cognition
	2	Meaningfully impaired in memory
	3	Predominant domain first recognised as changed
	4	Meaningful change in behaviour - Personality change
	5	Meaningful change in behaviour - Disinhibition
AD v VD	1	History of stroke
	2	Years from last stroke
	3	Stroke
	4	Hachinski ischemic score
	5	Focal neurological signs
DLB v Other	1	Meaningful change in behaviour - Visual hallucinations
	2	Parkinson's disease
	3	Years from Parkinson's disease diagnosis
	4	Eye movement abnormalities indicative of CNS disorder
	5	Meaningful changes in motor function - Slowness
VD v DLB	1	History of stroke
	2	Stroke
	3	Years from last stroke
	4	Hachinski ischemic score
	5	Mode of onset of motor symptoms
VD v Other	1	History of stroke
	2	Years from last stroke
	3	Stroke
	4	Transient ischemic attack
	5	Hachinski ischemic score

Table 4.5: Top five variables for each pairwise dementia subtype classifier.

CNS - central nervous system

number of variables appear to be associated with a particular subtype, such as the variable that indicates whether the subject currently manifests meaningful change in behaviour in the form of visual hallucinations (BEVHALL). This variable, in particular, is in the top five for three of the four DLB classifiers. Another example is the variable which specifies whether the subject currently manifests meaningful change in behaviour with respect to disinhibition (BEDISIN). It is also featured in the top five for three classifiers, but they were trained using FTD diagnoses. In addition, it seems the most important variables for the VD and DLB classifiers have a common theme, specifically stroke and Parkinson's disease (and other parkinsonian disorder) respectively. It should be noted, however, that combining subjects with a diagnosis of stroke and VD, and those with Parkinson's disease and DLB, (section 2.4) could have potentially inflated the importance of the variables associated with these themes. Nevertheless, it is extremely reassuring that the features and themes linked to specific subtypes generally reflect the current diagnostic criteria.

Figure 4.10 presents the complete set of variable importances for each pairwise dementia subtype classifier. The plots highlight the rate at which importance declines, as the variables have been arranged according to the rankings. In fact, it can be ascertained that importance decreases very quickly for the VD classifiers and relatively slowly for the 'other' classifiers (excluding VD v Other). The latter is unsurprising as the 'other' subjects had not been assigned a specific subtype diagnosis, so the defining features of these subjects were likely to be much more varied. The FTD v Other classifier, in particular, has the slowest decline, and the variables low in the rankings appear to still be of some importance. This could partly explain why the FTD v Other classifier was one of the poorest performers, and it may suggest that many of the diagnoses designated as 'other' exhibited similar symptoms to FTD.

To determine the key features for the differential diagnosis of dementia in more general terms, the 10 sets of variable importances were summed and then scaled so the most important variable had an importance of 100%. The resultant importances are referred to as the *differential importances*. Figure 4.11 shows the 60 most important variables for differential diagnosis, according to these differential importances.

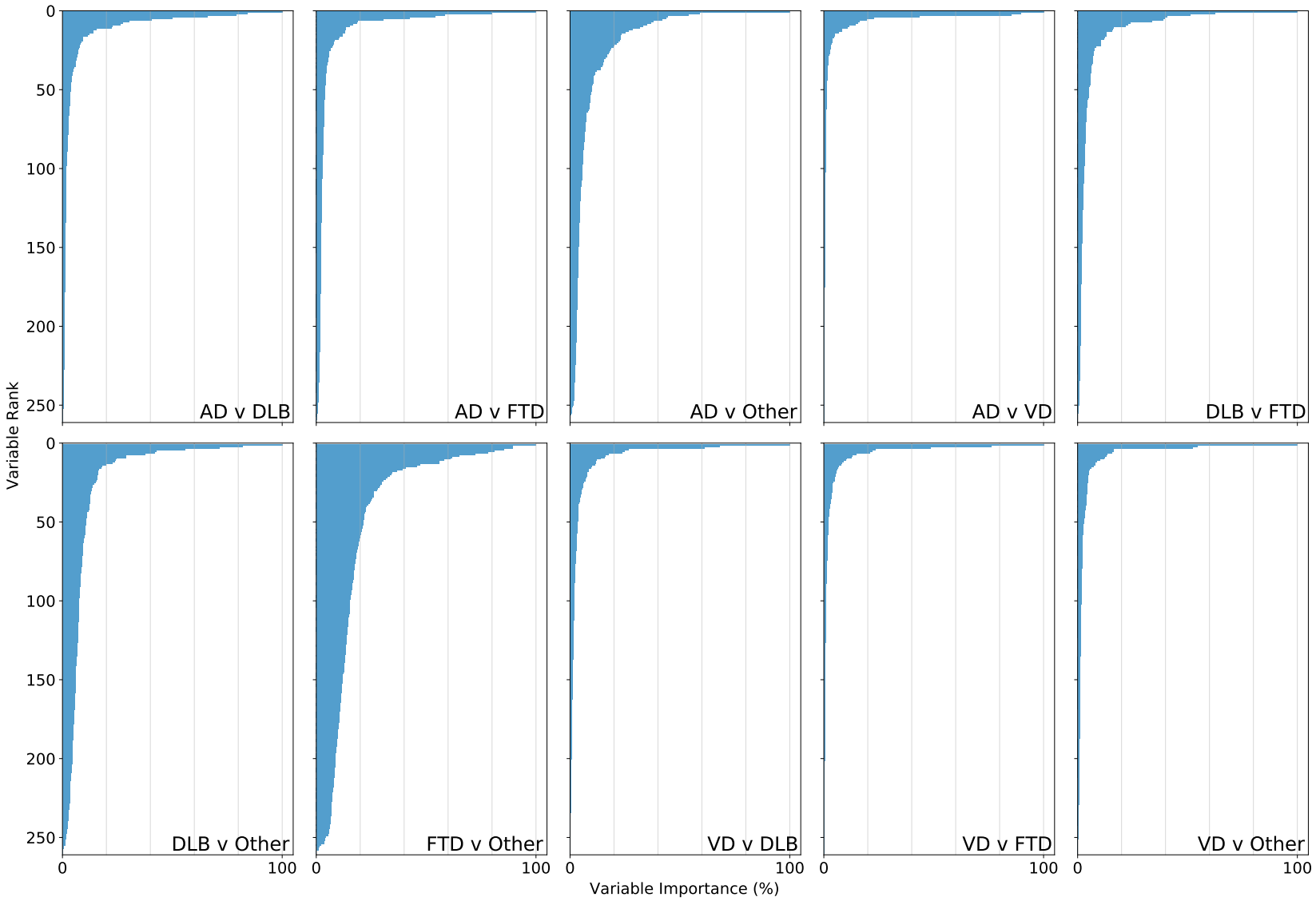


Figure 4.10: Importances for all 260 variables, in descending order, for each pairwise dementia subtype classifier.

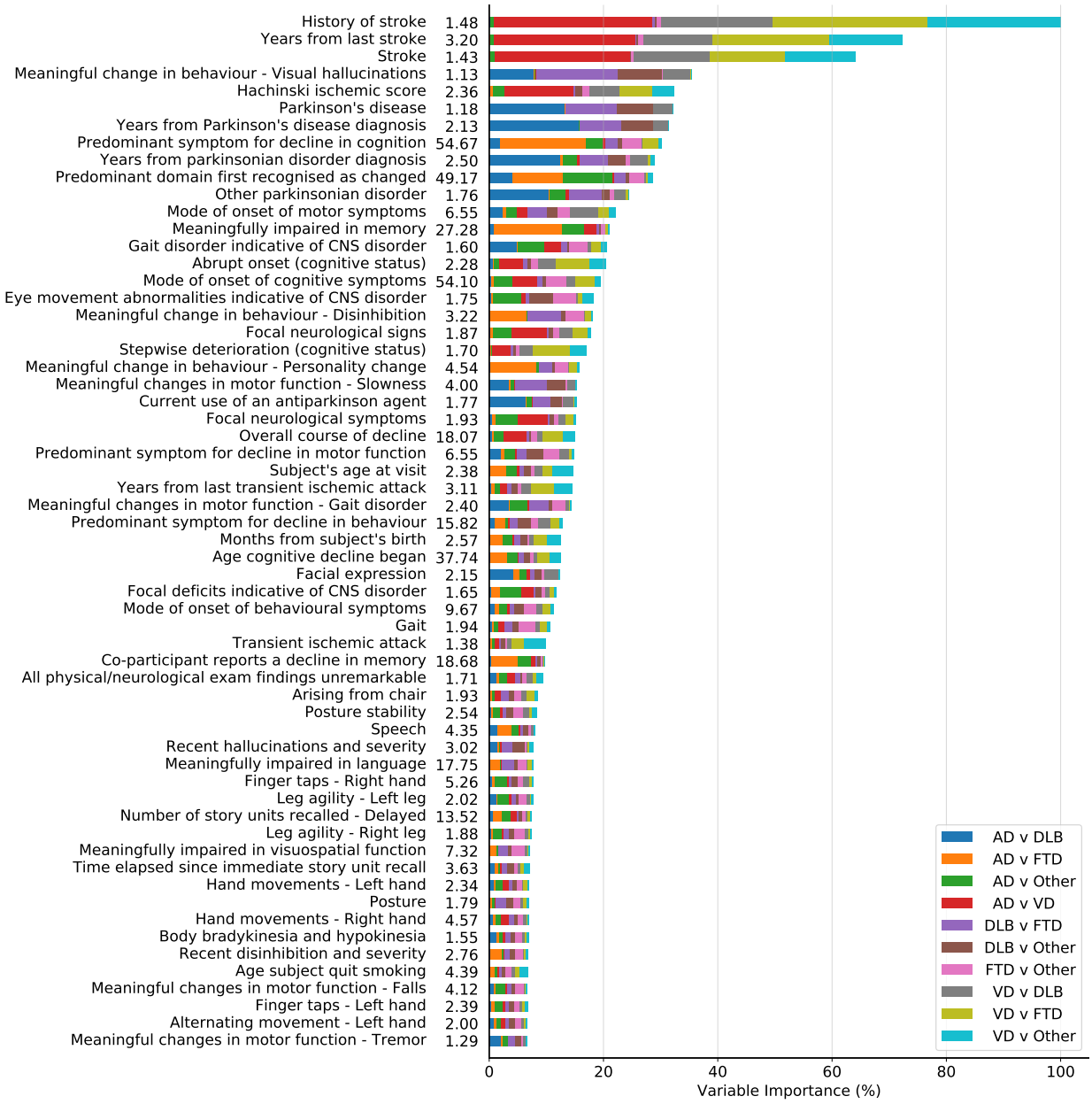


Figure 4.11: The 60 most important variables for the differential diagnosis of dementia according to the pairwise dementia subtype classifiers, along with their importances. The importance of each variable for diagnosing dementia, according to the dementia classifier, is also provided (to the right of the variable's description).

Similarly to figure 4.6, abridged descriptions of the variables are provided based on those from the researchers data dictionary (National Alzheimer's Coordinating Center, 2017). The multi-coloured bars indicate the differential importance of each variable; and every colour corresponds to a single pairwise dementia subtype classifier, showing its contribution.

The top three variables, which provide the subject's stroke history (HXSTROKE, CBSTROKE) and the number of years since their last stroke, if applicable (NACCSTYR_#YRS), seem to be considerably more important than all the others. Incidentally, the two variables pertaining to stroke history record the information in different ways. HXSTROKE, which is the most important, simply indicates whether the subject has previously had a stroke; this information was used to calculate the Hachinski ischemic score. CBSTROKE, on the other hand, goes into more detail and specifies whether the subject has had a clinical or silent stroke. As both variables are ranked very highly, it appears that their importance was not severely affected by their similarity. The rankings of the two variables could also suggest that a simple indication of stroke is the most valuable way to record this information. Interestingly, figure 4.11 shows that the top three variables almost exclusively got their importance from the VD classifiers. Likewise, those ranked fourth, sixth and seventh acquired almost all of their importance from the DLB classifiers. These three variables indicate whether the subject is having visual hallucinations (BEVHALL); has been diagnosed with Parkinson's disease (PD); and the number of years since the Parkinson's disease diagnosis, if applicable (PDYR_#YRS). The close connections apparent between these variables and subtypes echo the analysis of table 4.5.

The remaining variables comprising the top 10 important features for the differential diagnosis of dementia concern the Hachinski ischemic score (HACHIN); the predominant symptom for decline in cognition (NACCCOGF); the number of years since the subject's parkinsonian disorder diagnosis, if applicable (PDOTHRYR_#YRS); and the predominant domain first recognised as changed (FRSTCHG). NACCCOGF and FRSTCHG are the only variables also included in the top 10 features for diagnosing dementia, indicating a clear difference in the key features for diagnosing dementia

and differentiating between the subtypes. This is emphasised by the variable importances inferred using the dementia classifier (*diagnostic importances*) which are provided in figure 4.11 to the right of the variable descriptions. They also show that the vast majority of the variables which are important for both types of diagnosis pertain to cognition and, more generally, the nature of any changes the subject is experiencing. Furthermore, figure 4.12 has been included to demonstrate that the most important variables for diagnosing dementia and differentiating between the subtypes are different. In particular, it shows the differential importances for all 260 variables ordered according to their diagnostic importances.

Table B.1 also highlights the differences in diagnostic and differential importance. In addition, it indicates that most variables are of very little importance for differentiating between dementia subtypes, including the one constant and four synthetic variables discussed in section 4.1.3, as they were for diagnosing dementia.

4.3 Stacking Classifier

To determine whether the performance of the original dementia classifier could be improved upon (i.e. if more accurate diagnoses could be made), a stacking classifier was constructed to differentiate between subjects with and without dementia. Notably, a stacking classifier is a meta-classifier which, fundamentally, is trained using the outputs of a number of other classifiers. This section explains how the new classifier was built and details its performance. Two additional classifiers, one of which was constructed to aid in the analysis of the stacking classifier, are also discussed.

4.3.1 Construction

As shown in figure 4.13, the outputs of five classifiers, labelled C_{1-5}, were used to train the stacking classifier SC. These L_0 (level zero) classifiers had been trained to differentiate between a key dementia subtype (or 'other') and diagnoses indicating no dementia, in the same manner as the pairwise dementia subtype classifiers. Each classifier required their own unique training set, comprising subjects with a diagnosis

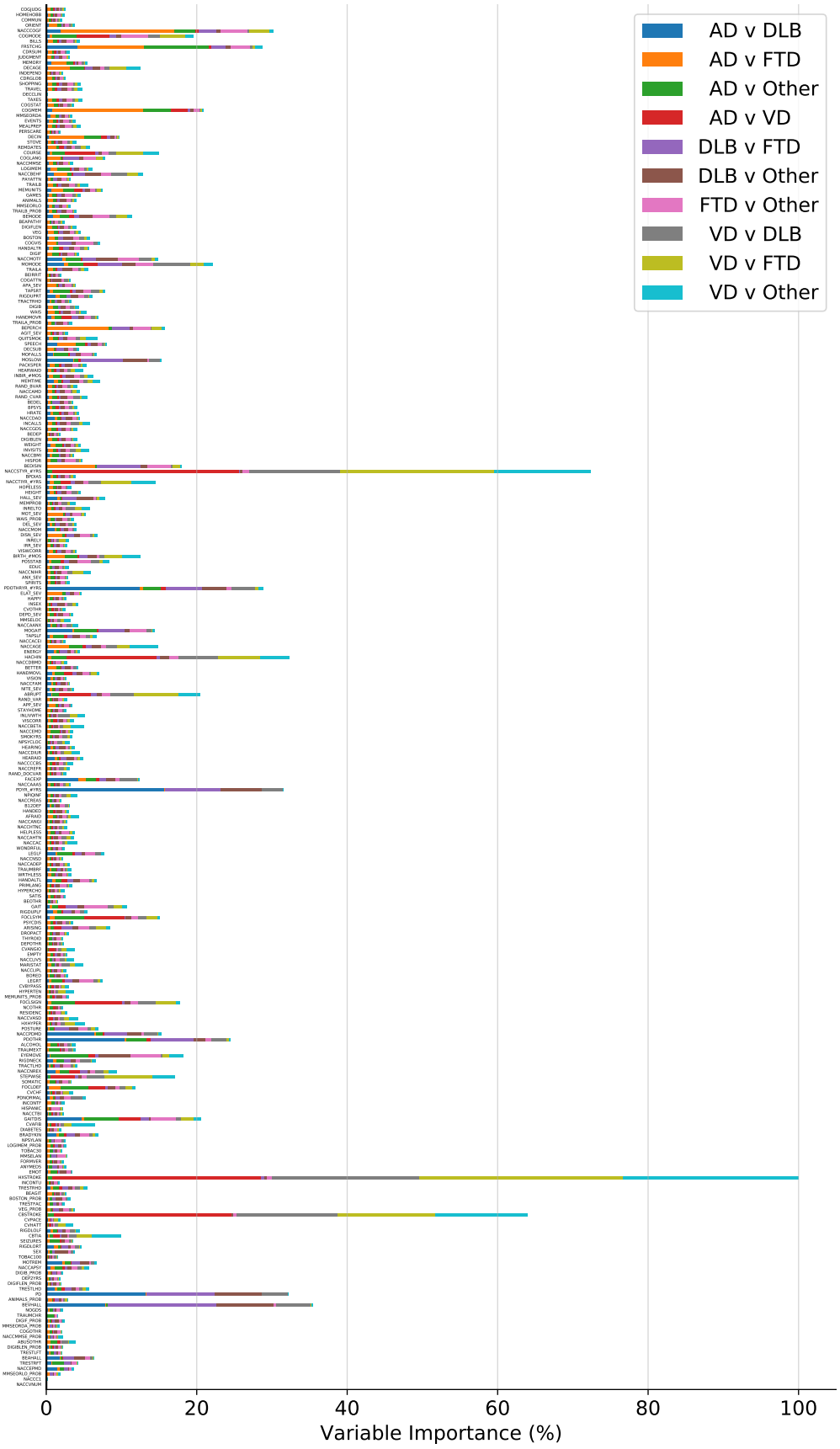


Figure 4.12: Differential importances for all 260 variables ordered according to their diagnostic importances. The most important variable for diagnosing dementia is at the top. The names of the variables are provided, but they are only readable if the plot is viewed electronically. Table B.1, however, provides the variable names in the same order.

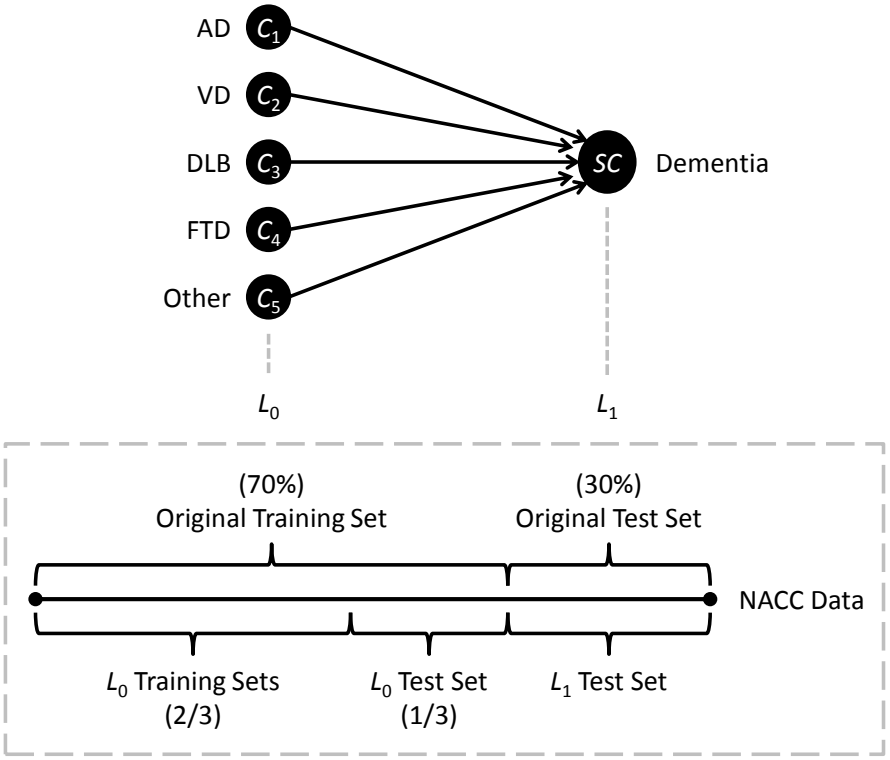


Figure 4.13: A simple diagram of the stacking classifier SC, showing that it was trained using the outputs of five subtype classifiers C_{1-5}. L_0 (level zero) and L_1 (level one) indicate the stage in the stacking process. A supplementary diagram, which provides an overview of how the NACC data was broken down into various training and test sets, is also provided.

of the subtype in question (including 'other') or without dementia, as well as a set of classification targets which, essentially, distinguished between cases of dementia and no dementia. The training sets were generated as described in section 4.2.1, by identifying the relevant subjects from two-thirds of the original training set ($N = 15,201$) in its imputed state. The same random sample (two-thirds) of subjects was used but primary cases of each subtype, namely the cases in which the subtype was the primary cause of the subject's cognitive impairment (i.e. dementia), were extracted instead of pure cases. Primary cases were utilised as these subtype classifiers were not being trained to differentiate between two subtypes, so mixed presentations were not an issue; this also meant more data could be used. Within the original training set there were 6,226 primary cases of AD, 161 of VD, 507 of DLB and 1,014 of FTD. There were also 592 of 'other', more specifically the cases in which a subject had been diagnosed with dementia but the primary cause of their cognitive

L_0 Classifier		Training Set			Test Set		
		DS	ND	Total	DS	ND	Total
AD	(C_1)	4141	9547	13688	2085	4754	6839
VD	(C_2)	123	9547	9670	38	4754	4792
DLB	(C_3)	340	9547	9887	167	4754	4921
FTD	(C_4)	676	9547	10223	338	4754	5092
Other	(C_5)	374	9547	9921	218	4754	4972

Table 4.6: Composition of the training (and test) sets for the L_0 (or subtype) classifiers. DS and ND are shorthand for dementia subtype and no dementia respectively. The subtypes considered were Alzheimer's disease (AD), vascular dementia (VD), dementia with Lewy bodies (DLB), frontotemporal dementia (FTD) and 'other'. The test sets described were only generated to ascertain how the classifiers performed in their own right.

impairment was not one of the four main subtypes.

Table 4.6 indicates the number of subjects in each of the five training sets, as well as those with a dementia subtype diagnosis (DS) and without dementia (ND). In addition, the table details five conditioned test sets, which were generated from the remaining one-third of the original training set ($N = 7,600$). These test sets were only created to determine how each of the subtype (L_0) classifiers performed in their own right, and were not pertinent to the construction of the stacking classifier. The individual performance of these classifiers is commented on in the next section.

Before the stacking or L_1 (level one) classifier could be constructed, the subtype classifiers each had to be employed for subjects comprised within a single test set. As indicated in figure 4.13, the L_0 test set amounted to the remaining one-third of the original training set. Of the 7,600 subjects, 2,846 had been diagnosed with dementia and 4,754 had not. The resultant sets of predictions (or outputs), which were converted into dementia probabilities and compiled into a data set, were used to train the stacking classifier in the usual way. As the L_0 classifiers had been tailored to different dementia subtypes, the dementia probabilities varied for the classifiers.

To enable the stacking classifier and original dementia classifier to be compared, a new version of the latter was built using the L_0 test set. A third classifier was also constructed with the purpose of differentiating between subjects with and without dementia. This classifier was trained using the L_0 test set and the dementia probabilities from the L_0 classifiers, to determine whether utilising the two sets of

Classifier	Accuracy % (no.)	Sensitivity TPR (no.)	Specificity TNR (no.)	FNR (no.)	FPR (no.)	AUC
Dementia	93.97 (9183)	0.92 (3363)	0.95 (5820)	0.08 (273)	0.05 (316)	0.99
Stacking	93.50 (9137)	0.92 (3349)	0.94 (5788)	0.08 (287)	0.06 (348)	0.98
Hybrid	94.12 (9197)	0.93 (3369)	0.95 (5828)	0.07 (267)	0.05 (308)	0.99

Table 4.7: Classification performance of the stacking and hybrid classifiers. The performance of the newly generated dementia classifier, which was trained with a subset of the subjects used for the original, is also detailed. Performance was measured in terms of the accuracy, sensitivity (true positive rate or TPR), specificity (true negative rate or TNR), false negative rate (FNR) and false positive rate (FPR) of the classifier, and the area under the receiver operating characteristic curve (AUC).

information in conjunction with one another could improve classification performance. As the classifier made use of the inputs to both the stacking classifier and the newly generated dementia classifier, it is referred to as the *hybrid classifier*. Analysis of the three classifiers' performance for the original test set ($N = 9,772$) is provided in the following section.

4.3.2 Classification Performance

As indicated in section 4.3.1, the stacking and hybrid classifiers, along with the newly generated dementia classifier, were assessed using the original test set ($N = 9,772$) in its imputed state. Of the 9,772 subjects, 3,636 had received a dementia diagnosis and 6,136 had not. Sets of predictions were easily ascertained from the dementia classifier, by simply passing the observations through each tree in the ensemble, but a precursory step was required for the stacking and hybrid classifiers. In particular, the observations had to be classified using the subtype (L_0) classifiers, so that dementia probabilities (or ensemble scores) could be generated to serve as input. Once sets of predictions had been obtained from each of the three classifiers, the ensemble votes (predicted classes) and scores could be determined. The predicted classes were used, along with the true classes, to calculate the accuracy, sensitivity, specificity, FNR (1 - sensitivity) and FPR (1 - specificity) of the classifiers (table 4.7). ROC curves were also produced (figure 4.14) using the ensemble scores and true classes, and their AUCs were computed (table 4.7).

Table 4.7 shows the new dementia classifier's performance is consistent with that of the original. To reiterate, the original dementia classifier had an accuracy of 94.21%, a sensitivity of 0.93, a specificity of 0.95, a FNR of 0.07, a FPR of 0.05 and an AUC of 0.99. The table also indicates that the stacking and hybrid classifiers perform very well; and suggests all three classifiers have comparable performance, which the ROC curves in figure 4.14 confirm. As a result, the stacking and hybrid classifiers fail to improve upon the performance of the dementia classifier.

It was stated in section 4.3.1 that conditioned test sets were created to assess the individual performance of the subtype (L_0) classifiers. Figure 4.15 presents the ROC curves generated. In particular, it shows that all five of the classifiers are able to perform well, although it highlights that the Other classifier (purple) is the worst performer. This is unsurprising due to the varied nature of the 'other' subjects.

4.4 Discussion

The purpose of this section is to put the work recounted over the course of this chapter into context. It does so by providing a relatively concise discussion of related work, focusing on the intersection of machine learning and dementia literature. It also briefly explores the clinical implications of the findings, which were identified, in part, by my clinical supervisors.

4.4.1 Related Work

Machine learning has been employed for a wide range of medical applications. Whilst reviewing the dementia literature it became apparent that machine learning has predominantly been utilised in conjunction with neuroimaging, largely for the purposes of diagnosing Alzheimer's disease. Pellegrini et al. (2018), along with Ahmed et al. (2019), provide a comprehensive review of machine learning approaches which make use of neuroimaging. Support vector machines are the most commonly used classifiers, as highlighted by Sørensen, Nielsen and Alzheimer's Disease Neuroimaging Initiative (2018), but ensembles of classifiers and random forests are becoming more popular; this is emphasised by the work of Sarica, Cerasa and Quattrone (2017).

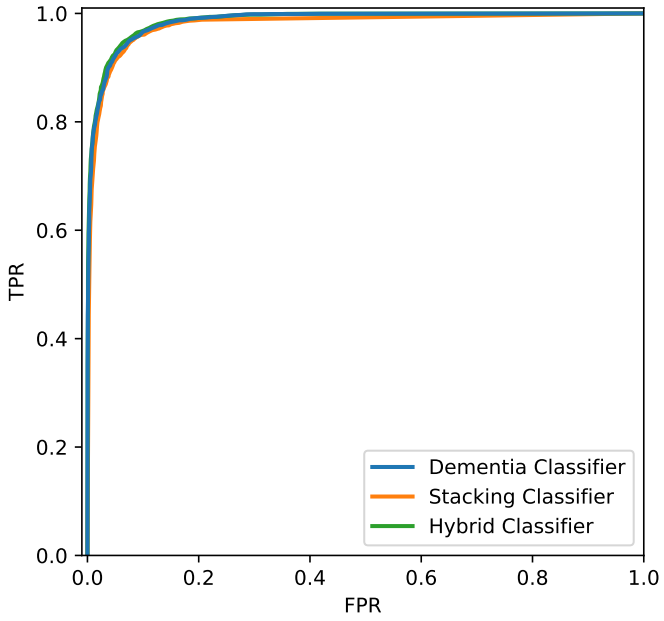


Figure 4.14: Receiver operating characteristic (ROC) curves for the stacking and hybrid classifiers. The ROC curve for the newly generated dementia classifier is also shown.

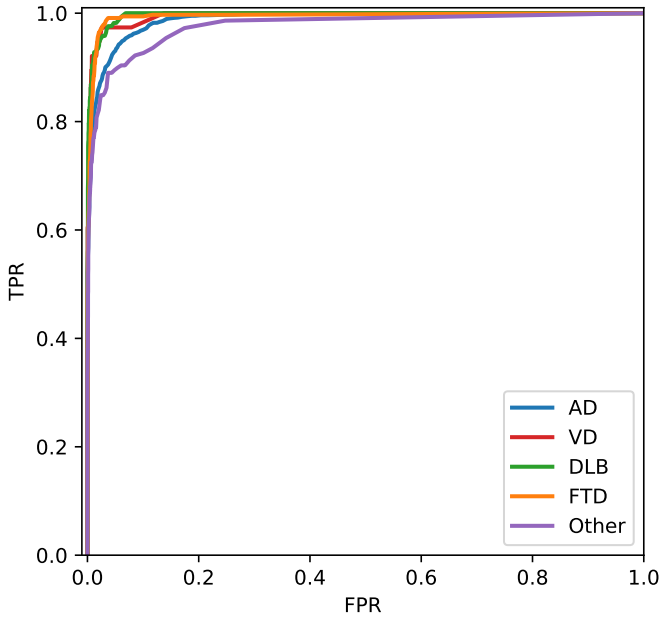


Figure 4.15: Receiver operating characteristic (ROC) curves for the L_0 (subtype) classifiers.

Pellegrini et al. (2018) highlight the need to diversify the data used. Jammeh et al. (2018) did just that by utilising primary care clinical data routinely collected over the course of two years at GP (general practitioner) surgeries in Devon. The aim of their study was to develop a tool which could identify people that may have dementia who had not received a formal diagnosis. In particular, 3,063 patients, all above the age of 65, were used to develop a naïve Bayes classifier to differentiate between patients with and without dementia. Of the 3,063 patients, 850 were thought to have dementia and 2,213 were deemed healthy (or cognitively normal). As all the variables were binary and simply highlighted the patients for which a specific risk factor, symptom or behaviour had been noted, there was no missing data. A combination of manual and automated feature selection was employed, prior to the construction of the classifier, in order to identify the subset of variables considered to be indicative of dementia. Notably, the variables used to identify patients with and without dementia were not included. 10-fold cross-validation was used to assess the performance of the classifier; and it achieved an accuracy of 86.06%, a sensitivity of 0.84, a specificity of 0.87 and an AUC of 0.87. The dementia classifier (section 4.1), in comparison, had an accuracy of 94.21%, a sensitivity of 0.93, a specificity of 0.95 and an AUC of 0.99.

Research with the aim of differentiating between subtypes of dementia using machine learning has been limited to date, but Jarrold et al. (2014) and Dauwan et al. (2016) have conducted work of this nature. In fact, Dauwan et al. (2016) utilised a random forest to differentiate between dementia with Lewy bodies (DLB) and Alzheimer's disease (AD). The classifier, which was built using the Random Forests algorithm (section 1.3.2) and consisted of 500 trees, was developed with 66 DLB and 66 AD patients from the Amsterdam Dementia Cohort (van der Flier et al., 2014). In particular, features were manually selected according to their availability, as well as their presence in the diagnostic criteria of the subtypes. 61 features of continuous and categorical type were used, which encapsulated electroencephalography (EEG) (47 features), clinical (2 features) and neuropsychological (5 features) data, along with information regarding neuroimaging and cerebrospinal fluid (CSF) biomarkers

(7 features). Missing data was dealt with by discarding the features which had greater than (or equal to) 33% of its values missing, and any remaining missing values were imputed with the average for the corresponding feature. The authors note it can be difficult to determine a meaningful average for a categorical feature. The resultant classifier had an accuracy of 87%, a sensitivity of 0.88 and a specificity of 0.86. Interestingly, a classifier was also developed using the clinical and neuropsychological data alone; it achieved an accuracy of 66%, a sensitivity of 0.65 and a specificity of 0.67. The latter classifier is more comparable to the AD v DLB pairwise dementia subtype classifier (section 4.2), which had an accuracy of 97%, a sensitivity of 1.00 and a specificity of 0.63.

Sarica, Cerasa and Quattrone (2017), along with Dauwan et al. (2016), discuss the fact that variable importances can be ascertained using random forests. They also highlight their significance for the domain, particularly in identifying features that are important for diagnosis. In fact, Dauwan et al. (2016) specifically highlight the importance of EEG features for differentiating between DLB and AD. EEG data, however, is not routinely collected in clinical practice for the purposes of diagnosing dementia. Furthermore, a modest amount of work has been carried out with the intention of identifying a small number of features that could be used to diagnose cognitive impairment to varying degrees. The work of Weakley et al. (2015) and Chiu et al. (2019) serve as examples, but neither of the studies made use of a large number of variables or random forests.

Machine learning has also been used to predict the prognosis of dementia or, in simpler terms, the likelihood that someone will develop dementia. Dallora et al. (2017) present a detailed review of machine learning approaches tailored for this purpose, highlighting that, once again, a significant proportion of research has focused on neuroimaging and Alzheimer's disease. Ritchie and Tuokko (2011) and Maroco et al. (2011) describe studies which predominantly aimed to predict the incidence of dementia using clinical and neuropsychological data, or just the latter. Maroco et al. (2011), in particular, conducted a comparison of a range of techniques and found that random forests, along with linear discriminant analysis, was the most

successful approach. Additionally, a study which utilised data from the NACC UDS in order to predict the incidence of mild cognitive impairment (MCI) is presented by Lin et al. (2018).

4.4.2 Clinical Implications

The findings presented over the course of this chapter have the potential to transform the way in which dementia is diagnosed. This is due to the lack of clarity in the prevailing diagnostic criteria, notably important assessments that facilitate an accurate diagnosis are absent; and the fact that it is currently difficult and time consuming to diagnose dementia reliably. The key features, which have been identified for the purposes of diagnosing dementia and differentiating between the main subtypes, could prove useful in redesigning and streamlining routine clinical practice, particularly in specialised centres; thus, reducing the time required to make a diagnosis, along with the associated costs. They may also help to reduce the variability in the diagnosis of dementia, as well as to improve the quality of diagnoses. In addition, there is the potential to further enhance clinical practice with a diagnostic aid, developed from the various classifiers constructed. It would, however, require appropriate validation (in a clinical trial, for example) and regulatory approval.

The research conducted is valuable, partly due to the size of the data set utilised, in terms of both subjects and features. It is also advantageous that the subjects were assessed at a number of different specialised centres, the features encompassed a wide range of clinical and neuropsychological data, and missingness was handled in what could be considered a robust manner. Nevertheless, there are inevitably limitations to the study. In fact, the degree to which the findings are applicable to other settings, such as low- and middle-income countries and less specialised diagnostic environments, is unknown. However, it may be possible to draw on what has been learnt via transfer learning (Pan and Yang, 2010). Furthermore, the true accuracy of each classifier could be marginally different to the accuracy provided here, as the diagnoses in the NACC UDS may have been subject to error. The classifiers also work under the assumption that the criteria used to make these diagnoses

accurately reflect the nature of dementia and its subtypes, which is investigated to some extent in the following chapter.

In addition to the avenues for future research suggested earlier in the chapter, such as predicting change in cognitive impairment, there are two others to mention. Firstly, it could be informative to investigate the effects of varying the amount of training observations, particularly on accuracy and the number of important features, not only for the dementia classifier but also the pairwise dementia subtype classifiers. It is probable that more subjects diagnosed with certain subtypes, including 'other', would be useful. Secondly, it may be beneficial to look into multi-label classifiers, which can associate an observation with more than one class (Tsoumakas and Katakis, 2007), for those subjects diagnosed with more than one dementia subtype (i.e. mixed dementia (National Institute on Aging, 2017)).

4.5 Summary

A dementia classifier was built using 22,801 subjects from the NACC UDS, specifically to identify people with and without dementia; it was subsequently tested using 9,772 subjects, also from the NACC UDS. For this test set, the classifier achieved an accuracy of 94.21%, a sensitivity of 0.93 and a specificity of 0.95. Moreover, the area under the receiver operating characteristic curve (AUC) was 0.99. In short, these results suggest that machine learning could be a useful tool for diagnosing dementia.

Using the proximity (similarity) matrix generated in accordance with the dementia classifier, spectral seriation was employed to gain an understanding of the subjects in terms of their similarity. It arranged similar subjects together and two distinct groups were formed. On the whole, these groups separated the subjects with dementia or mild cognitive impairment (MCI) from those with normal cognition, indicating that the subjects with dementia were more similar to each other than they were to those with normal cognition, and vice versa. In particular, the MCI subjects tended to be situated between those with normal cognition and dementia, meaning that all the subjects had essentially been arranged to form a spectrum of cognitive impairment. Based on these results, it may be more suitable to consider MCI as a

mild form of dementia rather than a condition in its own right.

The importance of each variable was inferred using the dementia classifier, to enable the key features for diagnosing dementia to be identified. The vast majority of the variables were of very little importance, but those found to be highly important were clinically relevant. The top two variables, in particular, indicated whether the subject was impaired in judgment, planning or problem-solving; and pertained to the home and hobbies category of the Clinical Dementia Rating (CDR).

An investigation was carried out to determine the number of variables required to match the performance of the dementia classifier, which used all 260 variables, taking their importance into account. Ultimately, the 42 most important variables were needed; and exactly half of them provided information concerning either the CDR, Functional Activities Questionnaire (FAQ) or Mini-Mental State Examination (MMSE). It is these 42 features that could be considered fundamental for the diagnosis of dementia, although just 19 may be sufficient if the small drop in performance is deemed inconsequential when the effort and resources required to collect the data encapsulated in the additional 23 variables are taken into account.

The importance of the CDR, FAQ and MMSE, as well as Form B9, was investigated further. Interestingly, the four assessments do not seem to be irreplaceable. Nonetheless, the CDR could be considered marginally more important than the other assessments and it may be unwise to overlook it.

In addition to the dementia classifier, 10 pairwise dementia subtype classifiers were constructed. They were trained to differentiate between two of the four key subtypes, or a key subtype and alternative dementia diagnoses ('other'), using subjects with pure cases of the relevant subtypes (including 'other'); these subjects were selected from two-thirds of the 22,801 previously utilised. Each classifier was subsequently tested using subjects from the remaining one-third, and the accuracies ranged from 80.77% (FTD v Other) to 99.40% (AD v VD). The sensitivities (subtype one) and specificities (subtype two) were also determined, and they ranged from 0.81 (VD v FTD and VD v Other) to 1.0 (AD v DLB, AD v Other and AD v VD) and 0.38 (AD v Other) to 1.0 (VD v FTD) respectively. Close inspection of the

sensitivities and specificities indicated there was some bias towards subtype one for a number of the classifiers, possibly resulting from class imbalance, but adjusting the classification thresholds could potentially improve performance. Incidentally, additional sets of subjects would be required to determine the optimal thresholds. Nonetheless, each classifier had a very high AUC, which could be considered a more reliable measure of performance. In fact, the AUCs ranged from 0.88 (AD v Other and FTD v Other) to 1.0 (VD v FTD), indicating machine learning could be used to differentiate between the key dementia subtypes.

Variable importances were ascertained for each pairwise dementia subtype classifier, and subsequently combined, to enable the key features for differentiating between the main subtypes of dementia to be identified. Most of the variables were of very little importance, once again, but those found to be highly important for specific subtypes generally corresponded with the current diagnostic criteria. The top three variables provided the subject's stroke history, along with the number of years since the subject's last stroke, if applicable; and almost exclusively acquired their importance from the VD (vascular dementia) classifiers. Interestingly, only two of the top 10 variables also featured in the top 10 for diagnosing dementia, indicating a clear difference between the important features for the two types of diagnosis.

Two additional classifiers were developed to determine whether the performance of the dementia classifier could be improved upon. The first was a stacking classifier, which was trained using the outputs of five other classifiers. These five classifiers were built to differentiate between a key dementia subtype (or 'other') and diagnoses indicating no dementia, specifically using primary cases of the subtypes. The second classifier was regarded as a hybrid classifier, as it was trained using the outputs of the same five classifiers, along with all 260 original features. Both of the classifiers were tested with the 9,772 subjects used to test the dementia classifier, and were found to perform very well. However, the performance of all the classifiers was comparable, meaning no improvement was made on the dementia classifier's performance.

On review of the dementia literature, it became apparent that machine learning has primarily been used with neuroimaging, mainly to diagnose Alzheimer's disease.

Support vector machines are typically employed, but ensembles of classifiers and random forests are gaining popularity. Even the benefit of being able to determine variable importances from a random forest has been noted by a modest number in the field. Researchers are aware of the need to diversify the data used, along with the applications; and work has been done on diagnosing dementia, in more general terms, and differentiating between subtypes. Nevertheless, very few studies could be considered comparable with the research that has been described.

In summary, these findings have the potential to transform the way in which dementia is diagnosed, despite there being limitations to the study. The key features identified, for both diagnosing dementia and differentiating between the main subtypes, could prove useful in redesigning and streamlining routine clinical practice. They may also help to improve dementia diagnosis, in more general terms, if the diagnostic criteria were updated accordingly. Furthermore, there is the potential to develop a diagnostic aid from the classifiers constructed.

Chapter 5

Clustering Mixed Data with Isolation Forests

In chapter 1 it was explained that one of the primary aims of the research was to gain an understanding of the inherent structure of dementia data, to ultimately investigate disease signatures (Stemmer et al., 2019). Clustering, which groups similar observations (or subjects) together in an unsupervised manner (i.e. without reference to any associated labels) to form clusters, was employed for this purpose. Notably, observations comprising a cluster are deemed to be similar to one another and dissimilar to those in other clusters. As Xu and Wunsch (2009) and Saxena et al. (2017) highlight, there are many different clustering methods which can broadly be categorised as hierarchical or partitional. Hierarchical methods group observations in one of two ways, either the complete set of observations is regarded as a cluster which is recursively partitioned (divisive hierarchical clustering) or each observation is considered as a cluster, all of which are merged into larger and larger clusters (agglomerative hierarchical clustering). Partitional methods, however, directly divide the observations into a number of clusters; it is this type of clustering that was used.

Regardless of the type of clustering employed, most algorithms make use of a proximity measure between observations, specifically a distance or similarity measure in this context (Xu and Wunsch, 2009). As indicated throughout the thesis, the data obtained from the National Alzheimer's Coordinating Center (NACC) contained variables (or features) of continuous, categorical, ordinal and binary type. Notably,

the latter two can be classed as categorical. Ahmad and Khan (2019) explain that clustering mixed data is challenging, primarily due to it being difficult to measure distance or similarity for variables with values that have no inherent order, and ensure measures for different variables are compatible and meaningful. Ultimately, an approach was developed that measures proximity by means of an isolation forest, essentially an ensemble of (unsupervised) decision trees that isolate each unique observation; thus, it can naturally draw on variables of mixed type. More specifically, proximity is based on the similarity of the paths taken by observations through each of the trees in an isolation forest. Despite the fact the approach was applied to NACC data, it is initially discussed without regard to it, following a brief discussion of related work predominantly on clustering categorical and mixed data. It was also tested on a variety of alternative data sets, for which results are provided later in the chapter, along with those for the NACC data.

5.1 Related Work

As James et al. (2017) explain, K-means clustering is probably one of the most well-known clustering methods. The algorithm was proposed by both Lloyd (1982) and Forgy (1965), and it is partitional in nature. In short, this simple method looks to minimise the variation within each cluster, of which there are K; and uses centroids to represent the clusters in order to do so. Notably, each centroid is the set of mean feature values for the observations in its cluster, and the clusters can be initialised by randomly assigning observations to them. After initialising the clusters and computing the centroids, two steps are iteratively performed until the clusters stabilise. Firstly, each observation is reassigned to the cluster with the closest centroid, which is typically identified using the (squared) Euclidean distance. Secondly, the centroids are updated to reflect the new cluster assignments. James et al. (2017) highlight that K-means finds a locally optimal clustering; thus, it is important to repeat the process a number of times with different cluster initialisations.

K-means clustering is unsuitable for categorical data, as Huang (1998) points out. However, Huang (1998) and Chaturvedi, Green and Carroll (2001) independently

repurposed K-means for categorical data, ultimately creating K-modes. Although the two K-modes algorithms differ, Huang and Ng (2003) show that they are equivalent. Huang (1998), specifically, substituted the mode for the mean, as well as the Hamming distance for the squared Euclidean distance, which could be considered an over-simplistic measure (Ahmad and Khan, 2019). In addition to K-modes, Huang (1998) proposed K-prototypes for clustering mixed data, which is essentially an amalgamation of K-means and K-modes.

A variety of alternative algorithms have been devised for the purposes of clustering categorical and mixed data, as evidenced by Elavarasi and Akilandeswari (2014), Ahmad and Khan (2019), Foss, Markatou and Ray (2019), Hendrickson (2014) and van de Velden, D'Enza and Markos (2019). Nonetheless, it is highlighted in the literature that it is common to use dummy coding (or one-hot encoding) for categorical variables in conjunction with a method such as K-means. In particular, dummy coding involves generating a binary variable for each category associated with a feature; this has the potential to significantly increase the dimensionality of the data, which could prove problematic (Keogh and Mueen, 2017). Foss, Markatou and Ray (2019) also explain that the values used for the binary variables, namely one or another scalar, can affect the contribution of the different types of variables to the clustering.

A few examples of algorithms for clustering categorical data are CACTUS (Ganti, Gehrke and Ramakrishnan, 1999), ROCK (Guha, Rastogi and Shim, 2000), COOLCAT (Barbará, Li and Couto, 2002), Squeezer (He, Xu and Deng, 2002), LIMBO (Andritsos et al., 2004) and AT-DC (Cesario, Manco and Ortale, 2007). Squeezer, in particular, is simple but efficient, as every observation is considered only once. In fact, the observations are considered in turn. The first forms its own cluster and then each observation is either added to an existing cluster or used to create a new one, depending on the similarity between the observation and every one of the existing clusters. In order to calculate the similarity, each of the observation's feature values is inspected and, essentially, the proportion of observations in the cluster with the same feature value is determined; these proportions are then summed. The

number of clusters is not specified at the outset, but a similarity threshold must be set. Notably, an observation is only added to a cluster, namely the cluster with the most similarity, if this threshold is exceeded. Once all the observations have been clustered, outliers are handled by discarding any very small clusters. He, Xu and Deng (2002) state that the algorithm is robust with regards to the order in which the observations are processed, but highlight that the similarity threshold can affect the quality of the clustering and the algorithm's execution time.

A significant portion of the research on clustering categorical and mixed data focuses on defining a new distance or similarity measure which can be put to use by an existing clustering method. For instance, Jia, Cheung and Liu (2016) proposed a new distance measure for categorical data that utilises the relative frequency of categorical values, along with the perceived importance of the variables and the correlation between variables. Boriah, Chandola and Kumar (2008), Alamuri, Surampudi and Negi (2014) and Elavarasi and Akilandeswari (2014) review a number of measures for categorical data, whilst Xu and Wunsch (2009) discuss measures for various types of features, as well as mixed data. Foss, Markatou and Ray (2019) also consider measures for mixed data, and highlight that Gower's distance (Gower, 1971) is popular. In short, Gower's distance is the weighted average over the set of features, where the distance between two categorical values for a variable is measured using the Hamming distance.

As previously stated, a clustering approach was developed which measures proximity by means of an isolation forest. Whilst re-examining related literature post hoc, it was discovered that Cortes (2019) had independently looked into using an isolation forest to measure proximity. In fact, Cortes (2019) proposed a distance measure based on the average separation depth across the trees in an isolation forest, which has been adapted to handle categorical variables and missing values. The isolation forest is similarly constructed, but proximity is measured differently. Crucially, the measure presented by Cortes (2019) works under the assumption that all observations are unique. Incidentally, unsupervised random forests have been used to measure proximity and perform clustering, which Shi and Horvath (2006)

discuss; an example of work of this nature is provided by Zhu, Loy and Gong (2014).

The clustering approach was applied to NACC data, in order to gain an understanding of its inherent structure. Ahlqvist et al. (2018) undertook similar research on diabetes data; and a limited number of related studies have been carried out for dementia, some of which are highlighted here. As a matter of fact, Viroli (2012) applied a factor mixture model to neuropsychological data from the Aging, Demographics, and Memory Study (ADAMS) (Langa et al., 2005) which largely clustered individuals according to cognitive impairment. Young et al. (2018), on the other hand, proposed the Subtype and Stage Inference (SuStaIn) model to group patients according to disease subtype, as well as stage of progression. It was tested on magnetic resonance imaging (MRI) data from the Genetic Frontotemporal dementia Initiative (GENFI) (The Genetic Frontotemporal dementia Initiative, 2020) and the Alzheimer's Disease Neuroimaging Initiative (ADNI) (Mueller et al., 2005), respectively; the objective was to identify subtypes of genetic frontotemporal dementia (FTD) and sporadic Alzheimer's disease (AD). In addition, Whitwell et al. (2009) investigated anatomical subtypes of the behavioural variant of FTD, specifically with differences in grey matter loss, whilst Mitelpunkt et al. (2020) and Qiu et al. (2018, 2019) conducted research into subtypes of AD. Notably, Qiu et al. (2018, 2019) utilised neuropsychological data from NACC. Finally, Cleret de Langavant, Bayen and Yaffe (2018) employed clustering to identify individuals which are highly likely to develop dementia, and Gamberger et al. (2017) investigated the prognosis of subjects with mild cognitive impairment (MCI) using a multilayered clustering algorithm that makes use of unsupervised random forests.

5.2 Clustering with Isolation Forests

To reiterate, a clustering approach was developed which measures proximity based on the similarity of the paths taken by observations through each tree of an isolation forest, primarily to exploit the nature of isolation forests and their intrinsic ability to handle mixed data. As a result, the approach can naturally draw on variables of mixed type, which is advantageous as clustering mixed data is challenging. In short,

Algorithm 8 Building an iTree (isolation tree)

1: **function build_iTree**(X)
2: **if** $|X| = 1$ or $X^i \; \forall i \leftarrow 1, \ldots, F$ constant **then**
3: **return** $\ell \leftarrow X$
4: **end if**
5: Randomly select an inconstant variable X^f
6: Randomly generate a split $S \triangleq \{X_L, X_R\}$ on X^f
7: $t_L \leftarrow$ **build_iTree**(X_L)
8: $t_R \leftarrow$ **build_iTree**(X_R)
9: Create η for S and attach t_L and t_R to form t
10: **return** t
11: **end function**

an isolation forest is constructed, which is ultimately used to measure the similarity of observations; and spectral clustering is applied to the matrix of similarities. Spectral clustering, which is similar to spectral seriation (section 4.1.2), is popular and simple to implement, as well as effective (von Luxburg, 2007). The remainder of this section details how an isolation forest is constructed, the various isolation forest proximity measures considered, and how spectral clustering is performed.

5.2.1 Isolation Forest

As explained, an isolation forest, which is essentially an ensemble of (unsupervised) decision trees that isolate each unique observation, is initially constructed. Notably, Liu, Ting and Zhou (2008) originally proposed building an isolation forest (or iForest) for the purposes of anomaly detection. The authors highlight that anomalies are few and far between, as well as different from normal instances; thus, they are more susceptible to isolation.

In chapter 1 it was explained that a decision tree consists of internal splitting and terminal nodes, along with edges. For each internal splitting node η of an isolation tree (or iTree) t, a variable X^f and a split S are randomly chosen (algorithm 8 lines 5–6). S partitions the data set X into X_L and X_R, which are the subsets of observations used to construct the left and right subtrees (t_L and t_R) respectively

(algorithm 8 lines 7–9). If at any point during the construction of t only a single observation remains or all the observations are equivalent, a terminal node ℓ is formed (algorithm 8 lines 2–4). It should be noted that if all the observations are equivalent, then all the variables will be constant (of which there are F). An isolation forest T is simply produced by building multiple (or M) trees in this manner.

Assuming X^f is free from missingness, S is based on a random cut-point or subset depending on the type of X^f, as for Extra-Trees (section 1.3.3). Missing (or conditionally missing) values can be handled, however. In fact, one of the three missingness incorporated in attributes (MIA) (section 3.1.3) splits is chosen at random if missing values are present; the sole exception being when there is just a single unique observed value, as the only option is to split X according to whether the value for X^f is missing or observed (S_{MIA_3}).

As indicated, there are similarities between the ways in which an isolation forest and a random forest (Extra-Trees algorithm) are constructed, but there are inevitably differences. For example, an isolation forest is built in an unsupervised manner, meaning classification targets are not required, whilst Extra-Trees is a supervised procedure. There are also differences between the isolation forest algorithm described and the original outlined by Liu, Ting and Zhou (2008). In particular, Liu, Ting and Zhou (2008) sample the data set prior to constructing each tree, which they claim improves anomaly detection; and limit the depth of the trees, as anomalies are likely to have relatively short paths. Furthermore, their algorithm does not deal with categorical variables and missing values.

There is scope to potentially improve the clustering approach by employing either an isolation forest with random rotations (rotated trees) or an extended isolation forest (Hariri, Kind and Brunner, 2018). In short, an isolation forest can only partition the feature space into axis-aligned hyperrectangles, whereas these alternatives are not restricted in this way. To construct an isolation forest with random rotations, the data is simply randomly rotated before each tree is built. The splitting procedure, however, is different for an extended isolation forest. In fact, a hyperplane with a random slope and intercept is chosen rather than a random

variable and split. Notably, these alternatives are unable to handle categorical variables; thus, a potential avenue for future research is to investigate how to extend them so that they can.

5.2.2 Isolation Forest Proximity Measures

Using the isolation forest, the similarity between observations in X can be ascertained. As for a random forest (section 3.2.3), the similarity of two observations is calculated based on their paths through the trees in the ensemble. The similarities are used to cluster the observations by means of an $N-\text{by}-N$ matrix P, where N is the number of observations. In fact, spectral clustering is applied to the similarity matrix, which ultimately results in a projection of the data being clustered using K-means.

As previously indicated, a number of (novel) isolation forest proximity measures were considered which enable the similarity between observations to be ascertained. The first to be discussed makes use of the Baire distance (Baire, 1909; de Bakker and de Vink, 1998), which is suitable for sequences and utilises the length of the common prefix. The path of an observation X_n through a tree t, denoted by ρ_n, can be represented using a binary string (i.e. a sequence), where each digit indicates whether the observation travelled left or right at each node. Thus, it follows that the proximity of two observations for a single tree t is measured using the Baire distance:

$$\bar{P}_t^{\,B}(X_i, X_j) = \begin{cases} 0 & \text{if } \rho_i = \rho_j\,, \\ 2^{-|\rho_i \cap \rho_j|} & \text{otherwise}\,. \end{cases} \tag{5.1}$$

Here, $|\rho_i \cap \rho_j|$ is the length of the common path (or prefix), which is equivalent to the depth of the last common node d_{ij}; and the proximity is zero if the paths of the observations are identical. It should be noted that a normalisation step is not required for the *Baire measure* due to the way in which proximity is defined for a single tree, hence $\bar{P}_t$ is used instead of P_t. In order to obtain the proximity of two observations for the isolation forest T, the average across the trees is calculated using

$$P(X_i, X_j) = \frac{1}{M} \sum_{t \in T} \bar{P}_t(X_i, X_j)\,, \tag{5.2}$$

where M is the size of T (i.e. the number of trees). Clearly, the result is a distance as opposed to a similarity, but it is possible to convert from the former to the latter by subtracting the value from one as $0 \leq P(X_i, X_j) \leq 1$. Incidentally, the conversion could be performed for each tree prior to averaging.

The two alternative measures considered also draw on the length of the common path, or depth of the last common node d_{ij}, for a tree. In fact, the proximity of two observations for a single tree is defined as

$$P_t^{MD}(X_i, X_j) = d_{ij} \tag{5.3}$$

for the *matching depth (MD) similarity measure* and

$$P_t^{QD}(X_i, X_j) = \frac{d_{ij}(d_{ij} + 1)}{2} \tag{5.4}$$

for the *quadratic depth (QD) similarity measure*; the latter is equivalent to the sum of the depths up to and including that of the last common node. As these values are similarities, they must be normalised to ensure $P_t(X_i, X_i) = 1 \; \forall i = 1, \ldots, N$, which is achieved using the following equation.

$$\bar{P}_t(X_i, X_j) = \frac{P_t(X_i, X_j)}{\sqrt{P_t(X_i, X_i) P_t(X_j, X_j)}} \tag{5.5}$$

In order to obtain the proximity of two observations for the isolation forest, the average across the trees is calculated using equation 5.2 for each measure.

Figure 5.1 shows how similarity would be assigned by the three proximity measures as a path of length 20 is traced through a tree. Crucially, it highlights the relative change in similarity as depth increases for each of the measures. In short, the change is constant for the matching depth similarity measure, whilst it decreases rapidly for the Baire measure and increases for the quadratic depth similarity measure. Notably, the local structure of the data is emphasised during clustering if the latter measure is employed, as a result of said increase. Consequently, the figure indicates that a variety of proximity measures were considered.

As Xu and Wunsch (2009) explain, a proximity measure must satisfy two conditions to be a similarity function, which are as follows:

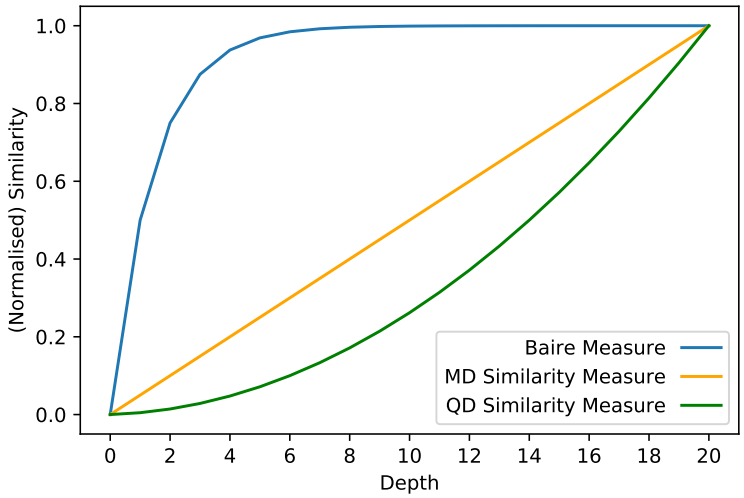


Figure 5.1: Similarity versus depth for the isolation forest proximity measures, namely the Baire measure, matching depth (MD) similarity measure and quadratic depth (QD) similarity measure.

$$P(X_i, X_j) = P(X_j, X_i) \quad \text{(symmetry)}\,, \tag{5.6a}$$

$$0 \leq P(X_i, X_j) \leq 1 \; \forall \; X_i, X_j \quad \text{(positivity)}\,. \tag{5.6b}$$

In order for it to be regarded as a similarity metric, it must also satisfy

$$P(X_i, X_j)P(X_j, X_k) \leq [P(X_i, X_j) + P(X_j, X_k)]P(X_i, X_k) \; \forall \; X_i, X_j, X_k\,, \tag{5.7a}$$

$$P(X_i, X_j) = 1 \text{ iff } X_i = X_j\,. \tag{5.7b}$$

The latter condition asserts that the proximity of two observations is equal to one if and only if the observations are identical, whilst the former is analogous to the triangle inequality. In brief, the three proximity measures considered are at least similarity functions, as the conditions described by equations 5.6a and 5.6b are satisfied by each of them. More work is required, however, to confirm whether or not they are similarity metrics, although the final condition (equation 5.7b) is satisfied by each measure. It should be noted that the Baire distance satisfies a strong form of the triangle inequality, among other conditions, making it an ultrametric (de Bakker and de Vink, 1998; Contreras and Murtagh, 2010), but it is unclear whether the Baire measure could be considered a similarity metric as a result.

5.2.3 Spectral Clustering

With the similarities in P, spectral clustering can be performed. As this is the final step of the approach, it produces a partition of X (i.e. clusters), from which it is possible to gain an understanding of the inherent structure of the data. This section provides a concise overview of spectral clustering and how it is performed, along with an explanation of K-means clustering, building on that given in section 5.1, as it plays a fundamental part. For an in-depth discussion of spectral clustering, the reader is directed to the work of von Luxburg (2007).

As von Luxburg (2007) explains, an undirected, weighted similarity graph, which consists of edges and vertices, can be constructed using the similarities; the problem of clustering can then be transformed into that of partitioning the graph. In general terms, spectral clustering achieves this by determining the Laplacian matrix (or graph Laplacian) of the weighted adjacency matrix of said graph, the latter matrix being positive and symmetric (as is P). It then projects the data onto the first K eigenvectors, assuming the eigenvalues are in ascending order, on which K-means clustering is performed.

A fully connected graph is used, thus the weighted adjacency matrix is simply the similarity matrix. It is common to employ a k-nearest neighbour graph, which emphasises the local structure of the data by only connecting observations (or data points) that are neighbours, but preliminary experiments suggested it may result in poorer performance for categorical and mixed data sets. Investigating its effects could be an interesting avenue for future research, however. The normalised Laplacian matrix related to a random walk, which is denoted by $\mathcal{L}_{rw}$, is also utilised on the recommendation of von Luxburg (2007). It is defined as

$$\mathcal{L}_{rw} = D^{-1}\mathcal{L} = I - D^{-1}P\,, \tag{5.8}$$

where D is the degree matrix (algorithm 9 line 2), $\mathcal{L}$ is the unnormalised Laplacian matrix (algorithm 9 line 3) and I is the identity matrix. Chung (1997) details properties of $\mathcal{L}_{rw}$, which von Luxburg (2007) highlights are pivotal to the success of spectral clustering, such as it is positive semi-definite.

Algorithm 9 Pseudocode for spectral clustering

1: **function spectral_clustering**(X,P,K,N)
2: $D \leftarrow \text{diag}(\sum_{j=1}^{N} P(X_i, X_j)\ \forall i \leftarrow 1, \ldots, N)$
3: $\mathcal{L} \leftarrow D - P$
4: Solve $\mathcal{L}u_i = \lambda_i D u_i\ \forall i \leftarrow 1, \ldots, K$ ▷ generalised eigenvalue problem
5: $U \leftarrow [u_1, \ldots, u_K]_{N,K}$ ▷ eigenvector matrix of size $N-\text{by}-K$
6: $\mathcal{P}_U \leftarrow$ **K-means**(U,K,N) where $\mathcal{P}_U \triangleq \{c_1, \ldots, c_K\}$
7: Convert $\mathcal{P}_U$ to $\mathcal{P}_X$ such that $c_i \leftarrow \{X_n\ \forall U_n \in c_i\}\ \forall i \leftarrow 1, \ldots, K$
8: **return** $\mathcal{P}_X$
9: **end function**

The most common spectral clustering algorithm for $\mathcal{L}_{rw}$ draws on the work of Shi and Malik (2000); this algorithm is employed. To begin with, the degree matrix, which has the sum of the similarities for every observation along the diagonal, and the unnormalised Laplacian matrix are computed (algorithm 9 lines 2–3). The first K (of N) generalised eigenvectors of $\mathcal{L}$ are then obtained by solving the generalised eigenvalue problem ($\mathcal{L}u_n = \lambda_n D u_n$); these are used as the columns of an $N-\text{by}-K$ matrix, denoted by U, which can be regarded as a projection of the data (algorithm 9 lines 4–5). Crucially, a generalised eigenvalue λ_n and generalised eigenvector u_n of $\mathcal{L}$ are an eigenvalue and eigenvector of $\mathcal{L}_{rw}$ (von Luxburg, 2007). K-means is subsequently applied to the rows of U (i.e. the projected data points), each denoted by U_n, which should be easier to cluster, and a partition $\mathcal{P}_U$ results (algorithm 9 line 6). It is possible to convert this partition into one pertaining to the original data, namely $\mathcal{P}_X$, by simply drawing on the indices of the projected data points in each cluster c_k (algorithm 9 lines 7–8).

K-means clustering was introduced in section 5.1. Its objective is to assign each data point X_n in a data set X to exactly one cluster, of which there are K, such that $\bigcup_{i=1}^{K} c_i = X$ and the variation within each cluster is minimised. Initially, the clusters are randomly assigned data points; they should each contain at least one (algorithm 10 line 2). The centroid μ_k of every cluster, namely the set of mean feature values, is then computed (algorithm 10 line 3). Subsequently, two steps are iteratively performed until the clusters converge, which essentially means there

Algorithm 10 Pseudocode for K-means clustering

1: **function K-means**(X,K,N)
2: Randomly assign X_i to one of K clusters $\{c_1, \ldots, c_K\}$ $\forall i \leftarrow 1, \ldots, N$
3: Compute the cluster centroids $\{\mu_1, \ldots, \mu_K\}$
such that $\mu_i \leftarrow \frac{1}{|c_i|} \sum_{X_n \in c_i} X_n$ $\forall i \leftarrow 1, \ldots, K$
4: **while** clusters not converged **do**
5: Update the cluster assignments such that $X_i \in c_k$ $\forall i \leftarrow 1, \ldots, N$
where $k \leftarrow \text{argmin}_{j \leftarrow 1,\ldots,K} ||X_i - \mu_j||_2^2$
6: Recompute the cluster centroids (see line 3)
7: **end while**
8: **return** $\mathcal{P} \leftarrow \{c_1, \ldots, c_K\}$
9: **end function**

is no change in cluster assignments (algorithm 10 line 4). During these two steps, each data point is reassigned to the cluster with the closest centroid in terms of the squared Euclidean distance, although other distances could be employed, and the cluster centroids are recomputed (algorithm 10 lines 5–6). Ultimately, a set of clusters, or partition $\mathcal{P}$, results (algorithm 10 line 8).

In section 5.1 it was explained that it is important to repeat the clustering process multiple times with different cluster initialisations, as K-means finds a locally optimal solution (i.e. partition). In order to select a partition from those produced, the within-cluster sum of squares (WCSS) is calculated, which is defined as

$$\text{WCSS} = \sum_{i=1}^{N} \min_{j=1,\ldots,K} ||X_i - \mu_j||_2^2 \,. \tag{5.9}$$

More simply, it is the sum of the squared Euclidean distances between the data points and their closest centroid. By minimising the WCSS, which is in fact locally minimised by K-means, the variation within each cluster is itself minimised.

5.3 Experiments

The clustering approach developed was tested on a variety of alternative data sets, which are detailed in section 5.3.1, to ascertain whether it could produce reasonable results. In particular, it was applied to every data set using each proximity measure

in turn, along with 1,000 trees, 100 repetitions of K-means and K (i.e. the number of clusters) set to what could be considered the true number of clusters. The same isolation forest was utilised for each of the proximity measures, enabling their results to be directly compared. The results for the quadratic depth similarity measure, which was found to be somewhat more successful than the other measures considered, are briefly discussed in section 5.3.2 but, to summarise, the approach produced reasonable clusters for all of the data sets.

As previously stated, the aim was to gain an understanding of the inherent structure of dementia data, to ultimately investigate disease signatures. Once the approach had been tested, a number of preliminary experiments were conducted on NACC data which built on the work discussed in chapter 4. More specifically, the approach was applied to subsets of NACC data focusing on cognitive status or the four main dementia subtypes, which are Alzheimer's disease (AD), vascular dementia (VD), dementia with Lewy bodies (DLB) and frontotemporal dementia (FTD). Following its success on alternative data sets, the approach was employed with the quadratic depth similarity measure using the parameters which had enabled reasonable clusters to be produced. Exactly how the subsets were generated is detailed in section 5.3.1 and the results of these preliminary experiments are discussed in section 5.3.2.

5.3.1 Data Sets

Six alternative (or trial) data sets of different types (continuous, categorical or mixed) and varying difficulty were considered. They are listed in table 5.1, along with the two subsets of NACC data; those which are two/three-dimensional are shown in figure 5.2 with what could be deemed their true classes (or clusters) indicated where available. It should be noted that four of these data sets are synthetic, whilst the remaining two comprise real-world data. This section describes all eight of the data sets in turn and how they were generated where relevant.

Old Faithful Data A version of the classic data set including 222 observations of the Old Faithful geyser in Yellowstone National Park (Wyoming, United States)

Data Set	Type	Variables (no.)	Observations (no.)
Old Faithful Data	Continuous	2	222
Two Rings Data	Continuous	2	400
Synthetic Categorical Data	Categorical	2	600
Synthetic Mixed Data	Mixed	2	200
Tube Data	Mixed	3	1000
Lymphography Data	Mixed	18	148
NACC Data - Cognitive Status	Mixed	50	2000
NACC Data - Dementia Subtypes	Mixed	39	396

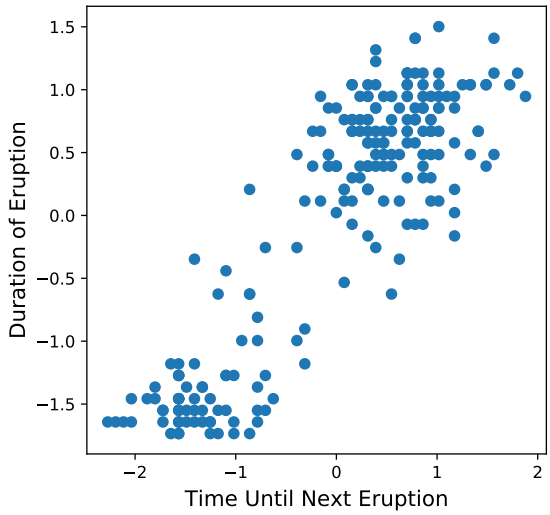
Table 5.1: Detailed list of data sets.

which were recorded in August 1978 and 1979 (Duke University, 2002). It has two continuous variables that provide the duration of the eruption and time until the next eruption (in minutes). The data, which is visualised in figure 5.2(a), was standardised prior to its use.

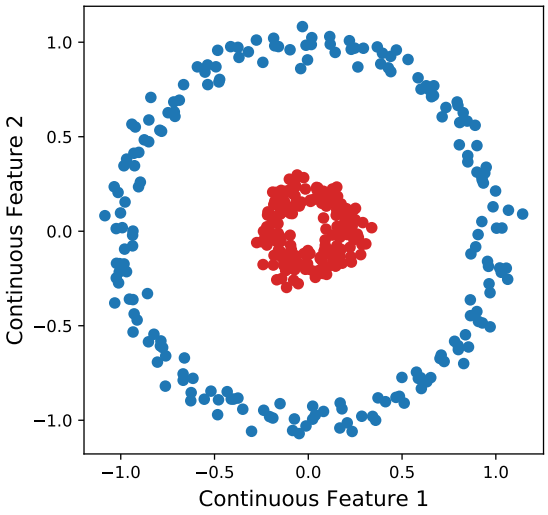
Two Rings Data A more challenging two-dimensional continuous data set which has proved popular for testing classification and clustering approaches; it consists of two rings, one of which is situated inside the other. This version, in particular, was generated using the `make_circles` function in scikit-learn (Pedregosa et al., 2011) and comprises 400 observations (or data points), namely 200 per ring. As figure 5.2(b) highlights, its true classes coincide with the two rings.

Synthetic Categorical Data This two-dimensional categorical data set is composed of 600 data points. Each variable has nine categories labelled with the integers in $[0, 8]$, which were grouped during the creation of the data set as figure 5.2(c) indicates. In fact, the points of three distinct three-by-three grids were randomly sampled in turn; the sample size was 200. Notably, these three grids correspond to the data set's three true classes and the category labels (or values) themselves are arbitrary, meaning they can be permuted without affecting the clustering.

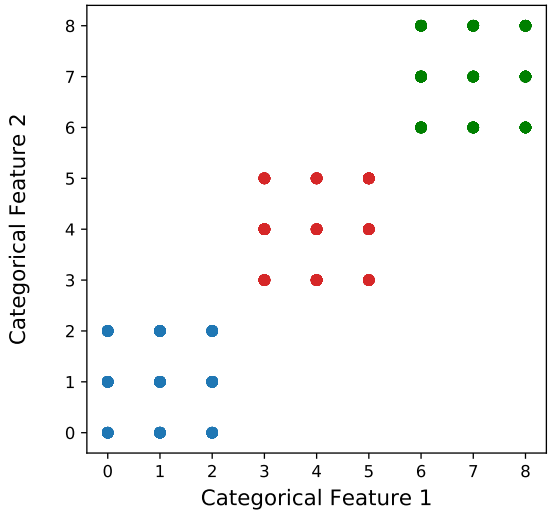
Synthetic Mixed Data A simple mixed data set with one continuous and one categorical variable including 200 data points. The continuous variable was generated by randomly sampling from two different normal distributions one by one with a sample size of 100. The categorical variable, on the other hand, was produced



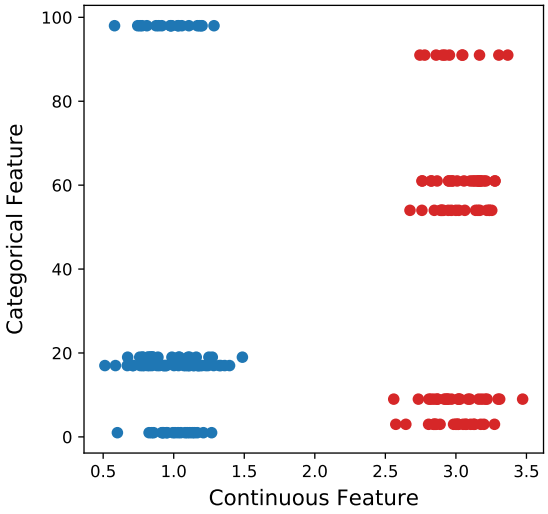
(a) Old Faithful Data



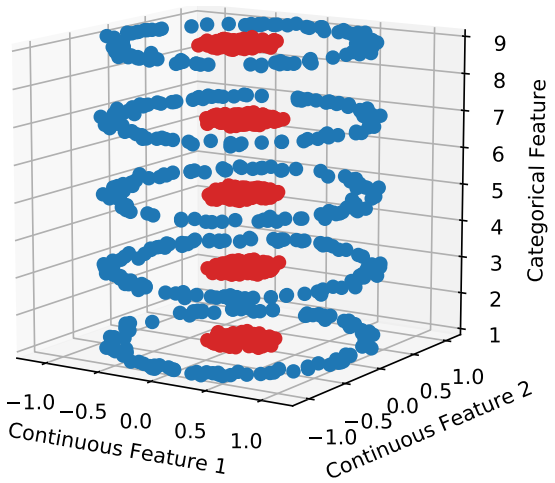
(b) Two Rings Data



(c) Synthetic Categorical Data



(d) Synthetic Mixed Data



(e) Tube Data

Figure 5.2: Two/three-dimensional alternative data sets with true classes indicated (using colour) where available.

by randomly selecting 10 categories in $[1, 100]$ with replacement and drawing two samples from a multinomial distribution, each of which represented the frequencies for five categories for 100 data points. It should be noted that the category values themselves were irrelevant and the categories had equal probability of occurring. Ultimately, the data set's two true classes were formed by pairing the samples from each of the normal distributions with five of the categories. Figure 5.2(d) presents the data and shows the categorical variable has nine unique categories.

Tube Data Essentially, this is an elongated version of the two rings data with two continuous variables, one categorical variable and 1,000 data points. It is a more testing mixed data set to successfully cluster into its two true classes, each of which comprise 500 data points, as its categorical variable does not aid in their separation (see figure 5.2(e)). Initially, the two continuous variables were generated using scikit-learn's `make_circles` function. The categorical variable was then created by randomly selecting five unique categories labelled with integers in $[1, 10]$ and drawing two samples from a multinomial distribution; these two samples represented the frequencies of the five categories for the 500 points of the inner and outer rings respectively. As for the synthetic mixed data, the category values themselves were irrelevant and the categories had equal probability of occurring.

Lymphography Data A data set encapsulating the findings of a medical imaging technique known as lymphography, which is used to visualise the lymphatic system. It was obtained from the UCI (University of California, Irvine) Machine Learning Repository (Dua and Graff, 2019), whilst the data itself was from the University Medical Centre, Institute of Oncology, Ljubljana, Yugoslavia (Zwitter et al., 1988). Notably, the data set includes 148 (complete) observations and 18 variables of mixed type. Every observation is also associated with one of four true classes, namely normal find, metastases, malign lymph and fibrosis; the number of observations in each class is 2, 81, 61 and 4 respectively.

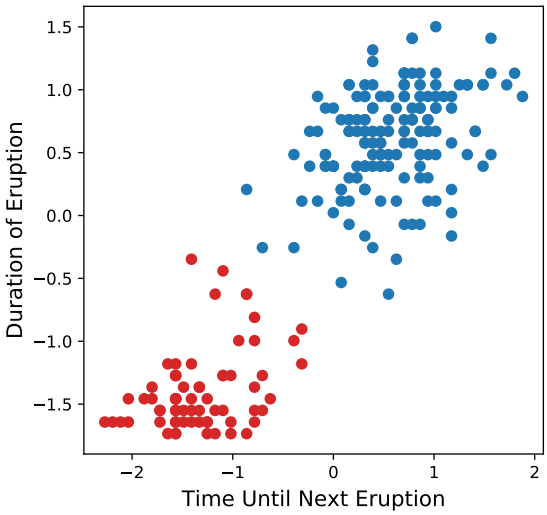
NACC Data - Cognitive Status A subset of NACC data which is focused on cognitive status. In short, the subset was extracted from the training set, comprising 22,801 randomly selected subjects and 260 variables, in which missing values had been

imputed using the approach outlined in chapter 3, namely proximity imputation with MIA; conditionally missing values were still present, however. The 50 most important features for diagnosing dementia, according to the dementia classifier, were included in the subset, building on the work discussed in chapter 4. The dimensionality of the data was reduced significantly because the majority of the 260 variables were found to be of very little importance for diagnosing dementia. A sample of 2,000 subjects was also used. These subjects were chosen such that 1,000 had been diagnosed with dementia and 1,000 had been diagnosed with either MCI (332) or normal cognition (668), although the subjects themselves were selected at random. Notably, these true classes may be subject to error, as explained in chapter 4.

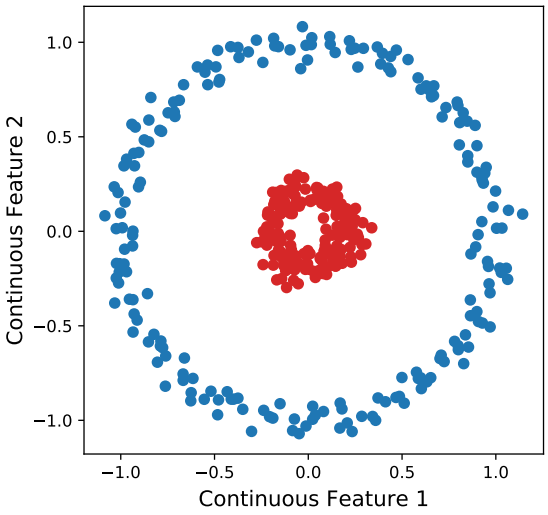
NACC Data - Dementia Subtypes This is a subset of NACC data which is focused on the four main dementia subtypes (AD, VD, DLB and FTD); it was extracted from the imputed training set, similarly to the cognitive status subset. Once again, the dimensionality of the data was reduced significantly, but this time by drawing on the fact that most of the 260 variables were found to be of very little importance for the differential diagnosis of dementia (see chapter 4). In particular, the 39 most important features for the differential diagnosis of dementia, according to the pairwise dementia subtype classifiers, were included in the subset, along with a sample of 396 subjects. This sample comprised subjects with 'pure' cases of the four main dementia subtypes, thus each subject had a primary diagnosis of one of the main subtypes and no supplementary diagnoses of any of the others. Ultimately, all those with a pure diagnosis of VD were included as there were only 96, whilst 100 subjects were chosen at random for the remaining subtypes (AD, DLB and FTD) so there was approximately an equal number of pure cases for each of the subtypes. Inevitably, these true classes may also be subject to error.

5.3.2 Results

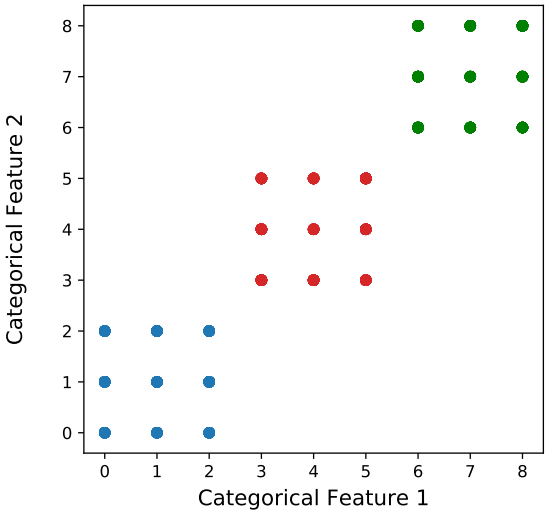
Figure 5.3 presents the results (i.e. clusters) for the two/three-dimensional alternative data sets. When considered in conjunction with figure 5.2, it is clear that three of the data sets were successfully clustered into their true classes, namely the two



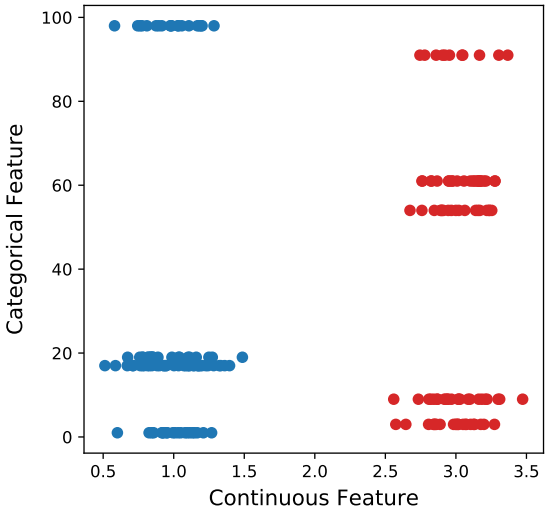
(a) Old Faithful Data



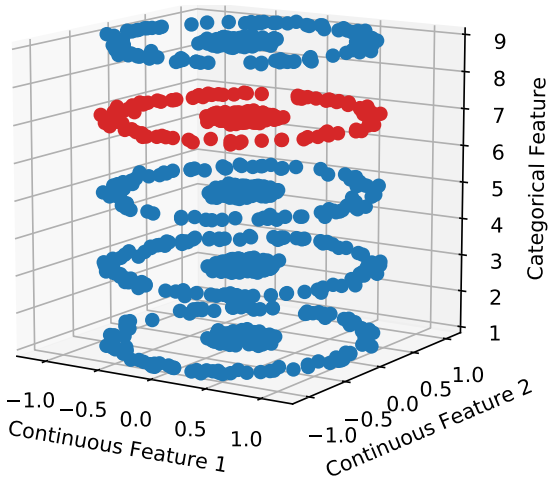
(b) Two Rings Data



(c) Synthetic Categorical Data



(d) Synthetic Mixed Data



(e) Tube Data

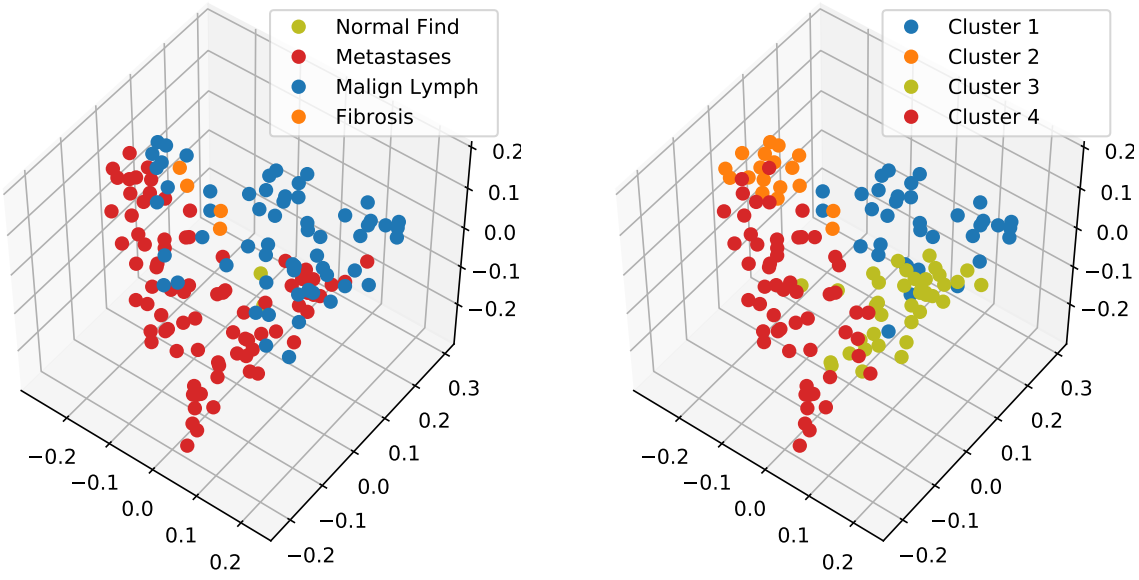
Figure 5.3: Clustering results for the two/three-dimensional alternative data sets.

rings, synthetic categorical and synthetic mixed data. This was verified with the normalised mutual information (NMI) (Strehl and Ghosh, 2002), which is a widely used external measure that evaluates the clusters (or partition $\mathcal{P}$) against the true classes Ω, disregarding any labels that have been assigned. It produces a value between zero and one, where the latter indicates perfect agreement between the clusters and true classes. The measure can be defined as follows:

$$\text{NMI}(\mathcal{P}, \Omega) = \frac{\text{MI}(\mathcal{P}, \Omega)}{mean(H(\mathcal{P}), H(\Omega))}\,, \tag{5.10}$$

where $\text{MI}(\cdot)$ is the mutual information, $H(\cdot)$ is the entropy and $mean(\cdot)$ is the arithmetic mean (see scikit-learn documentation (scikit-learn developers, 2020a) for more details). Incidentally, the Old Faithful data was also successfully clustered but no true classes were available.

It was highlighted in section 5.3.1 that the tube data is challenging as its categorical variable does not aid in the separation of its true classes, and the lymphography data has two very small true classes ($|\text{normal find}| = 2$, $|\text{fibrosis}| = 4$). In contrast to the four alternative data sets already mentioned, the clusters generated for these two do not match their respective true classes; they are not unreasonable, however. Figure 5.3(e) shows the clustering results for the tube data, which suggest that five clusters may be found to be more appropriate than two. Figure 5.4 shows the clusters, along with the true classes, for the lymphography data, which were visualised using metric multidimensional scaling (MDS) (Abdi, 2007). In short, MDS attempts to produce a low-dimensional embedding in which distances are preserved. Unlike the tube data, the clusters bear some resemblance to the true classes. In fact, those corresponding to the very small true classes are simply larger. As for the other alternative data sets, the NMI was calculated. It was approximately zero for the tube data and 0.26 for the lymphography data; this highlights that the NMI simply measures how closely the clusters match the true classes rather than indicating whether the clusters are reasonable or not. Ultimately, it could be concluded that the clustering approach developed can produce reasonable clusters for a variety of data sets.

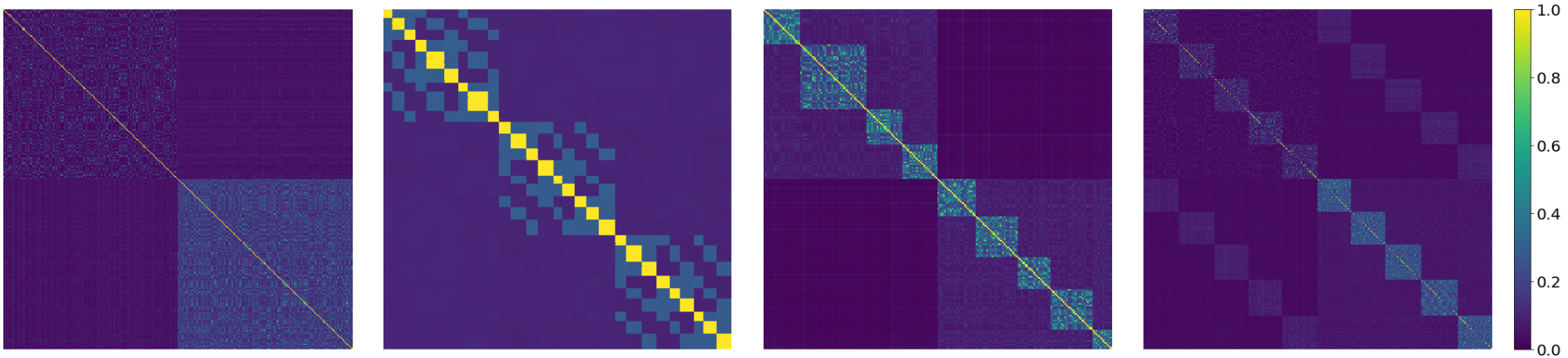


(a) True Classes

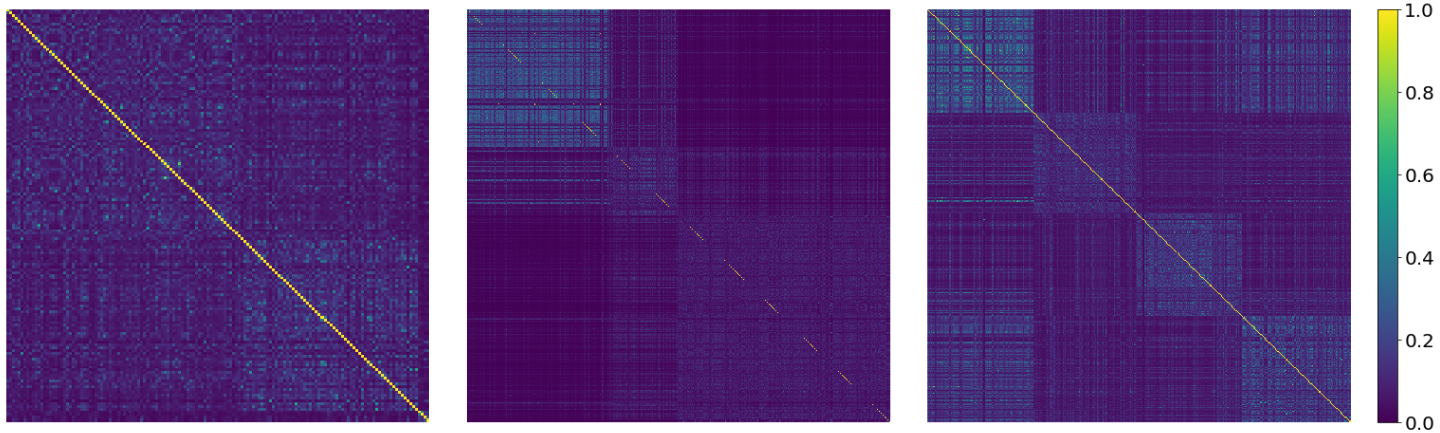
(b) Clusters

Figure 5.4: Clustering results for the lymphography data visualised using metric multi-dimensional scaling (MDS).

Figure 5.5 presents the similarity matrices for five of the six alternative data sets, as well as the two NACC subsets. The matrices are ordered such that the observations are grouped according to their true classes, and categories in some cases, hence the matrix for the Old Faithful data is not included. It should be noted that imperfections, such as the broken diagonal in figure 5.5(f), are artefacts of the plotting process. Figures 5.5(a), 5.5(b) and 5.5(c), in particular, show the similarity matrices for the three alternative data sets which were successfully clustered into their true classes and, unsurprisingly, there is little similarity between the classes. Reassuringly, the points comprising the inner ring of the two rings data are more similar to each other than those in the outer ring, similarity has been assigned in accordance with the number of matching categories for the synthetic categorical data, and there is approximately uniform similarity between categories for the synthetic mixed data. Figures 5.5(d) and 5.5(e) show the similarity matrices for the tube and lymphography data respectively, and provide some explanation as to the clusters produced. The former matrix appears to almost be an amalgamation of those for the two rings and synthetic mixed data, although there is some similarity between the classes for the categories; this similarity influenced the clustering. The latter, however, is unlike any



(a) Two Rings Data (b) Synthetic Categorical Data (c) Synthetic Mixed Data (d) Tube Data



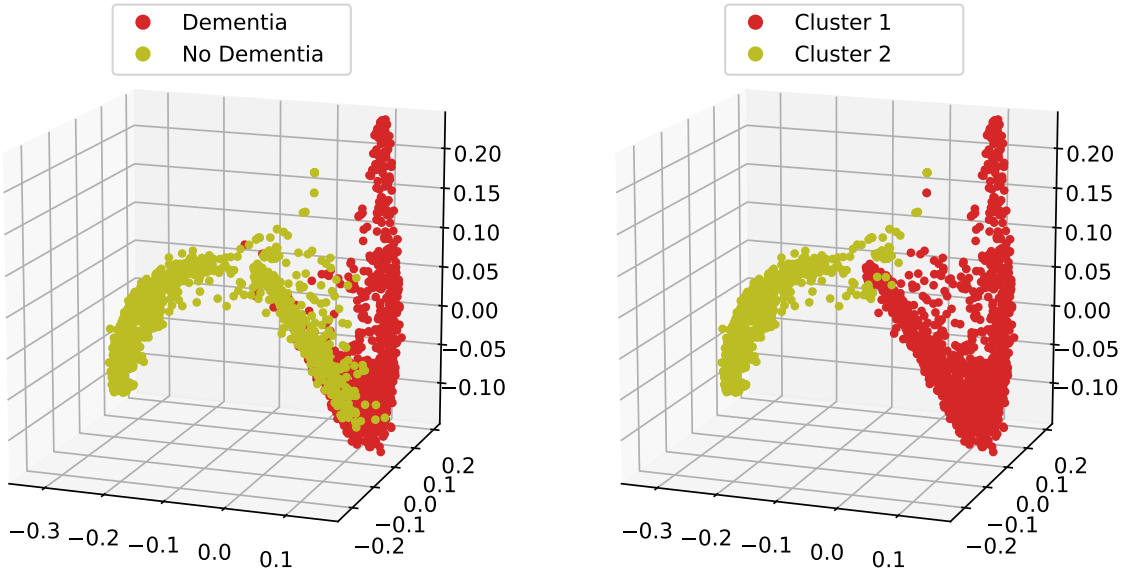
(e) Lymphography Data (f) NACC Data - Cognitive Status (g) NACC Data - Dementia Subtypes

Figure 5.5: Ordered similarity matrices for the data sets considered, excluding the Old Faithful data. It should be noted that imperfections, such as the broken diagonal in (f), are artefacts of the plotting process.

of the others. The matrix has some structure but it is considerably less pronounced.

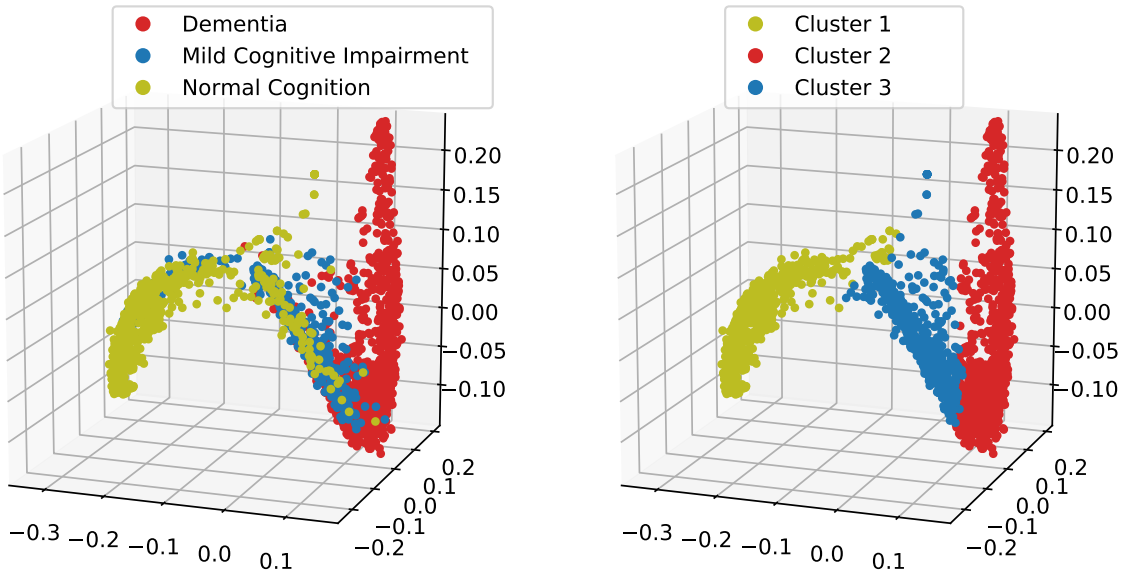
It was stated that figure 5.5 also presents the similarity matrices for the two NACC subsets. Figure 5.5(f), specifically, shows the matrix for the cognitive status subset which, from left to right, includes subjects diagnosed with normal cognition, MCI and dementia. Clearly, there are three blocks that correspond to the three categories of cognitive status, and relatively little similarity between the subjects with normal cognition and dementia. However, there is some similarity between the MCI subjects and those with normal cognition and dementia, suggesting the clusters may not simply corroborate NACC's diagnoses (i.e. the true classes). Furthermore, the matrix indicates that the cognitively normal subjects are more similar to each other than those with MCI and dementia. Figure 5.5(g), on the other hand, shows the matrix for the dementia subtypes subset, for which the subjects are ordered by subtype (from left to right: AD, VD, DLB and FTD). Once again, the matrix has structure, but there is some similarity between the subtypes, particularly AD and FTD. As a result, the clusters may differ somewhat from NACC's diagnoses. Interestingly, the subjects with AD are more similar to each other, as those with normal cognition are in the matrix for the cognitive status subset.

Figure 5.6 provides the clustering results for the cognitive status subset. Clustering was, in fact, performed with $K = 2$ and $K = 3$, where K is the number of clusters, as cognitive status can also be considered in terms of dementia and no dementia; the true classes (i.e. NACC's diagnoses) and clusters are shown for each case. MDS was used to visualise the true classes and clusters, as for the lymphography data, revealing some underlying structure consisting of two parts. Figure 5.6(c) shows that, essentially, one part encompasses the subjects diagnosed with normal cognition, whilst the other comprises those with MCI and dementia. Consequently, the no dementia class in figure 5.6(a), which consists of subjects with normal cognition and MCI, extends across the two parts. In short, the three clusters produced are relatively well-matched to their true classes (NMI = 0.54), but the two clusters are less so (NMI = 0.44). Nonetheless, both sets of clusters, along with the underlying structure, suggest that MCI may be a mild form of dementia as opposed to a clinical



(a) True Classes

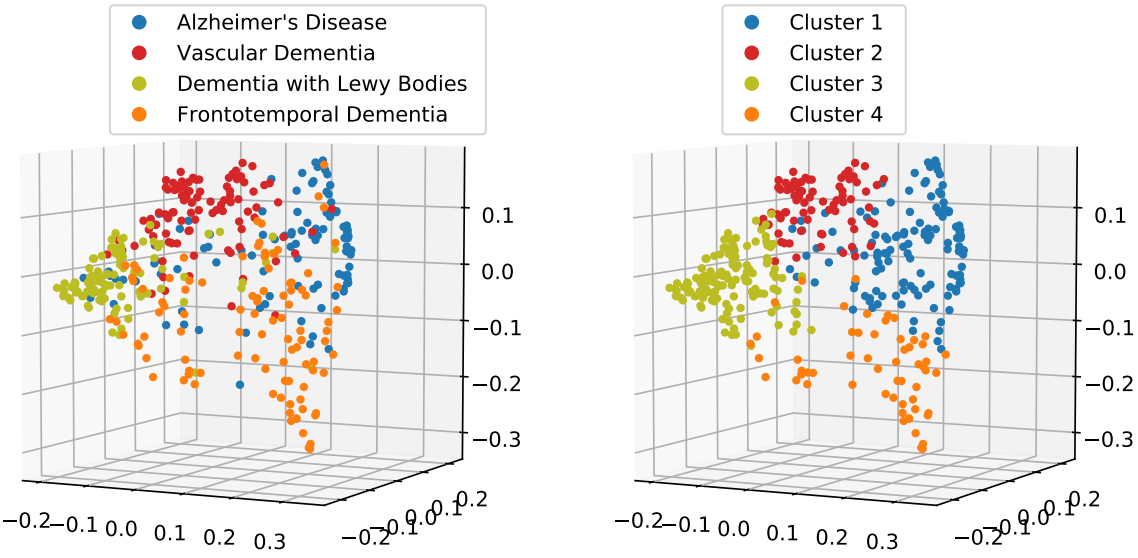
(b) Clusters



(c) True Classes

(d) Clusters

Figure 5.6: Clustering results for the subset of NACC data which is focused on cognitive status visualised using metric multidimensional scaling (MDS). Notably, two and three clusters are considered.



(a) True Classes

(b) Clusters

Figure 5.7: Clustering results for the subset of NACC data which is focused on the four main dementia subtypes visualised using metric multidimensional scaling (MDS).

entity (i.e. a condition in its own right), which it was explained in chapter 4 there is much debate over. The similarity matrix (figure 5.5(f)) also lends weight to this view, as it could be considered indicative of a spectrum of cognitive impairment.

Finally, figure 5.7 presents the true classes (i.e. NACC's diagnoses) and clusters for the dementia subtypes subset, which were visualised using MDS. Despite the NMI being 0.38, the figure indicates the four main subtypes (AD, VD, DLB and FTD) were basically recovered as clusters, suggesting there may be evidence for the current subtypes. In fact, the only real difference between the true classes and clusters appears to be the degree of overlap between the subtypes. The confusion matrix, which is presented in the form of table 5.2 and provides a more detailed view of how the subjects were clustered with respect to their true classes, does not dispute this. To ascertain whether this would be the case if all 260 variables or primary cases of the subtypes were utilised, two additional NACC subsets were experimented with. It should be noted that a primary case of a subtype is one in which the subject received a primary diagnosis of said subtype, and 150 subjects with primary diagnoses were randomly selected for each subtype. Ultimately, the clusters generated with all 260 variables bore some resemblance to the true classes,

		True Class				
		AD	VD	DLB	FTD	Total
Cluster	(1) AD	75	13	8	30	126
	(2) VD	3	69	2	5	79
	(3) DLB	18	11	88	16	133
	(4) FTD	4	3	2	49	58
	Total	100	96	100	100	396

Table 5.2: Confusion matrix for the subset of NACC data which is focused on the four main dementia subtypes.

but considerably less so as indicated by the NMI which was 0.15, whilst the subtypes were, once again, essentially recovered as clusters using primary rather than pure cases (NMI = 0.34).

An alternative external measure that evaluates the clusters (or partition $\mathcal{P}$) against the true classes Ω is the adjusted Rand index (ARI) (Hubert and Arabie, 1985), which was also used, albeit post hoc, as it could be considered preferable to the NMI due to it being corrected (or adjusted) for chance. As a result, the ARI is close to zero when observations have been randomly assigned to clusters. The measure can be defined as follows:

$$\mathrm{ARI}(\mathcal{P}, \Omega) = \frac{\mathrm{RI}(\mathcal{P}, \Omega) - E[\mathrm{RI}(\mathcal{P}, \Omega)]}{\max \mathrm{RI}(\mathcal{P}, \Omega) - E[\mathrm{RI}(\mathcal{P}, \Omega)]}, \tag{5.11}$$

where $\mathrm{RI}(\cdot)$ is the Rand index (Rand, 1971), $E[\mathrm{RI}(\mathcal{P}, \Omega)]$ is the expected value and $\max \mathrm{RI}(\mathcal{P}, \Omega)$ is the maximum value (see scikit-learn documentation (scikit-learn developers, 2020b) for more details). Once again, permuting any labels that have been allocated has no effect, and a value of one indicates perfect agreement between the clusters and true classes. In contrast to the NMI, negative values are possible, although not common in practice. Ultimately, the ARI was found to equal the NMI (to two decimal places) for all but two of the data sets, namely the lymphography data (ARI = 0.25, NMI = 0.26) and the cognitive status subset of NACC data ($K = 2$: ARI = 0.42, NMI = 0.44; $K = 3$: ARI = 0.6, NMI = 0.54), for which there was very little difference between the two measures.

To summarise, it could be concluded that the clustering approach developed

can produce reasonable clusters for a variety of data sets. Moreover, the preliminary experiments on NACC data suggested MCI may be a mild form of dementia as opposed to a clinical entity, and there could be evidence for the current dementia subtypes. With regards to future research, the main focus should be conducting more exploratory experiments to enable disease signatures to be investigated and a clinical conclusion to be drawn. For example, the clustering approach could be applied to all the data using a range of K. In addition, it may be worth considering substituting a fuzzy (or soft) clustering method for K-means, which permits an observation to belong to more than one cluster (Bezdek, 1981); this would allow for mixed presentations of subtypes.

5.3.3 Supplementary Investigation

As a precursor to the future research outlined at the end of the previous section, a supplementary investigation was conducted with the aim of revealing potential sub-subtypes of dementia. To be brief, agglomerative hierarchical clustering was carried out on the dementia subtypes subset of NACC data (see section 5.3.1 for details), enabling a dendrogram to be produced; this type of tree diagram, which is used to visualise the hierarchy of clusters, can prove useful in understanding the structure of data.

Agglomerative hierarchical clustering initially considers each observation as a cluster and repeatedly merges pairs of clusters until one remains. In particular, clusters are merged such that a linkage criterion, which draws on the distances between the observations, is minimised. For this investigation, the similarity matrix populated using the quadratic depth similarity measure (visualised in figure 5.5) was converted to a distance matrix and utilised in conjunction with average linkage (Sokal and Michener, 1958). Single and complete linkage were considered but the former failed to find meaningful clusters and the latter showed no improvement on that which was used. For explanations of the various linkage criteria, the reader is referred to the work of Everitt et al. (2011).

Figure 5.8 shows the resultant dendrogram; the dementia subtype (AD, VD,

DLB or FTD) diagnosed for each of the 396 subjects in the data set is indicated below the relevant edge (or branch). A truncated version of the dendrogram, which is easier to interpret, is also presented in figure 5.9. Notably, this version of the dendrogram is annotated with subtype frequencies (x = 0), as well as coloured boxes that indicate the predominant subtype for three major clusters (from left to right: VD, AD and FTD, DLB). These three clusters, along with the cluster situated on the far left comprising only 10 subjects, can be compared to the four found using spectral clustering (visualised in figure 5.7). As previously discussed, spectral clustering basically recovered the four main subtypes, whereas agglomerative hierarchical clustering had difficulty differentiating between AD and FTD. This outcome was not unexpected due to the fact that there was some similarity between these subtypes in the similarity matrix, as explained in section 5.3.2. Regardless, the dendrogram, in either form, does not clearly highlight potential sub-subtypes of dementia.

5.4 Summary

One of the primary aims of the research was to gain an understanding of the inherent structure of dementia data, specifically (mixed) data obtained from the National Alzheimer's Coordinating Center (NACC), to ultimately investigate disease signatures; clustering was employed for this purpose. Most clustering algorithms make use of a proximity (distance or similarity) measure between observations, but measuring proximity appropriately when the data is of mixed type is difficult. Consequently, clustering mixed data, in general, is challenging. Ultimately, an approach was developed that measures proximity by means of an isolation forest, so it is able to naturally draw on variables of mixed type.

On review of related work predominantly concerning clustering categorical and mixed data, it became apparent that a variety of algorithms have been devised, including K-modes (categorical), K-prototypes (mixed) and Squeezer (categorical). Nevertheless, a significant portion of the research focuses on defining a new proximity measure which can be put to use by an existing clustering method. At the time of

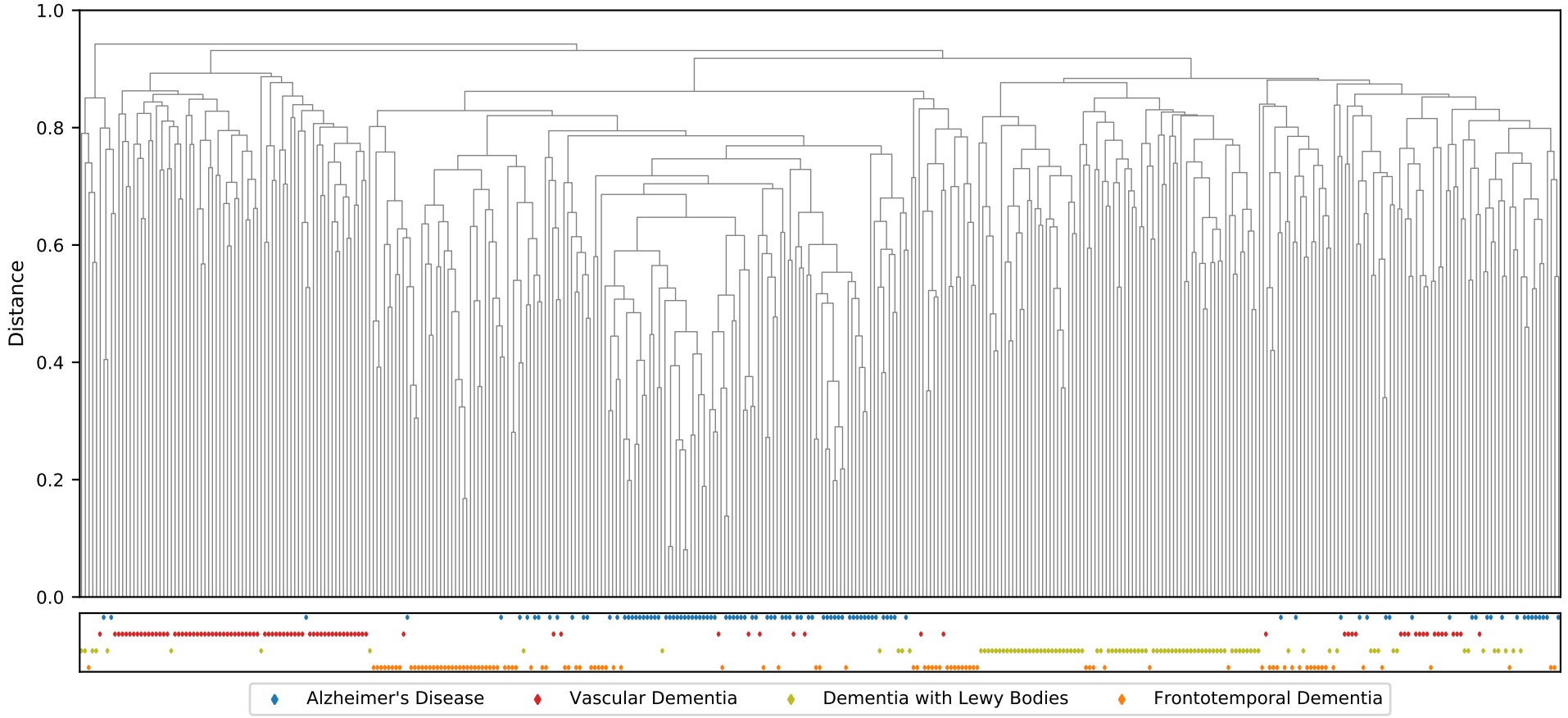


Figure 5.8: Dendrogram for the subset of NACC data which is focused on the four main dementia subtypes.

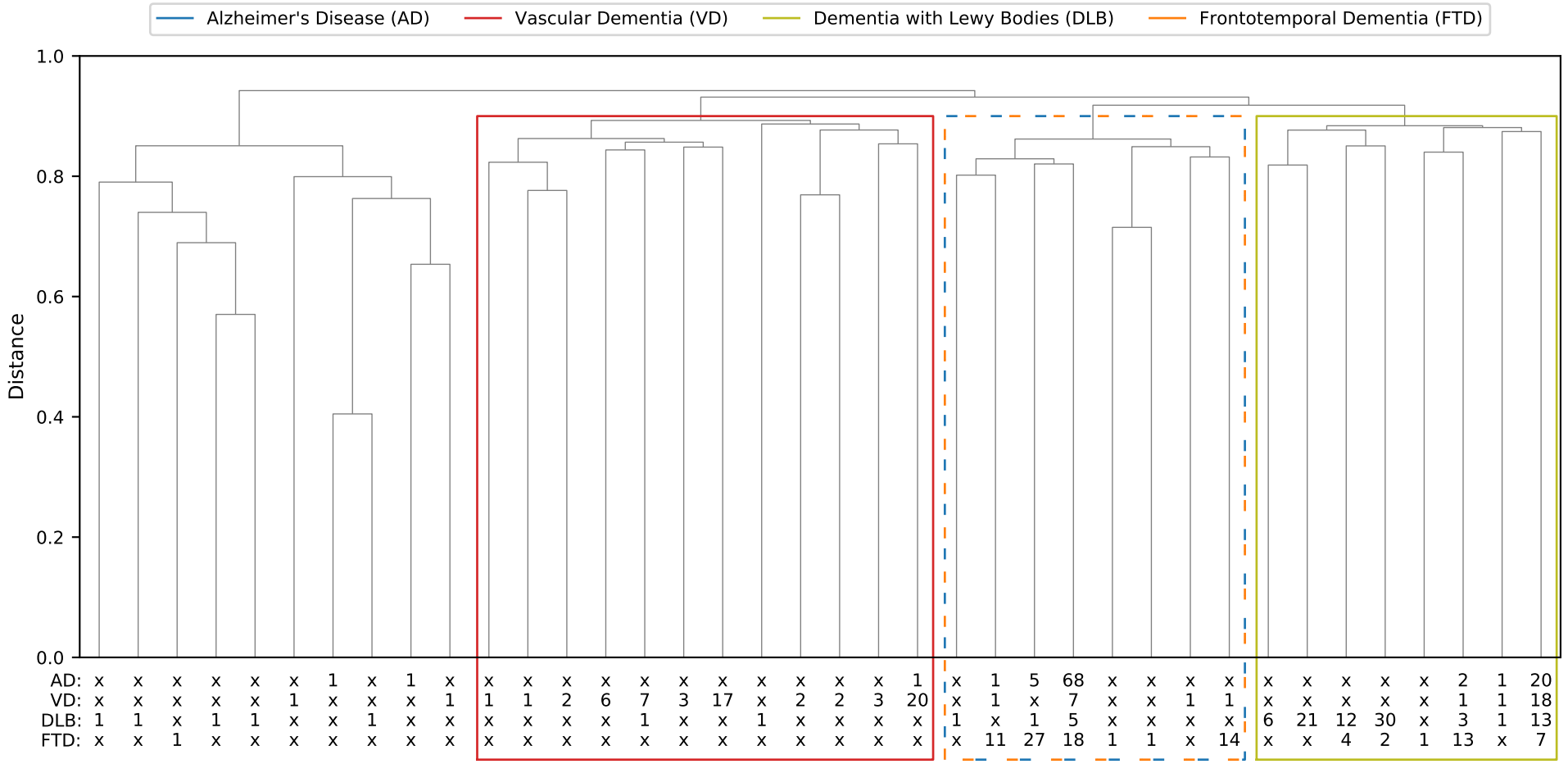


Figure 5.9: Truncated version of figure 5.8. Subtype frequencies are provided for each edge (or branch) of the dendrogram, and the predominant subtype is indicated by means of a coloured box for three major clusters.

writing, it was discovered another researcher had independently looked into using an isolation forest to measure proximity. The distance measure they proposed, however, works under the assumption that all observations are unique, which is less than ideal. Furthermore, related studies on dementia data were reviewed. In brief, it was revealed there are a limited number which primarily focus on identifying subtypes of Alzheimer's disease and frontotemporal dementia.

The clustering approach developed specifically measures proximity based on the similarity of the paths taken by observations through each tree of an isolation forest. Initially, an isolation forest is constructed in a manner that enables categorical variables, as well as missing (or conditionally missing) values, to be handled, then, ultimately, the similarity between observations is ascertained. Various (novel) isolation forest proximity measures were considered, but the quadratic depth similarity measure was found to be somewhat more successful during testing. Notably, the local structure of the data is emphasised during clustering if this measure is utilised. Finally, spectral clustering is applied to the matrix of similarities, using a fully connected graph and the Laplacian matrix related to a random walk. In short, clusters are produced using K-means on a projection of the data, from which it is possible to gain an understanding of the data's inherent structure. Crucially, K-means is repeated multiple times with different cluster initialisations and a set of clusters (i.e. a partition) is selected by minimising the within-cluster sum of squares.

As previously indicated, the approach was tested. In fact, it was tested on six alternative data sets of different types (continuous, categorical and mixed) and varying difficulty to ascertain whether it could produce reasonable results. It has already been highlighted that the quadratic depth similarity measure was somewhat more successful than the others, but with it the approach produced reasonable clusters for all of the data sets. Consequently, it could be concluded that the clustering approach developed can produce reasonable clusters for a variety of data sets.

Once the approach had been tested, a number of preliminary experiments were conducted on NACC data which built on the work discussed in chapter 4. As a matter of fact, the approach was applied to subsets of NACC data, extracted from the

imputed training set, focusing on cognitive status or the four main dementia subtypes. More specifically, the cognitive status subset included the 50 most important features for diagnosing dementia, according to the dementia classifier, as well as 2,000 subjects selected based on their cognitive status (normal cognition, mild cognitive impairment (MCI) or dementia). Contrastingly, the dementia subtypes subset included the 39 most important features for the differential diagnosis of dementia, according to the pairwise dementia subtype classifiers, along with 396 subjects chosen such that there was approximately an equal number of pure cases for each of the subtypes. Ultimately, these preliminary experiments suggested that MCI may be a mild form of dementia as opposed to a clinical entity, over which there is much debate; and there could be evidence for the current subtypes.

Chapter 6

Summary, Conclusions and Future Research

It has been predicted that the prevalence of dementia will increase significantly over the coming years; this, along with the considerable economic and social burden associated with dementia, is concerning. The thesis discussed research conducted with two primary aims, largely motivated by these factors and that it is currently difficult and time consuming to diagnose dementia reliably. The first aim was to investigate the use of machine learning for distinguishing between people with and without dementia, as well as differentiating between key dementia subtypes where appropriate. Notably, the four main subtypes are Alzheimer's disease (AD), vascular dementia (VD), dementia with Lewy bodies (DLB) and frontotemporal dementia (FTD). The second aim was to gain an understanding of the inherent structure of dementia data, to ultimately investigate disease signatures; it allowed for some investigation into whether the prevailing diagnostic criteria accurately reflect the nature of dementia and its subtypes. These aims were tackled using classification and clustering respectively.

The Uniform Data Set (UDS) was acquired from the National Alzheimer's Coordinating Center (NACC) for the purposes of this research. It comprises clinical and neuropsychological data from visits to Alzheimer's Disease Centers in the United States, during which the visitor (or subject) is assessed according to a standardised evaluation, specifically to ascertain a diagnosis that essentially indicates whether

they have dementia, along with the type of dementia if appropriate. The variables (or features) concerning diagnosis were extracted from the data set so labels, primarily for classification, could be generated. The remaining data was cleansed, resulting in a data set composed of 32,573 initial visits or subjects (i.e. observations) and 260 variables of mixed type (continuous, categorical, ordinal and binary). Crucially, the data set included variables in relations with one another and two types of missingness. The genuinely missing values, which arose when data was unexpectedly not recorded, were imputed where possible. The conditionally missing values, on the other hand, which arose when the information was irrelevant or unobtainable for a known reason, were handled during classification and clustering.

Two machine learning approaches were developed for this research. Firstly, an imputation approach was developed, which simultaneously builds a random forest classifier whilst handling conditionally missing values; it is termed proximity imputation with MIA (missingness incorporated in attributes). In fact, it is an amalgamation of the proximity imputation method (Breiman and Cutler, 2004), the Extra-Trees algorithm (Geurts, Ernst and Wehenkel, 2006) and MIA (Twala, Jones and Hand, 2008). Notably, the proximity imputation method was specifically tailored to maintain the known relations between variables in the NACC data set, so far as possible. To summarise, proximity imputation with MIA begins by crudely imputing the missing values in the data set (or training set) to enable a random forest to be constructed. Extra-Trees and MIA are subsequently employed to build the ensemble of decision trees, using the imputed data set. By inspecting the paths of the observations through every tree, the similarity of each pair of observations can be ascertained. These similarities (or proximities) are used to populate a proximity matrix, which is then utilised to impute the missing values for a second time. It is at this stage that precautions must be taken to ensure that the known relations between variables are maintained, so far as possible. Once a newly imputed data set has been formed, another random forest is built and the process is repeated for a number of iterations. Incidentally, the approach can also be used to impute test cases with a few alterations. Crucially, the imputed values are generated based on

the imputed training cases alone.

A clustering approach was also developed, which measures the proximity (distance or similarity, in this context) between observations based on the similarity of their paths through each tree of an isolation forest. Various (novel) isolation forest proximity measures were considered, but the quadratic depth similarity measure was found to be somewhat more successful during testing. Initially, an isolation forest is constructed in a manner that enables categorical variables, as well as missing (or conditionally missing) values, to be handled, then, ultimately, the similarity between observations is ascertained. Finally, spectral clustering is applied to the matrix of similarities, using a fully connected graph and the Laplacian matrix related to a random walk. In short, clusters are produced using K-means on a projection of the data, but it should be noted that K-means is repeated multiple times with different cluster initialisations and a set of clusters (i.e. a partition) is selected by minimising the within-cluster sum of squares.

The research produced a dementia classifier with an accuracy of 94.21%, a sensitivity of 0.93, a specificity of 0.95 and an area under the receiver operating characteristic curve (AUC) of 0.99, suggesting machine learning could be a useful tool for diagnosing dementia. It also produced 10 pairwise dementia subtype classifiers with AUCs ranging from 0.88 to 1.0 (rounded to two decimal places), indicating machine learning could be used to differentiate between the main dementia subtypes. Using these classifiers, it was possible to identify the key features for diagnosing dementia, as well as differentiating between the main subtypes of dementia; there is a clear difference between the important features for the two types of diagnosis. Furthermore, preliminary experiments conducted using the clustering approach developed suggested that mild cognitive impairment (MCI) may be a mild form of dementia as opposed to a clinical entity (i.e. a condition in its own right), over which there is much debate. They also suggested that there could be evidence for the current subtypes (AD, VD, DLB and FTD).

To summarise, the research prompted the development of two machine learning approaches and gave rise to what could be deemed valuable findings concerning

dementia and its diagnosis. It is hoped that these approaches will continue to prove useful and the findings ultimately help to improve the diagnosis of dementia. Nonetheless, there are numerous possible avenues for future research, some of which have been highlighted throughout the thesis. With regards to the clustering approach developed, there is scope to potentially improve it, specifically by employing either an isolation forest with random rotations (rotated trees) or an extended isolation forest (Hariri, Kind and Brunner, 2018). However, these isolation forest alternatives are unable to handle categorical variables, so it would be necessary to investigate how to extend them so that they can. It may also be interesting to investigate the effects of a k-nearest neighbour graph as opposed to a fully connected graph on clustering performance. In relation to the primary aims of the research, the main focus should be conducting more exploratory experiments using the clustering approach to enable disease signatures to be investigated and a clinical conclusion to be drawn. It is also vitally important that progress is made towards tangible changes in the diagnosis of dementia; developing a diagnostic aid from the classifiers produced could be a good place to start.

Appendix A

Data Cleansing Specifics

This appendix provides a more in-depth view of the data cleansing process, which was discussed in section 2.3, by detailing how each variable in the National Alzheimer's Coordinating Center Uniform Data Set (NACC UDS) was handled, excluding the diagnostic variables associated with Form D1. What follows was compiled using the researchers data dictionary (National Alzheimer's Coordinating Center, 2017) as a basis. However, it was also necessary to refer to the original forms (ADC Clinical Task Force and National Alzheimer's Coordinating Center, 2006c, 2014c, 2017c), coding guidebooks (ADC Clinical Task Force and National Alzheimer's Coordinating Center, 2006a, 2014a, 2017a) and data element dictionaries (ADC Clinical Task Force and National Alzheimer's Coordinating Center, 2006b, 2014b, 2017b) provided by NACC for the initial visits within the UDS.

The 18 forms and the form header are considered in turn. Firstly, information pertaining to the form or form header itself is provided. This includes the original number of variables; the number utilised; and a diagram showing the dependencies between the variables used, similar to the one in figure 2.2. For the vast majority of the forms, whether it is required to be completed according to versions 1.2 and 2.0 of the UDS is indicated. In addition to this, the proportion of the subjects considered for which the form was not completed is provided. This information is not applicable to the form header, and is not included for three forms (C2, D2, Milestones) as no variables were used from them.

Each of the variables utilised is subsequently considered in detail. Their name

is provided, along with the versions of the UDS they are in, or associated with for those which were newly derived. All of the variables are linked to versions 1.2 and 2.0, as a direct result of the way in which variables were selected for analysis. Whether the variable was derived is also indicated. There are two types of derived variable in the UDS, as defined by NACC, namely those encapsulating data collected differently across versions, and those resulting from some form of analysis by NACC (National Alzheimer's Coordinating Center, 2017). Only the latter are highlighted, along with those derived as part of this work to consolidate data and provide it in a more suitable format. A brief description of the variable, and its corresponding data type, is given next. This is followed by the possible values for the variable, and those identified as either missing (M) or conditionally missing (CM) are indicated. The proportions of the subjects considered with these types of values were calculated, and are stated. Brief comments on individual values are also provided where necessary, typically explaining why it was deemed missing or conditionally missing.

For those variables acting as the parent in a dependency, the value which causes the child to be conditionally missing is given; this is denoted the dependency trigger. A parent variable is immediately followed by any children it has, each of which is preceded with a dashed line. For any variables which can be determined by a relationship, said relationship is declared.

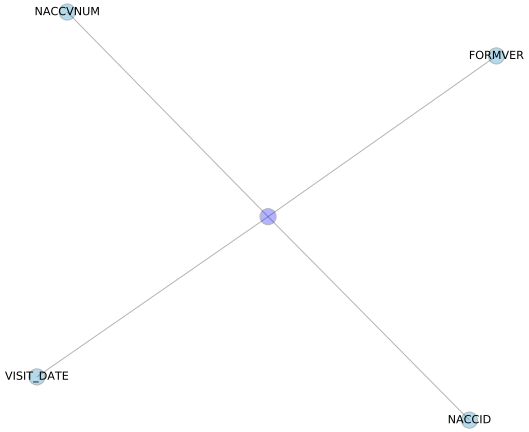
Missing values were imputed, and the variables requiring imputation are highlighted. It was important that the dependencies and relationships were maintained during imputation, so far as possible; thus, two additional actions are also considered. These are whether the parent should be inspected, if the variable is the child in a dependency; and whether an attempt should be made to calculate or derive the variable from others, if it can be determined by a relationship. These actions relate to the update steps discussed in section 3.2.5 in chapter 3, within which the imputation procedure is fully explained.

Further information is supplied for a number of the variables in the form of general comments. These can indicate any constraints on the variable, such as whether it would be appropriate for a conditionally missing value to be considered

as a potential fill value (i.e. imputed value); and inconsistencies between NACC's documentation and the data set, particularly in reference to dependencies and relationships. The general comments can also provide explanations for certain decisions, and the original variables used to derive a new one.

Finally, the variables that were dropped from the data set are listed. The name of each variable is provided, along with a brief description and the predominant reason(s) for its exclusion. There were a variety of the latter, but the most common ones were the variable contained free-text, and it was not in versions 1.2 and 2.0 of the UDS.

Form Header



Number of Variables	12
Number of Variables Used	4

NACCID

Form Versions	Derived (NACC)	Derived (New)
1.2, 2.0, 3.0	✗	✗

Description Subject identification number
Data Type NACC identifier

Values		Comments
Original	**Replacement**	
Prefix NACC followed by 0-10 numbers	-	-

Dependency Trigger -
Relationship -

Inspect Parent	Calculate/Derive	Impute
✗	✗	✗

General Comments Only to be used to identify subjects. Should be excluded when analysis is applied to the data set.

FORMVER

Form Versions	Derived (NACC)	Derived (New)
1.2, 2.0, 3.0	✗	✗

Description Form version number
Data Type Continuous

Values		Comments
Original	**Replacement**	
1-3	-	-

Dependency Trigger -
Relationship -

Inspect Parent	Calculate/Derive	Impute
✗	✗	✗

General Comments -

VISIT_DATE

Form Versions	Derived (NACC)	Derived (New)
1.2, 2.0, 3.0	✗	✓

Description Form date (year, month, day)
Data Type Date/Time object

Values		Comments
Original	Replacement	
Date from 2005 to 2016	-	-

Dependency Trigger -
Relationship -

Inspect Parent	Calculate/Derive	Impute
✗	✗	✗

General Comments Derived from VISITMO, VISITDAY and VISITYR. Only to be used to put other date variables into context. Should be excluded when analysis is applied to the data set.

NACCVNUM

Form Versions	Derived (NACC)	Derived (New)
1.2, 2.0, 3.0	✗	✗

Description UDS visit number (order)
Data Type Continuous

Values		Comments
Original	Replacement	
1-20	-	-

Dependency Trigger -
Relationship -

Inspect Parent	Calculate/Derive	Impute
✗	✗	✗

General Comments Constant variable as only initial visits considered. Retained for testing purposes.

Dropped Variables

Name NACCADC
Description Center identification number
Reason Information provided deemed irrelevant to the research.

Name PACKET
Description Packet code
Reason Constant variable as only initial visits considered.

Name VISITMO
Description Form date - month
Reason Replaced by the derived variable VISIT_DATE.

Name VISITDAY
Description Form date - day
Reason Replaced by the derived variable VISIT_DATE.

Name VISITYR
Description Form date - year
Reason Replaced by the derived variable VISIT_DATE.

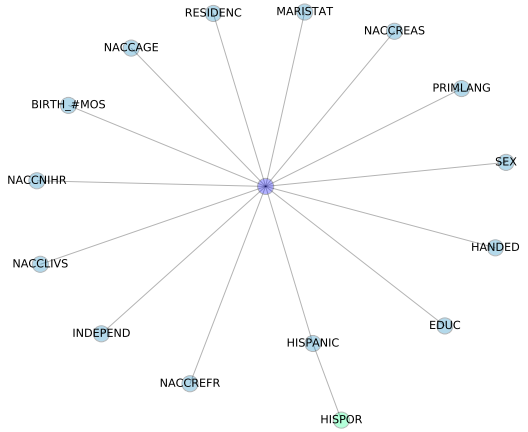
Name NACCAVST
Description Total number of all UDS visits made
Reason Variable is constant across visits. Information provided would not be available at an initial visit.

Name NACCNVST
Description Number of in-person UDS visits made
Reason Variable is constant across visits. Information provided would not be available at an initial visit.

Name NACCDAYS
Description Days from initial visit to most recent visit
Reason Variable is constant across visits. Information provided would not be available at an initial visit.

Name NACCFDYS
Description Days from initial visit to each follow-up visit
Reason Constant variable as only initial visits considered.

A1 - Subject Demographics



Number of Variables	25
Number of Variables Used	15
Form Required?	✓
Form Missingness	0.00%

NACCREAS

Form Versions	Derived (NACC)	Derived (New)
1.2, 2.0, 3.0	✗	✗

Description Primary reason for coming to an Alzheimer's Disease Center (ADC)
Data Type Categorical

Values		Comments
Original	**Replacement**	
1-2	-	-
7	-	-
9	M	(0.10%)
-4	M	Missing as form required and variable independent. (0.00%)

Dependency Trigger -
Relationship -

Inspect Parent	Calculate/Derive	Impute
✗	✗	✓

General Comments -

NACCREFR

Form Versions	Derived (NACC)	Derived (New)
1.2, 2.0, 3.0	✗	✗

Description Principal referral source
Data Type Categorical

Values		Comments
Original	**Replacement**	
1-2	-	-
8	-	-
9	M	(2.54%)
-4	M	Missing as form required and variable independent. (0.00%)

Dependency Trigger -
Relationship -

Inspect Parent	Calculate/Derive	Impute
✗	✗	✓

General Comments -

BIRTH_#MOS

Form Versions	Derived (NACC)	Derived (New)
1.2, 2.0, 3.0	✗	✓

Description Months from subject's month/year of birth to month/year of visit
Data Type Continuous

Values		Comments
Original	**Replacement**	
Positive integer	-	Year of birth from 1875 to 2001.

Dependency Trigger -
Relationship -

Inspect Parent	Calculate/Derive	Impute
✗	✗	✗

General Comments Derived from BIRTHMO and BIRTHYR.

SEX

Form Versions	Derived (NACC)	Derived (New)
1.2, 2.0, 3.0	✗	✗

Description Subject's sex
Data Type Binary

Values		Comments
Original	**Replacement**	
1-2	-	Binary due to number of options available.

Dependency Trigger -
Relationship -

Inspect Parent	Calculate/Derive	Impute
✗	✗	✗

General Comments -

HISPANIC

Form Versions	Derived (NACC)	Derived (New)
1.2, 2.0, 3.0	✗	✗

Description Hispanic/Latino ethnicity
Data Type Binary

Values		Comments
Original	Replacement	
0-1	-	-
9	M	(0.41%)

Dependency Trigger 0
Relationship -

Inspect Parent	Calculate/Derive	Impute
✗	✗	✓

General Comments -

HISPOR

Form Versions	Derived (NACC)	Derived (New)
1.2, 2.0, 3.0	✗	✗

Description Hispanic origins
Data Type Categorical

Values		Comments
Original	Replacement	
1-6	-	-
50	-	-
88	CM	(92.36%)
99	M	(0.21%)
-4	CM	Conditionally missing even though form required as variable dependent. (0.00%)

Dependency Trigger -
Relationship -

Inspect Parent	Calculate/Derive	Impute
✓	✗	✓

General Comments Values should be updated if parent imputed. A conditionally missing value should not be used as a potential fill value if child imputed.

PRIMLANG

Form Versions	Derived (NACC)	Derived (New)
1.2, 2.0, 3.0	✗	✗

Description Primary language
Data Type Categorical

Values		Comments
Original	Replacement	
1-6	-	-
8	-	-
9	M	(0.10%)

Dependency Trigger -
Relationship -

Inspect Parent	Calculate/Derive	Impute
✗	✗	✓

General Comments -

EDUC

Form Versions	Derived (NACC)	Derived (New)
1.2, 2.0, 3.0	✗	✗

Description Years of education
Data Type Continuous

Values		Comments
Original	Replacement	
0-36	-	-
99	M	(0.71%)

Dependency Trigger -
Relationship -

Inspect Parent	Calculate/Derive	Impute
✗	✗	✓

General Comments -

NACCLIVS

Form Versions	Derived (NACC)	Derived (New)
1.2, 2.0, 3.0	✗	✗

Description Living situation
Data Type Categorical

Values		Comments
Original	Replacement	
1-5	-	-
9	M	(0.25%)

Dependency Trigger -

Relationship -

Inspect Parent	Calculate/Derive	Impute
✗	✗	✓

General Comments -

INDEPEND

Form Versions	Derived (NACC)	Derived (New)
1.2, 2.0, 3.0	✗	✗

Description Level of independence
Data Type Ordinal

Values		Comments
Original	Replacement	
1-4	-	-
9	M	(0.38%)

Dependency Trigger -
Relationship -

Inspect Parent	Calculate/Derive	Impute
✗	✗	✓

General Comments -

RESIDENC

Form Versions	Derived (NACC)	Derived (New)
1.2, 2.0, 3.0	✗	✗

Description Type of residence
Data Type Categorical

Values		Comments
Original	Replacement	
1-4	-	-
9	-	Not replaced as 'other or unknown'.

Dependency Trigger -
Relationship -

Inspect Parent	Calculate/Derive	Impute
✗	✗	✗

General Comments -

MARISTAT

Form Versions	Derived (NACC)	Derived (New)
1.2, 2.0, 3.0	✗	✗

Description Marital status
Data Type Categorical

Values		Comments
Original	**Replacement**	
1-6	-	-
9	-	Not replaced as 'other or unknown'.

Dependency Trigger -
Relationship -

Inspect Parent	Calculate/Derive	Impute
✗	✗	✗

General Comments -

HANDED

Form Versions	Derived (NACC)	Derived (New)
1.2, 2.0, 3.0	✗	✗

Description Is the subject left- or right-handed?
Data Type Categorical

Values		Comments
Original	**Replacement**	
1-3	-	-
9	CM	Conditionally missing as cannot be sensibly imputed. (0.53%)

Dependency Trigger -
Relationship -

Inspect Parent	Calculate/Derive	Impute
✗	✗	✗

General Comments -

NACCAGE

Form Versions	Derived (NACC)	Derived (New)
1.2, 2.0, 3.0	✓	✗

Description Subject's age at visit
Data Type Continuous

Values		Comments
Original	Replacement	
18-120	-	-

Dependency Trigger -
Relationship -

Inspect Parent	Calculate/Derive	Impute
✗	✗	✗

General Comments -

NACCNIHR

Form Versions	Derived (NACC)	Derived (New)
1.2, 2.0, 3.0	✓	✗

Description Derived National Institutes of Health (NIH) race definitions
Data Type Categorical

Values		Comments
Original	Replacement	
1-6	-	-
99	M	(1.63%)

Dependency Trigger -
Relationship -

Inspect Parent	Calculate/Derive	Impute
✗	✗	✓

General Comments -

Dropped Variables

Name BIRTHMO
Description Subject's month of birth
Reason Replaced by the derived variable BIRTH_#MOS.

Name BIRTHYR
Description Subject's year of birth
Reason Replaced by the derived variable BIRTH_#MOS.

Name HISPORX
Description Hispanic origins, other - specify
Reason Free-text variable.

Name RACE
Description Race
Reason Information provided used to generate NACCNIHR.

Name RACEX
Description Race, other - specify
Reason Free-text variable and information provided used to generate NACCNIHR.

Name RACESEC
Description Second race
Reason Information provided used to generate NACCNIHR.

Name RACESECX
Description Second race, other - specify
Reason Free-text variable and information provided used to generate NACCNIHR.

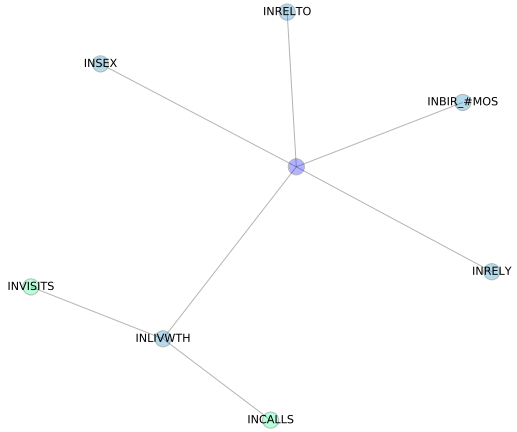
Name RACETER
Description Third race
Reason Information provided used to generate NACCNIHR.

Name RACETERX
Description Third race, other - specify
Reason Free-text variable and information provided used to generate NACCNIHR.

Name PRIMLANX
Description Primary language, other - specify
Reason Free-text variable.

Name NACCAGEB
Description Subject's age at initial visit
Reason Equivalent to NACCAGE as only initial visits considered.

A2 - Co-participant Demographics



Number of Variables	22
Number of Variables Used	7
Form Required?	✗
Form Missingness	6.32%

INBIR_#MOS

Form Versions	Derived (NACC)	Derived (New)
1.2, 2.0, 3.0	✗	✓

Description Months from co-participant's month/year of birth to month/year of visit
Data Type Continuous

Values		Comments
Original	**Replacement**	
Positive integer	-	Year of birth from 1875 to 2001.
CM	-	Conditionally missing as cannot be sensibly imputed and form not required. (5.16% (9999) + 6.32% (-4))

Dependency Trigger -
Relationship -

Inspect Parent	Calculate/Derive	Impute
✗	✗	✗

General Comments Derived from INBIRMO and INBIRYR. Missingness based on that of INBIRYR. 'Average' month (June) used if only year provided.

INSEX

Form Versions	Derived (NACC)	Derived (New)
1.2, 2.0, 3.0	✗	✗

Description Co-participant's sex
Data Type Binary

Values		Comments
Original	**Replacement**	
1-2	-	Binary due to number of options available.
-4	CM	Conditionally missing as form not required. (6.32%)

Dependency Trigger -
Relationship -

Inspect Parent	Calculate/Derive	Impute
✗	✗	✗

General Comments -

INRELTO

Form Versions	Derived (NACC)	Derived (New)
1.2, 2.0, 3.0	✗	✗

Description Co-participant's relationship to subject
Data Type Categorical

Values		Comments
Original	**Replacement**	
1-7	-	-
-4	CM	Conditionally missing as form not required. (6.32%)

Dependency Trigger -
Relationship -

Inspect Parent	Calculate/Derive	Impute
✗	✗	✗

General Comments -

INLIVWTH

Form Versions	Derived (NACC)	Derived (New)
1.2, 2.0, 3.0	✗	✗

Description Does the co-participant live with the subject?
Data Type Binary

Values		Comments
Original	**Replacement**	
0-1	-	-
-4	CM	Conditionally missing as form not required. (6.32%)

Dependency Trigger 1
Relationship -

Inspect Parent	Calculate/Derive	Impute
✗	✗	✗

General Comments -

INVISITS

Form Versions	Derived (NACC)	Derived (New)
1.2, 2.0, 3.0	✗	✗

Description If no, approximate frequency of in-person visits?
Data Type Ordinal

Values		Comments
Original	**Replacement**	
1-6	-	-
8	CM	(58.96%)
-4	CM	Conditionally missing as form not required. (6.32%)

Dependency Trigger -
Relationship -

Inspect Parent	Calculate/Derive	Impute
✗	✗	✗

General Comments No need to inspect parent as no missing values in parent.

INCALLS

Form Versions	Derived (NACC)	Derived (New)
1.2, 2.0, 3.0	✗	✗

Description If no, approximate frequency of telephone contact?
Data Type Ordinal

Values		Comments
Original	**Replacement**	
1-6	-	-
8	CM	(58.96%)
-4	CM	Conditionally missing as form not required. (6.32%)

Dependency Trigger -
Relationship -

Inspect Parent	Calculate/Derive	Impute
✗	✗	✗

General Comments No need to inspect parent as no missing values in parent.

INRELY

Form Versions	Derived (NACC)	Derived (New)
1.2, 2.0, 3.0	✗	✗

Description Is there a question about the co-participant's reliability?
Data Type Binary

	Values		Comments
	Original	**Replacement**	
	0-1	-	-
	-4	CM	Conditionally missing as form not required. (6.32%)

Dependency Trigger -
Relationship -

Inspect Parent	Calculate/Derive	Impute
✗	✗	✗

General Comments -

Dropped Variables

Name INBIRMO
Description Co-participant's month of birth
Reason Replaced by the derived variable INBIR_#MOS.

Name INBIRYR
Description Co-participant's year of birth
Reason Replaced by the derived variable INBIR_#MOS.

Name INHISP
Description Co-participant Hispanic/Latino ethnicity
Reason Information provided deemed irrelevant to the research.

Name INHISPOR
Description Co-participant's Hispanic origins
Reason Information provided deemed irrelevant to the research.

Name INHISPOX
Description Co-participant of Hispanic origins, other - specify
Reason Free-text variable and information provided deemed irrelevant to the research.

Name INRACE
Description Co-participant race
Reason Information provided deemed irrelevant to the research.

Name INRACEX
Description Co-participant race, other - specify
Reason Free-text variable and information provided deemed irrelevant to the research.

Name INRASEC
Description Co-participant second race
Reason Information provided deemed irrelevant to the research.

Name INRASECX
Description Co-participant second race, other - specify
Reason Free-text variable and information provided deemed irrelevant to the research.

Name INRATER
Description Co-participant third race
Reason Information provided deemed irrelevant to the research.

Name INRATERX
Description Co-participant third race, other - specify
Reason Free-text variable and information provided deemed irrelevant to the research.

Name INEDUC
Description Co-participant's years of education
Reason Information provided deemed irrelevant to the research.

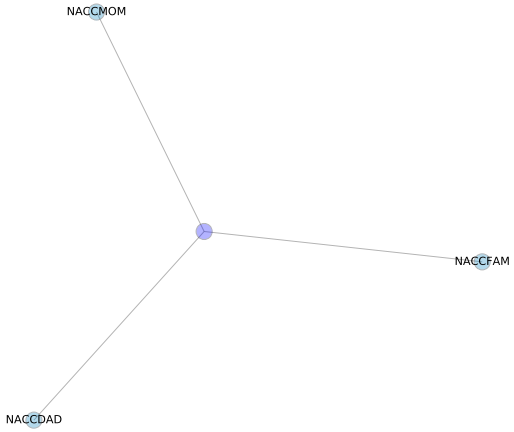
Name INRELTOX
Description Co-participant relationship, other - specify
Reason Free-text variable.

Name NACCNINR
Description Derived National Institutes of Health (NIH) race definitions
Reason Information provided deemed irrelevant to the research.

Name INKNOWN
Description How long has the co-participant known the subject?
Reason Only available in version 3.0 of the UDS.

Name NEWINF
Description Is this a new co-participant - i.e. one who was not a co-participant at any past UDS visit?
Reason Information provided would not be available at an initial visit.

A3 - Family History



Number of Variables	15
Number of Variables Used	3
Form Required?	✗
Form Missingness	1.15%

NACCMOM

Form Versions	Derived (NACC)	Derived (New)
1.2, 2.0, 3.0	✓	✗

Description Indicator of mother with cognitive impairment
Data Type Binary

Values		Comments
Original	**Replacement**	
0-1	-	-
9	M	(4.81%)
-4	CM	Conditionally missing as form not required. (1.15%)

Dependency Trigger -
Relationship -

Inspect Parent	Calculate/Derive	Impute
✗	✗	✓

General Comments A conditionally missing value should not be used as a potential fill value if imputed.

NACCDAD

Form Versions	Derived (NACC)	Derived (New)
1.2, 2.0, 3.0	✓	✗

Description Indicator of father with cognitive impairment
Data Type Binary

Values		Comments
Original	**Replacement**	
0-1	-	-
9	M	(7.37%)
-4	CM	Conditionally missing as form not required. (1.15%)

Dependency Trigger -
Relationship -

Inspect Parent	Calculate/Derive	Impute
✗	✗	✓

General Comments A conditionally missing value should not be used as a potential fill value if imputed.

NACCFAM

Form Versions	Derived (NACC)	Derived (New)
1.2, 2.0, 3.0	✓	✗

Description Indicator of first-degree family member with cognitive impairment
Data Type Binary

Values		Comments
Original	**Replacement**	
0-1	-	-
9	M	(9.85%)
-4	CM	Conditionally missing as form not required. (1.15%)

Dependency Trigger Relationship -
NACCFAM = 1 if NACCMOM and/or NACCDAD = 1

Inspect Parent	Calculate/Derive	Impute
✗	✓	✓

General Comments Only missing values should be updated if NACCMOM and/or NACCDAD imputed. A conditionally missing value should not be used as a potential fill value if variable imputed.

Dropped Variables

Name NACCAM
Description In this family, is there evidence for an Alzheimer's disease (AD) mutation?
Reason Only available in version 3.0 of the UDS.

Name NACCAMX
Description If yes, other - specify
Reason Free-text variable and only available in version 3.0 of the UDS.

Name NACCAMS
Description Source of evidence for AD mutation
Reason Only available in version 3.0 of the UDS.

Name NACCAMSX
Description If other - specify
Reason Free-text variable and only available in version 3.0 of the UDS.

Name NACCFM
Description In this family, is there evidence for a frontotemporal lobar degeneration (FTLD) mutation?
Reason Only available in version 3.0 of the UDS.

Name NACCFMX
Description If yes, other - specify
Reason Free-text variable and only available in version 3.0 of the UDS.

Name NACCFMS
Description Source of evidence for FTLD mutation
Reason Only available in version 3.0 of the UDS.

Name NACCFMSX
Description If other - specify
Reason Free-text variable and only available in version 3.0 of the UDS.

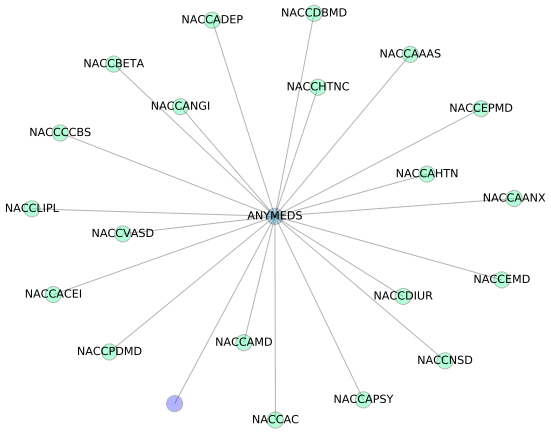
Name NACCOM
Description In this family, is there evidence for a mutation other than an AD or FTLD mutation?
Reason Only available in version 3.0 of the UDS.

Name NACCOMX
Description If yes - specify
Reason Free-text variable and only available in version 3.0 of the UDS.

Name NACCOMS
Description Source of evidence for other mutation
Reason Only available in version 3.0 of the UDS.

Name NACCOMSX
Description If other - specify
Reason Free-text variable and only available in version 3.0 of the UDS.

A4 - Medications



Number of Variables	62
Number of Variables Used	21
Form Required?	✗
Form Missingness	1.09%

ANYMEDS

Form Versions	Derived (NACC)	Derived (New)
1.2, 2.0, 3.0	✗	✗

Description Subject taking any medications
Data Type Binary

Values		Comments
Original	**Replacement**	
0-1	-	-
-4	CM	Conditionally missing as form not required. (1.09%)

Dependency Trigger 0
Relationship -

Inspect Parent	Calculate/Derive	Impute
✗	✗	✗

General Comments -

NACCAMD

Form Versions	Derived (NACC)	Derived (New)
1.2, 2.0, 3.0	✓	✗

Description Total number of medications reported at each visit
Data Type Continuous

Values		Comments
Original	**Replacement**	
0-40	-	-
CM	-	Replaced 0 and -4 values to indicate dependence on ANYMEDS. (8.14%)
-4	CM	Conditionally missing as form not required. (1.09%)

Dependency Trigger -
Relationship -

Inspect Parent	Calculate/Derive	Impute
✗	✗	✗

General Comments No need to inspect parent as no missing values in parent.

NACCHTNC

Form Versions	Derived (NACC)	Derived (New)
1.2, 2.0, 3.0	✓	✗

Description Reported current use of an antihypertensive combination therapy
Data Type Binary

Values		Comments
Original	**Replacement**	
0-1	-	-
CM	-	Replaced 0 and -4 values to indicate dependence on ANYMEDS. (8.14%)
-4	CM	Conditionally missing as form not required. (1.09%)

Dependency Trigger -
Relationship -

Inspect Parent	Calculate/Derive	Impute
✗	✗	✗

General Comments No need to inspect parent as no missing values in parent.

NACCACEI

Form Versions	Derived (NACC)	Derived (New)
1.2, 2.0, 3.0	✓	✗

Description Reported current use of an angiotensin converting enzyme (ACE) inhibitor
Data Type Binary

Values		Comments
Original	**Replacement**	
0-1	-	-
CM	-	Replaced 0 and -4 values to indicate dependence on ANYMEDS. (8.14%)
-4	CM	Conditionally missing as form not required. (1.09%)

Dependency Trigger -
Relationship -

Inspect Parent	Calculate/Derive	Impute
✗	✗	✗

General Comments No need to inspect parent as no missing values in parent.

NACCAAAS

Form Versions	Derived (NACC)	Derived (New)
1.2, 2.0, 3.0	✓	✗

Description Reported current use of an antiadrenergic agent
Data Type Binary

Values		Comments
Original	**Replacement**	
0-1	-	-
CM	-	Replaced 0 and -4 values to indicate dependence on ANYMEDS. (8.14%)
-4	CM	Conditionally missing as form not required. (1.09%)

Dependency Trigger -
Relationship -

Inspect Parent	Calculate/Derive	Impute
✗	✗	✗

General Comments No need to inspect parent as no missing values in parent.

NACCBETA

Form Versions	Derived (NACC)	Derived (New)
1.2, 2.0, 3.0	✓	✗

Description Reported current use of a beta-adrenergic blocking agent (Beta-Blocker)
Data Type Binary

Values		Comments
Original	**Replacement**	
0-1	-	-
CM	-	Replaced 0 and -4 values to indicate dependence on ANYMEDS. (8.14%)
-4	CM	Conditionally missing as form not required. (1.09%)

Dependency Trigger -
Relationship -

Inspect Parent	Calculate/Derive	Impute
✗	✗	✗

General Comments No need to inspect parent as no missing values in parent.

NACCCCBS

Form Versions	Derived (NACC)	Derived (New)
1.2, 2.0, 3.0	✓	✗

Description Reported current use of a calcium channel blocking agent
Data Type Binary

Values		Comments
Original	**Replacement**	
0-1	-	-
CM	-	Replaced 0 and -4 values to indicate dependence on ANYMEDS. (8.14%)
-4	CM	Conditionally missing as form not required. (1.09%)

Dependency Trigger -
Relationship -

Inspect Parent	Calculate/Derive	Impute
✗	✗	✗

General Comments No need to inspect parent as no missing values in parent.

NACCDIUR

Form Versions	Derived (NACC)	Derived (New)
1.2, 2.0, 3.0	✓	✗

Description Reported current use of a diuretic
Data Type Binary

Values		Comments
Original	**Replacement**	
0-1	-	-
CM	-	Replaced 0 and -4 values to indicate dependence on ANYMEDS. (8.14%)
-4	CM	Conditionally missing as form not required. (1.09%)

Dependency Trigger -
Relationship -

Inspect Parent	Calculate/Derive	Impute
✗	✗	✗

General Comments No need to inspect parent as no missing values in parent.

NACCVASD

Form Versions	Derived (NACC)	Derived (New)
1.2, 2.0, 3.0	✓	✗

Description Reported current use of a vasodilator
Data Type Binary

Values		Comments
Original	**Replacement**	
0-1	-	-
CM	-	Replaced 0 and -4 values to indicate dependence on ANYMEDS. (8.14%)
-4	CM	Conditionally missing as form not required. (1.09%)

Dependency Trigger -
Relationship -

Inspect Parent	Calculate/Derive	Impute
✗	✗	✗

General Comments No need to inspect parent as no missing values in parent.

NACCANGI

Form Versions	Derived (NACC)	Derived (New)
1.2, 2.0, 3.0	✓	✗

Description Reported current use of an angiotensin II inhibitor
Data Type Binary

Values		Comments
Original	**Replacement**	
0-1	-	-
CM	-	Replaced 0 and -4 values to indicate dependence on ANYMEDS. (8.14%)
-4	CM	Conditionally missing as form not required. (1.09%)

Dependency Trigger -
Relationship -

Inspect Parent	Calculate/Derive	Impute
✗	✗	✗

General Comments No need to inspect parent as no missing values in parent.

NACCAHTN

Form Versions	Derived (NACC)	Derived (New)
1.2, 2.0, 3.0	✓	✗

Description Reported current use of any type of antihypertensive or blood pressure medication
Data Type Binary

Values		Comments
Original	**Replacement**	
0-1	-	-
CM	-	Replaced 0 and -4 values to indicate dependence on ANYMEDS. (8.14%)
-4	CM	Conditionally missing as form not required. (1.09%)

Dependency Trigger Relationship -
NACCAHTN = 1 if NACCHTNC, NACCACEI, NACCAAAS, NACCBETA, NACCCCBS, NACCDIUR, NACCVASD and/or NACCANGI = 1 else 0

Inspect Parent	Calculate/Derive	Impute
✗	✗	✗

General Comments No need to inspect parent or calculate/derive variable as no missing values in any of the variables involved.

NACCLIPL

Form Versions	Derived (NACC)	Derived (New)
1.2, 2.0, 3.0	✓	✗

Description Reported current use of lipid lowering medication
Data Type Binary

Values		Comments
Original	**Replacement**	
0-1	-	-
CM	-	Replaced 0 and -4 values to indicate dependence on ANYMEDS. (8.14%)
-4	CM	Conditionally missing as form not required. (1.09%)

Dependency Trigger Relationship -
-

Inspect Parent	Calculate/Derive	Impute
✗	✗	✗

General Comments No need to inspect parent as no missing values in parent.

NACCNSD

Form Versions	Derived (NACC)	Derived (New)
1.2, 2.0, 3.0	✓	✗

Description Reported current use of nonsteroidal anti-inflammatory medication
Data Type Binary

Values		Comments
Original	**Replacement**	
0-1	-	-
CM	-	Replaced 0 and -4 values to indicate dependence on ANYMEDS. (8.14%)
-4	CM	Conditionally missing as form not required. (1.09%)

Dependency Trigger -
Relationship -

Inspect Parent	Calculate/Derive	Impute
✗	✗	✗

General Comments No need to inspect parent as no missing values in parent.

NACCAC

Form Versions	Derived (NACC)	Derived (New)
1.2, 2.0, 3.0	✓	✗

Description Reported current use of an anticoagulant or antiplatelet agent
Data Type Binary

Values		Comments
Original	**Replacement**	
0-1	-	-
CM	-	Replaced 0 and -4 values to indicate dependence on ANYMEDS. (8.14%)
-4	CM	Conditionally missing as form not required. (1.09%)

Dependency Trigger -
Relationship -

Inspect Parent	Calculate/Derive	Impute
✗	✗	✗

General Comments No need to inspect parent as no missing values in parent.

NACCADEP

Form Versions	Derived (NACC)	Derived (New)
1.2, 2.0, 3.0	✓	✗

Description Reported current use of an antidepressant
Data Type Binary

Values		Comments
Original	**Replacement**	
0-1	-	-
CM	-	Replaced 0 and -4 values to indicate dependence on ANYMEDS. (8.14%)
-4	CM	Conditionally missing as form not required. (1.09%)

Dependency Trigger -
Relationship -

Inspect Parent	Calculate/Derive	Impute
✗	✗	✗

General Comments No need to inspect parent as no missing values in parent.

NACCAPSY

Form Versions	Derived (NACC)	Derived (New)
1.2, 2.0, 3.0	✓	✗

Description Reported current use of an antipsychotic agent
Data Type Binary

Values		Comments
Original	**Replacement**	
0-1	-	-
CM	-	Replaced 0 and -4 values to indicate dependence on ANYMEDS. (8.14%)
-4	CM	Conditionally missing as form not required. (1.09%)

Dependency Trigger -
Relationship -

Inspect Parent	Calculate/Derive	Impute
✗	✗	✗

General Comments No need to inspect parent as no missing values in parent.

NACCAANX

Form Versions	Derived (NACC)	Derived (New)
1.2, 2.0, 3.0	✓	✗

Description Reported current use of an anxiolytic, sedative or hypnotic agent
Data Type Binary

Values		Comments
Original	**Replacement**	
0-1	-	-
CM	-	Replaced 0 and -4 values to indicate dependence on ANYMEDS. (8.14%)
-4	CM	Conditionally missing as form not required. (1.09%)

Dependency Trigger -
Relationship -

Inspect Parent	Calculate/Derive	Impute
✗	✗	✗

General Comments No need to inspect parent as no missing values in parent.

NACCPDMD

Form Versions	Derived (NACC)	Derived (New)
1.2, 2.0, 3.0	✓	✗

Description Reported current use of an antiparkinson agent
Data Type Binary

Values		Comments
Original	**Replacement**	
0-1	-	-
CM	-	Replaced 0 and -4 values to indicate dependence on ANYMEDS. (8.14%)
-4	CM	Conditionally missing as form not required. (1.09%)

Dependency Trigger -
Relationship -

Inspect Parent	Calculate/Derive	Impute
✗	✗	✗

General Comments No need to inspect parent as no missing values in parent.

NACCEMD

Form Versions	Derived (NACC)	Derived (New)
1.2, 2.0, 3.0	✓	✗

Description Reported current use of estrogen hormone therapy
Data Type Binary

Values		Comments
Original	**Replacement**	
0-1	-	-
CM	-	Replaced 0 and -4 values to indicate dependence on ANYMEDS. (8.14%)
-4	CM	Conditionally missing as form not required. (1.09%)

Dependency Trigger -
Relationship -

Inspect Parent	Calculate/Derive	Impute
✗	✗	✗

General Comments No need to inspect parent as no missing values in parent.

NACCEPMD

Form Versions	Derived (NACC)	Derived (New)
1.2, 2.0, 3.0	✓	✗

Description Reported current use of estrogen + progestin hormone therapy
Data Type Binary

Values		Comments
Original	**Replacement**	
0-1	-	-
CM	-	Replaced 0 and -4 values to indicate dependence on ANYMEDS. (8.14%)
-4	CM	Conditionally missing as form not required. (1.09%)

Dependency Trigger -
Relationship -

Inspect Parent	Calculate/Derive	Impute
✗	✗	✗

General Comments No need to inspect parent as no missing values in parent.

NACCDBMD

Form Versions	Derived (NACC)	Derived (New)
1.2, 2.0, 3.0	✓	✗

Description Reported current use of a diabetes medication
Data Type Binary

Values		Comments
Original	**Replacement**	
0-1	-	-
CM	-	Replaced 0 and -4 values to indicate dependence on ANYMEDS. (8.14%)
-4	CM	Conditionally missing as form not required. (1.09%)

Dependency Trigger -
Relationship -

Inspect Parent	Calculate/Derive	Impute
✗	✗	✗

General Comments No need to inspect parent as no missing values in parent.

Dropped Variables

Name DRUG1
Description Name of medication used within two weeks of UDS visit
Reason Free-text variable.

Name DRUG2
Description Name of medication used within two weeks of UDS visit
Reason Free-text variable.

Name DRUG3
Description Name of medication used within two weeks of UDS visit
Reason Free-text variable.

Name DRUG4
Description Name of medication used within two weeks of UDS visit
Reason Free-text variable.

Name DRUG5
Description Name of medication used within two weeks of UDS visit
Reason Free-text variable.

Name DRUG6
Description Name of medication used within two weeks of UDS visit
Reason Free-text variable.

Name DRUG7
Description Name of medication used within two weeks of UDS visit
Reason Free-text variable.

Name DRUG8
Description Name of medication used within two weeks of UDS visit
Reason Free-text variable.

Name DRUG9
Description Name of medication used within two weeks of UDS visit
Reason Free-text variable.

Name DRUG10
Description Name of medication used within two weeks of UDS visit
Reason Free-text variable.

Name DRUG11
Description Name of medication used within two weeks of UDS visit
Reason Free-text variable.

Name DRUG12
Description Name of medication used within two weeks of UDS visit
Reason Free-text variable.

Name DRUG13
Description Name of medication used within two weeks of UDS visit
Reason Free-text variable.

Name DRUG14
Description Name of medication used within two weeks of UDS visit
Reason Free-text variable.

Name DRUG15
Description Name of medication used within two weeks of UDS visit
Reason Free-text variable.

Name DRUG16
Description Name of medication used within two weeks of UDS visit
Reason Free-text variable.

Name DRUG17
Description Name of medication used within two weeks of UDS visit
Reason Free-text variable.

Name DRUG18
Description Name of medication used within two weeks of UDS visit
Reason Free-text variable.

Name DRUG19
Description Name of medication used within two weeks of UDS visit
Reason Free-text variable.

Name DRUG20
Description Name of medication used within two weeks of UDS visit
Reason Free-text variable.

Name DRUG21
Description Name of medication used within two weeks of UDS visit
Reason Free-text variable.

Name DRUG22
Description Name of medication used within two weeks of UDS visit
Reason Free-text variable.

Name DRUG23
Description Name of medication used within two weeks of UDS visit
Reason Free-text variable.

Name DRUG24
Description Name of medication used within two weeks of UDS visit
Reason Free-text variable.

Name DRUG25
Description Name of medication used within two weeks of UDS visit
Reason Free-text variable.

Name DRUG26
Description Name of medication used within two weeks of UDS visit
Reason Free-text variable.

Name DRUG27
Description Name of medication used within two weeks of UDS visit
Reason Free-text variable.

Name DRUG28
Description Name of medication used within two weeks of UDS visit
Reason Free-text variable.

Name DRUG29
Description Name of medication used within two weeks of UDS visit
Reason Free-text variable.

Name DRUG30
Description Name of medication used within two weeks of UDS visit
Reason Free-text variable.

Name DRUG31
Description Name of medication used within two weeks of UDS visit
Reason Free-text variable.

Name DRUG32
Description Name of medication used within two weeks of UDS visit
Reason Free-text variable.

Name DRUG33
Description Name of medication used within two weeks of UDS visit
Reason Free-text variable.

Name DRUG34
Description Name of medication used within two weeks of UDS visit
Reason Free-text variable.

Name DRUG35
Description Name of medication used within two weeks of UDS visit
Reason Free-text variable.

Name DRUG36
Description Name of medication used within two weeks of UDS visit
Reason Free-text variable.

Name DRUG37
Description Name of medication used within two weeks of UDS visit
Reason Free-text variable.

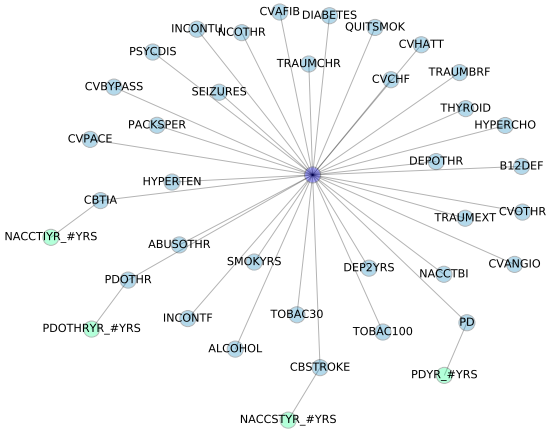
Name DRUG38
Description Name of medication used within two weeks of UDS visit
Reason Free-text variable.

Name DRUG39
Description Name of medication used within two weeks of UDS visit
Reason Free-text variable.

Name DRUG40
Description Name of medication used within two weeks of UDS visit
Reason Free-text variable.

Name NACCADMD
Description Reported current use of FDA-approved medication for Alzheimer's disease (AD) symptoms
Reason Information provided may indicate AD has previously been diagnosed.

A5 - Health History



Number of Variables	75
Number of Variables Used	38
Form Required?	✓
Form Missingness	0.00%

CVHATT

Form Versions	Derived (NACC)	Derived (New)
1.2, 2.0, 3.0	✗	✗

Description Heart attack/cardiac arrest
Data Type Categorical

Values		Comments
Original	**Replacement**	
0-2	-	-
9	M	(0.35%)
-4	M	Missing as form required and variable independent. (0.00%)

Dependency Trigger -
Relationship -

Inspect Parent	Calculate/Derive	Impute
✗	✗	✓

General Comments -

CVAFIB

Form Versions	Derived (NACC)	Derived (New)
1.2, 2.0, 3.0	✗	✗

Description Atrial fibrillation
Data Type Categorical

Values		Comments
Original	**Replacement**	
0-2	-	-
9	M	(0.51%)
-4	M	Missing as form required and variable independent. (0.00%)

Dependency Trigger -
Relationship -

Inspect Parent	Calculate/Derive	Impute
✗	✗	✓

General Comments -

CVANGIO

Form Versions	Derived (NACC)	Derived (New)
1.2, 2.0, 3.0	✗	✗

Description Angioplasty/endarterectomy/stent
Data Type Categorical

Values		Comments
Original	**Replacement**	
0-2	-	-
9	M	(0.14%)
-4	M	Missing as form required and variable independent. (0.00%)

Dependency Trigger -
Relationship -

Inspect Parent	Calculate/Derive	Impute
✗	✗	✓

General Comments -

CVBYPASS

Form Versions	Derived (NACC)	Derived (New)
1.2, 2.0, 3.0	✗	✗

Description Cardiac bypass procedure
Data Type Categorical

Values		Comments
Original	**Replacement**	
0-2	-	-
9	M	(0.11%)
-4	M	Missing as form required and variable independent. (0.00%)

Dependency Trigger -
Relationship -

Inspect Parent	Calculate/Derive	Impute
✗	✗	✓

General Comments -

CVPACE

Form Versions	Derived (NACC)	Derived (New)
1.2, 2.0	✗	✗

Description Pacemaker
Data Type Categorical

Values		Comments
Original	**Replacement**	
0-2	-	-
9	M	(0.09%)
-4	M	Missing as form required and variable independent. (0.00%)

Dependency Trigger -
Relationship -

Inspect Parent	Calculate/Derive	Impute
✗	✗	✓

General Comments -

CVCHF

Form Versions	Derived (NACC)	Derived (New)
1.2, 2.0, 3.0	✗	✗

Description Congestive heart failure
Data Type Categorical

Values		Comments
Original	**Replacement**	
0-2	-	-
9	M	(0.25%)
-4	M	Missing as form required and variable independent. (0.00%)

Dependency Trigger -
Relationship -

Inspect Parent	Calculate/Derive	Impute
✗	✗	✓

General Comments -

CVOTHR

Form Versions	Derived (NACC)	Derived (New)
1.2, 2.0, 3.0	✗	✗

Description Other cardiovascular disease
Data Type Categorical

Values		Comments
Original	Replacement	
0-2	-	-
9	M	(0.65%)
-4	M	Missing as form required and variable independent. (0.00%)

Dependency Trigger -
Relationship -

Inspect Parent	Calculate/Derive	Impute
✗	✗	✓

General Comments -

CBSTROKE

Form Versions	Derived (NACC)	Derived (New)
1.2, 2.0, 3.0	✗	✗

Description Stroke
Data Type Categorical

Values		Comments
Original	Replacement	
0-2	-	-
9	M	(0.40%)
-4	M	Missing as form required and variable independent. (0.00%)

Dependency Trigger 0
Relationship -

Inspect Parent	Calculate/Derive	Impute
✗	✗	✓

General Comments -

NACCSTYR_#YRS

Form Versions	Derived (NACC)	Derived (New)
1.2, 2.0, 3.0	✗	✓

Description Years from most recently reported year of stroke as of the initial visit to year of visit
Data Type Continuous

Values		Comments
Original	**Replacement**	
Positive integer	-	Year of stroke from 1900 to 2016. One negative (-1) value converted to zero.
CM	-	Conditionally missing even though form required as variable dependent and cannot be sensibly imputed. (94.41% (8888) + 1.41% (9999) + 0.40% (-4))

Dependency Trigger -
Relationship -

Inspect Parent	Calculate/Derive	Impute
✗	✗	✗

General Comments Derived from NACCSTYR. No need to inspect parent as cannot be sensibly imputed.

CBTIA

Form Versions	Derived (NACC)	Derived (New)
1.2, 2.0, 3.0	✗	✗

Description Transient ischemic attack (TIA)
Data Type Categorical

Values		Comments
Original	**Replacement**	
0-2	-	-
9	M	(0.92%)
-4	M	Missing as form required and variable independent. (0.00%)

Dependency Trigger 0
Relationship -

Inspect Parent	Calculate/Derive	Impute
✗	✗	✓

General Comments -

NACCTIYR_#YRS

Form Versions	Derived (NACC)	Derived (New)
1.2, 2.0, 3.0	✗	✓

Description Years from most recently reported year of TIA as of the initial visit to year of visit
Data Type Continuous

Values		Comments
Original	**Replacement**	
Positive integer	-	Year of TIA from 1900 to 2016.
CM	-	Conditionally missing even though form required as variable dependent and cannot be sensibly imputed. (94.08% (8888) + 0.75% (9999) + 0.92% (-4))

Dependency Trigger -
Relationship -

Inspect Parent	Calculate/Derive	Impute
✗	✗	✗

General Comments Derived from NACCTIYR. No need to inspect parent as cannot be sensibly imputed.

PD

Form Versions	Derived (NACC)	Derived (New)
1.2, 2.0, 3.0	✗	✗

Description Parkinson's disease (PD)
Data Type Binary

Values		Comments
Original	**Replacement**	
0-1	-	-
9	M	(0.30%)
-4	M	Missing as form required and variable independent. (0.00%)

Dependency Trigger 0
Relationship -

Inspect Parent	Calculate/Derive	Impute
✗	✗	✓

General Comments -

PDYR_#YRS

Form Versions	Derived (NACC)	Derived (New)
1.2, 2.0, 3.0	✗	✓

Description Years from year of PD diagnosis to year of visit
Data Type Continuous

Values		Comments
Original	**Replacement**	
Positive integer	-	Year of diagnosis from 1900 to 2016. Six negative (-1) values converted to zero.
CM	-	Conditionally missing even though form required as variable dependent and cannot be sensibly imputed. (97.66% (8888) + 0.20% (9999) + 0.00% (-4))

Dependency Trigger -
Relationship -

Inspect Parent	Calculate/Derive	Impute
✗	✗	✗

General Comments Derived from PDYR. No need to inspect parent as cannot be sensibly imputed.

PDOTHR

Form Versions	Derived (NACC)	Derived (New)
1.2, 2.0, 3.0	✗	✗

Description Other parkinsonian disorder
Data Type Binary

Values		Comments
Original	**Replacement**	
0-1	-	-
9	M	(0.33%)
-4	M	Missing as form required and variable independent. (0.00%)

Dependency Trigger 0
Relationship -

Inspect Parent	Calculate/Derive	Impute
✗	✗	✓

General Comments -

PDOTHRYR_#YRS

Form Versions	Derived (NACC)	Derived (New)
1.2, 2.0, 3.0	✗	✓

Description Years from year of parkinsonian disorder diagnosis to year of visit
Data Type Continuous

Values		Comments
Original	**Replacement**	
Positive integer	-	Year of diagnosis from 1900 to 2016.
CM	-	Conditionally missing even though form required as variable dependent and cannot be sensibly imputed. (96.95% (8888) + 0.47% (9999) + 0.00% (-4))

Dependency Trigger -
Relationship -

Inspect Parent	Calculate/Derive	Impute
✗	✗	✗

General Comments Derived from PDOTHRYR. No need to inspect parent as cannot be sensibly imputed.

SEIZURES

Form Versions	Derived (NACC)	Derived (New)
1.2, 2.0, 3.0	✗	✗

Description Seizures
Data Type Categorical

Values		Comments
Original	**Replacement**	
0-2	-	-
9	M	(0.40%)
-4	M	Missing as form required and variable independent. (0.00%)

Dependency Trigger -
Relationship -

Inspect Parent	Calculate/Derive	Impute
✗	✗	✓

General Comments -

TRAUMBRF

Form Versions	Derived (NACC)	Derived (New)
1.2, 2.0	✗	✗

Description Brain trauma - brief unconsciousness
Data Type Categorical

Values		Comments
Original	**Replacement**	
0-2	-	-
9	M	(1.17%)
-4	M	Missing as form required and variable independent. (0.00%)

Dependency Trigger -
Relationship -

Inspect Parent	Calculate/Derive	Impute
✗	✗	✓

General Comments -

TRAUMEXT

Form Versions	Derived (NACC)	Derived (New)
1.2, 2.0	✗	✗

Description Brain trauma - extended unconsciousness
Data Type Categorical

Values		Comments
Original	**Replacement**	
0-2	-	-
9	M	(0.89%)
-4	M	Missing as form required and variable independent. (0.00%)

Dependency Trigger -
Relationship -

Inspect Parent	Calculate/Derive	Impute
✗	✗	✓

General Comments -

TRAUMCHR

Form Versions	Derived (NACC)	Derived (New)
1.2, 2.0	✗	✗

Description Brain trauma - chronic deficit
Data Type Categorical

Values		Comments
Original	Replacement	
0-2	-	-
9	M	(0.68%)
-4	M	Missing as form required and variable independent. (0.00%)

Dependency Trigger -
Relationship -

Inspect Parent	Calculate/Derive	Impute
✗	✗	✓

General Comments -

NCOTHR

Form Versions	Derived (NACC)	Derived (New)
1.2, 2.0	✗	✗

Description Other neurological condition
Data Type Categorical

Values		Comments
Original	Replacement	
0-2	-	-
9	M	(0.83%)
-4	M	Missing as form required and variable independent. (0.00%)

Dependency Trigger -
Relationship -

Inspect Parent	Calculate/Derive	Impute
✗	✗	✓

General Comments -

HYPERTEN

Form Versions	Derived (NACC)	Derived (New)
1.2, 2.0, 3.0	✗	✗

Description Hypertension
Data Type Categorical

Values		Comments
Original	**Replacement**	
0-2	-	-
9	M	(0.37%)
-4	M	Missing as form required and variable independent. (0.00%)

Dependency Trigger -
Relationship -

Inspect Parent	Calculate/Derive	Impute
✗	✗	✓

General Comments -

HYPERCHO

Form Versions	Derived (NACC)	Derived (New)
1.2, 2.0, 3.0	✗	✗

Description Hypercholesterolemia
Data Type Categorical

Values		Comments
Original	**Replacement**	
0-2	-	-
9	M	(1.25%)
-4	M	Missing as form required and variable independent. (0.00%)

Dependency Trigger -
Relationship -

Inspect Parent	Calculate/Derive	Impute
✗	✗	✓

General Comments -

DIABETES

Form Versions	Derived (NACC)	Derived (New)
1.2, 2.0, 3.0	✗	✗

Description Diabetes
Data Type Categorical

Values		Comments
Original	**Replacement**	
0-2	-	-
9	M	(0.38%)
-4	M	Missing as form required and variable independent. (0.00%)

Dependency Trigger -
Relationship -

Inspect Parent	Calculate/Derive	Impute
✗	✗	✓

General Comments -

B12DEF

Form Versions	Derived (NACC)	Derived (New)
1.2, 2.0, 3.0	✗	✗

Description Vitamin B12 deficiency
Data Type Categorical

Values		Comments
Original	**Replacement**	
0-2	-	-
9	M	(2.04%)
-4	M	Missing as form required and variable independent. (0.00%)

Dependency Trigger -
Relationship -

Inspect Parent	Calculate/Derive	Impute
✗	✗	✓

General Comments -

THYROID

Form Versions	Derived (NACC)	Derived (New)
1.2, 2.0, 3.0	✗	✗

Description Thyroid disease
Data Type Categorical

Values		Comments
Original	**Replacement**	
0-2	-	-
9	M	(0.84%)
-4	M	Missing as form required and variable independent. (0.00%)

Dependency Trigger -
Relationship -

Inspect Parent	Calculate/Derive	Impute
✗	✗	✓

General Comments -

INCONTU

Form Versions	Derived (NACC)	Derived (New)
1.2, 2.0, 3.0	✗	✗

Description Incontinence - urinary
Data Type Categorical

Values		Comments
Original	**Replacement**	
0-2	-	-
9	M	(0.29%)
-4	M	Missing as form required and variable independent. (0.00%)

Dependency Trigger -
Relationship -

Inspect Parent	Calculate/Derive	Impute
✗	✗	✓

General Comments -

INCONTF

Form Versions	Derived (NACC)	Derived (New)
1.2, 2.0, 3.0	✗	✗

Description Incontinence - bowel
Data Type Categorical

Values		Comments
Original	**Replacement**	
0-2	-	-
9	M	(0.27%)
-4	M	Missing as form required and variable independent. (0.00%)

Dependency Trigger -
Relationship -

Inspect Parent	Calculate/Derive	Impute
✗	✗	✓

General Comments -

DEP2YRS

Form Versions	Derived (NACC)	Derived (New)
1.2, 2.0, 3.0	✗	✗

Description Active depression in the last two years
Data Type Binary

Values		Comments
Original	**Replacement**	
0-1	-	-
9	M	(0.76%)
-4	M	Missing as form required and variable independent. (0.00%)

Dependency Trigger -
Relationship -

Inspect Parent	Calculate/Derive	Impute
✗	✗	✓

General Comments -

DEPOTHR

Form Versions	Derived (NACC)	Derived (New)
1.2, 2.0, 3.0	✗	✗

Description Depression episodes more than two years ago
Data Type Binary

Values		Comments
Original	Replacement	
0-1	-	-
9	M	(1.72%)
-4	M	Missing as form required and variable independent. (0.00%)

Dependency Trigger -
Relationship -

Inspect Parent	Calculate/Derive	Impute
✗	✗	✓

General Comments -

ALCOHOL

Form Versions	Derived (NACC)	Derived (New)
1.2, 2.0, 3.0	✗	✗

Description Alcohol abuse - clinically significant impairment occurring over a 12-month period manifested in one of the following areas: work, driving, legal or social
Data Type Categorical

Values		Comments
Original	Replacement	
0-2	-	-
9	M	(0.37%)
-4	M	Missing as form required and variable independent. (0.00%)

Dependency Trigger -
Relationship -

Inspect Parent	Calculate/Derive	Impute
✗	✗	✓

General Comments -

TOBAC30

Form Versions	Derived (NACC)	Derived (New)
1.2, 2.0, 3.0	✗	✗

Description Smoked cigarettes in last 30 days
Data Type Binary

Values		Comments
Original	Replacement	
0-1	-	-
9	M	(0.62%)
-4	M	Missing as form required and variable independent. (0.00%)

Dependency Trigger -
Relationship -

Inspect Parent	Calculate/Derive	Impute
✗	✗	✓

General Comments -

TOBAC100

Form Versions	Derived (NACC)	Derived (New)
1.2, 2.0, 3.0	✗	✗

Description Smoked more than 100 cigarettes in life
Data Type Binary

Values		Comments
Original	Replacement	
0-1	-	-
9	M	(1.36%)
-4	M	Missing as form required and variable independent. (0.00%)

Dependency Trigger -
Relationship -

Inspect Parent	Calculate/Derive	Impute
✗	✗	✓

General Comments -

SMOKYRS

Form Versions	Derived (NACC)	Derived (New)
1.2, 2.0, 3.0	✗	✗

Description Total years smoked cigarettes
Data Type Continuous

Values		Comments
Original	**Replacement**	
0-87	-	-
88	CM	(0.26%)
99	M	(3.47%)
-4	M	Missing as form required and variable independent. (0.00%)

Dependency Trigger -
Relationship -

Inspect Parent	Calculate/Derive	Impute
✗	✗	✓

General Comments NACC's coding guidebooks state variable is dependent on TOBAC100, but dependency does not hold and is omitted from NACC's data element dictionaries for versions 1.2 and 2.0. A conditionally missing value should be used as a potential fill value if imputed.

PACKSPER

Form Versions	Derived (NACC)	Derived (New)
1.2, 2.0, 3.0	✗	✗

Description Average number of packs smoked per day
Data Type Ordinal

Values		Comments
Original	**Replacement**	
0	CM	Omitted from original forms and NACC's coding guidebooks. (53.66%)
1-5	-	-
8	CM	(1.34%)
9	M	(2.59%)
-4	M	Missing as form required and variable independent. (0.00%)

Dependency Trigger -
Relationship -

Inspect Parent	Calculate/Derive	Impute
✗	✗	✓

General Comments NACC's coding guidebooks state variable is dependent on TOBAC100, and NACC's researchers data dictionary states variable is dependent on TOBAC100 and SMOKYRS, but dependencies do not hold and are omitted from NACC's data element dictionaries for versions 1.2 and 2.0. A conditionally missing value should be used as a potential fill value if imputed.

QUITSMOK

Form Versions	Derived (NACC)	Derived (New)
1.2, 2.0, 3.0	✗	✗

Description If the subject quit smoking, age at which he/she last smoked (i.e. quit)
Data Type Continuous

Values		Comments
Original	**Replacement**	
7-110	-	-
888	CM	Indicates no significant smoking history or subject still smokes. (57.50%)
999	M	(3.87%)
-4	M	Missing as form required and variable independent. (0.00%)

Dependency Trigger -
Relationship -

Inspect Parent	Calculate/Derive	Impute
✗	✗	✓

General Comments NACC's coding guidebooks state variable is dependent on TOBAC100, but dependency does not hold and is omitted from NACC's data element dictionaries for versions 1.2 and 2.0. A conditionally missing value should be used as a potential fill value if imputed.

ABUSOTHR

Form Versions	Derived (NACC)	Derived (New)
1.2, 2.0, 3.0	✗	✗

Description Other abused substances - clinically significant impairment occurring over a 12-month period manifested in one of the following areas: work, driving, legal or social
Data Type Categorical

Values		Comments
Original	**Replacement**	
0-2	-	-
9	M	(0.36%)
-4	M	Missing as form required and variable independent. (0.00%)

Dependency Trigger -
Relationship -

Inspect Parent	Calculate/Derive	Impute
✗	✗	✓

General Comments -

PSYCDIS

Form Versions	Derived (NACC)	Derived (New)
1.2, 2.0, 3.0	✗	✗

Description Other psychiatric disorder
Data Type Categorical

Values		Comments
Original	**Replacement**	
0-2	-	-
9	M	(0.45%)
-4	M	Missing as form required and variable independent. (0.00%)

Dependency Trigger -
Relationship -

Inspect Parent	Calculate/Derive	Impute
✗	✗	✓

General Comments -

NACCTBI

Form Versions	Derived (NACC)	Derived (New)
1.2, 2.0, 3.0	✓	✗

Description History of traumatic brain injury (TBI)
Data Type Binary

Values		Comments
Original	**Replacement**	
0-1	-	-
9	M	(1.20%)
-4	M	Missing as form required. (0.00%)

Dependency Trigger -
Relationship NACCTBI = 1 if TRAUMBRF, TRAUMEXT and/or TRAUMCHR = 1 or 2 else 0 (NACCTBI = M if all M or all M and 0)

Inspect Parent	Calculate/Derive	Impute
✗	✓	✗

General Comments Values should be updated if TRAUMBRF, TRAUMEXT and/or TRAUMCHR imputed.

Dropped Variables

Name CVOTHRX
Description Other cardiovascular disease - specify
Reason Free-text variable and, even though present in original forms and NACC's coding guidebooks for all versions, NACC's researchers data dictionary states only available in version 3.0 of the UDS.

Name NACCSTYR
Description Most recently reported year of stroke as of the initial visit
Reason Replaced by the derived variable NACCSTYR_#YRS.

Name NACCTIYR
Description Most recently reported year of TIA as of the initial visit
Reason Replaced by the derived variable NACCTIYR_#YRS.

Name PDYR
Description Year of PD diagnosis
Reason Replaced by the derived variable PDYR_#YRS.

Name PDOTHRYR
Description Year of parkinsonian disorder diagnosis
Reason Replaced by the derived variable PDOTHRYR_#YRS.

Name NCOTHRX
Description Other neurological condition - specify
Reason Free-text variable.

Name ABUSX
Description If reported other abused substances - specify abused substance(s)
Reason Free-text variable.

Name PSYCDISX
Description If recent/active or remote/inactive psychiatric disorder - specify disorder
Reason Free-text variable.

Name HATTMULT
Description More than one heart attack/cardiac arrest?
Reason Only available in version 3.0 of the UDS.

Name HATTYEAR
Description Year of most recent heart attack
Reason Only available in version 3.0 of the UDS.

Name CVPACDEF
Description Pacemaker and/or defibrillator
Reason Only available in version 3.0 of the UDS.

Name CVANGINA
Description Angina
Reason Only available in version 3.0 of the UDS.

Name CVHVALVE
Description Heart valve replacement or repair
Reason Only available in version 3.0 of the UDS.

Name STROKMUL
Description More than one stroke reported as of the initial visit

Reason Only available in version 3.0 of the UDS.

Name TIAMULT
Description More than one TIA reported as of the initial visit
Reason Only available in version 3.0 of the UDS.

Name TBI
Description Traumatic brain injury (TBI)
Reason Only available in version 3.0 of the UDS.

Name TBIBRIEF
Description TBI with brief loss of consciousness
Reason Only available in version 3.0 of the UDS.

Name TBIEXTEN
Description TBI with extended loss of consciousness - 5 minutes or longer
Reason Only available in version 3.0 of the UDS.

Name TBIWOLOS
Description TBI without loss of consciousness - as might result from military detonations or sports injury
Reason Only available in version 3.0 of the UDS.

Name TBIYEAR
Description Year of most recent TBI
Reason Only available in version 3.0 of the UDS.

Name DIABTYPE
Description If recent/active or remote/inactive diabetes, which type?
Reason Only available in version 3.0 of the UDS.

Name ARTHRIT
Description Arthritis
Reason Only available in version 3.0 of the UDS.

Name ARTHTYPE
Description Type of arthritis
Reason Only available in version 3.0 of the UDS.

Name ARTHTYPX
Description Other arthritis - specify
Reason Free-text variable and only available in version 3.0 of the UDS.

Name ARTHUPEX
Description Arthritis, region affected - upper extremity
Reason Only available in version 3.0 of the UDS.

Name ARTHLOEX
Description Arthritis, region affected - lower extremity
Reason Only available in version 3.0 of the UDS.

Name ARTHSPIN
Description Arthritis, region affected - spine
Reason Only available in version 3.0 of the UDS.

Name ARTHUNK
Description Arthritis, region affected - unknown
Reason Only available in version 3.0 of the UDS.

Name APNEA

Description Sleep apnea history reported at initial visit
Reason Only available in version 3.0 of the UDS.

Name RBD
Description REM sleep behaviour disorder (RBD) history reported at initial visit
Reason Only available in version 3.0 of the UDS.

Name INSOMN
Description Hyposomnia/insomnia history reported at initial visit
Reason Only available in version 3.0 of the UDS.

Name OTHSLEEP
Description Other sleep disorder history reported at initial visit
Reason Only available in version 3.0 of the UDS.

Name OTHSLEEX
Description Other sleep disorder - specify
Reason Free-text variable and only available in version 3.0 of the UDS.

Name ALCOCCAS
Description In the past three months, has the subject consumed any alcohol?
Reason Only available in version 3.0 of the UDS.

Name ALCFREQ
Description During the past three months, how often did the subject have at least one drink of any alcoholic beverage such as wine, beer, malt liquor or spirits?
Reason Only available in version 3.0 of the UDS.

Name PTSD
Description Post-traumatic stress disorder (PTSD)
Reason Only available in version 3.0 of the UDS.

Name BIPOLAR
Description Bipolar disorder
Reason Only available in version 3.0 of the UDS.

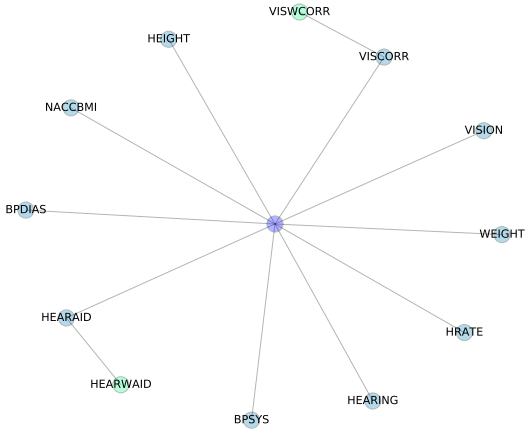
Name SCHIZ
Description Schizophrenia
Reason Only available in version 3.0 of the UDS.

Name ANXIETY
Description Anxiety
Reason Only available in version 3.0 of the UDS.

Name OCD
Description Obsessive-compulsive disorder (OCD)
Reason Only available in version 3.0 of the UDS.

Name NPSYDEV
Description Developmental neuropsychiatric disorders (e.g. autism spectrum disorder (ASD), attention-deficit hyperactivity disorder (ADHD), dyslexia)
Reason Only available in version 3.0 of the UDS.

B1 - Physical Examination



Number of Variables	12
Number of Variables Used	12
Form Required?	✗
Form Missingness	1.42%

HEIGHT

Form Versions	Derived (NACC)	Derived (New)
1.2, 2.0, 3.0	✗	✗

Description Subject's height (inches)
Data Type Continuous

Values		Comments
Original	**Replacement**	
36.0-87.9	-	-
88.8	M	99.9 in original forms and NACC's coding guidebooks for versions 1.2 and 2.0. (9.63%)
-4.0	CM	-4.0 rather than -4 due to data type. Conditionally missing as form not required. (1.42%)

Dependency Trigger -
Relationship -

Inspect Parent	Calculate/Derive	Impute
✗	✗	✓

General Comments A conditionally missing value should not be used as a potential fill value if imputed.

WEIGHT

Form Versions	Derived (NACC)	Derived (New)
1.2, 2.0, 3.0	✗	✗

Description Subject's weight (lbs)
Data Type Continuous

Values		Comments
Original	**Replacement**	
50-400	-	-
888	M	999 in original forms and NACC's coding guidebooks for versions 1.2 and 2.0. (7.08%)
-4	CM	Conditionally missing as form not required. (1.42%)

Dependency Trigger -
Relationship -

Inspect Parent	Calculate/Derive	Impute
✗	✗	✓

General Comments A conditionally missing value should not be used as a potential fill value if imputed.

BPSYS

Form Versions	Derived (NACC)	Derived (New)
1.2, 2.0, 3.0	✗	✗

Description Subject blood pressure (sitting), systolic
Data Type Continuous

Values		Comments
Original	**Replacement**	
70-230	-	-
888	M	999 in original forms and NACC's coding guidebooks for versions 1.2 and 2.0. (7.62%)
-4	CM	Conditionally missing as form not required. (1.42%)

Dependency Trigger -
Relationship -

Inspect Parent	Calculate/Derive	Impute
✗	✗	✓

General Comments A conditionally missing value should not be used as a potential fill value if imputed.

BPDIAS

Form Versions	Derived (NACC)	Derived (New)
1.2, 2.0, 3.0	✗	✗

Description Subject blood pressure (sitting), diastolic
Data Type Continuous

Values		Comments
Original	**Replacement**	
30-140	-	-
888	M	999 in original forms and NACC's coding guidebooks for versions 1.2 and 2.0. (7.64%)
-4	CM	Conditionally missing as form not required. (1.42%)

Dependency Trigger -
Relationship -

Inspect Parent	Calculate/Derive	Impute
✗	✗	✓

General Comments A conditionally missing value should not be used as a potential fill value if imputed.

HRATE

Form Versions	Derived (NACC)	Derived (New)
1.2, 2.0, 3.0	✗	✗

Description Subject resting heart rate (pulse)
Data Type Continuous

Values		Comments
Original	**Replacement**	
33-160	-	-
888	M	999 in original forms and NACC's coding guidebooks for versions 1.2 and 2.0. (8.08%)
-4	CM	Conditionally missing as form not required. (1.42%)

Dependency Trigger -
Relationship -

Inspect Parent	Calculate/Derive	Impute
✗	✗	✓

General Comments A conditionally missing value should not be used as a potential fill value if imputed.

VISION

Form Versions	Derived (NACC)	Derived (New)
1.2, 2.0, 3.0	✗	✗

Description Without corrective lenses, is the subject's vision functionally normal?
Data Type Binary

Values		Comments
Original	**Replacement**	
0-1	-	-
9	M	(2.12%)
-4	CM	Conditionally missing as form not required. (1.42%)

Dependency Trigger -
Relationship -

Inspect Parent	Calculate/Derive	Impute
✗	✗	✓

General Comments A conditionally missing value should not be used as a potential fill value if imputed.

VISCORR

Form Versions	Derived (NACC)	Derived (New)
1.2, 2.0, 3.0	✗	✗

Description Does the subject usually wear corrective lenses?
Data Type Binary

Values		Comments
Original	Replacement	
0-1	-	-
9	CM	Conditionally missing as cannot be sensibly imputed. (1.70%)
-4	CM	Conditionally missing as form not required. (1.42%)

Dependency Trigger 0
Relationship -

Inspect Parent	Calculate/Derive	Impute
✗	✗	✗

General Comments -

VISWCORR

Form Versions	Derived (NACC)	Derived (New)
1.2, 2.0, 3.0	✗	✗

Description If the subject usually wears corrective lenses, is the subject's vision functionally normal with corrective lenses?
Data Type Binary

Values		Comments
Original	Replacement	
0-1	-	-
8	CM	554 values converted from -4 to 8 to indicate dependence on VISCORR. (24.06%)
9	CM	Conditionally missing as cannot be sensibly imputed. (1.19%)
-4	CM	Conditionally missing as form not required. (1.42%)

Dependency Trigger -
Relationship -

Inspect Parent	Calculate/Derive	Impute
✗	✗	✗

General Comments No need to inspect parent as parent and child cannot be sensibly imputed.

HEARING

Form Versions	Derived (NACC)	Derived (New)
1.2, 2.0, 3.0	✗	✗

Description Without a hearing aid(s), is the subject's hearing functionally normal?
Data Type Binary

Values		Comments
Original	Replacement	
0-1	-	-
9	M	(1.65%)
-4	CM	Conditionally missing as form not required. (1.42%)

Dependency Trigger -
Relationship -

Inspect Parent	Calculate/Derive	Impute
✗	✗	✓

General Comments A conditionally missing value should not be used as a potential fill value if imputed.

HEARAID

Form Versions	Derived (NACC)	Derived (New)
1.2, 2.0, 3.0	✗	✗

Description Does the subject usually wear a hearing aid(s)?
Data Type Binary

Values		Comments
Original	Replacement	
0-1	-	-
9	CM	Conditionally missing as cannot be sensibly imputed. (1.69%)
-4	CM	Conditionally missing as form not required. (1.42%)

Dependency Trigger 0
Relationship -

Inspect Parent	Calculate/Derive	Impute
✗	✗	✗

General Comments -

HEARWAID

Form Versions	Derived (NACC)	Derived (New)
1.2, 2.0, 3.0	✗	✗

Description If the subject usually wears a hearing aid(s), is the subject's hearing functionally normal with a hearing aid(s)?
Data Type Binary

Values		Comments
Original	Replacement	
0-1	-	-
8	CM	550 values converted from -4 to 8 to indicate dependence on HEARAID. (86.03%)
9	CM	Conditionally missing as cannot be sensibly imputed. (0.41%)
-4	CM	Conditionally missing as form not required. (1.42%)

Dependency Trigger -
Relationship -

Inspect Parent	Calculate/Derive	Impute
✗	✗	✗

General Comments No need to inspect parent as parent and child cannot be sensibly imputed.

NACCBMI

Form Versions	Derived (NACC)	Derived (New)
1.2, 2.0, 3.0	✓	✗

Description Body mass index (BMI)
Data Type Continuous

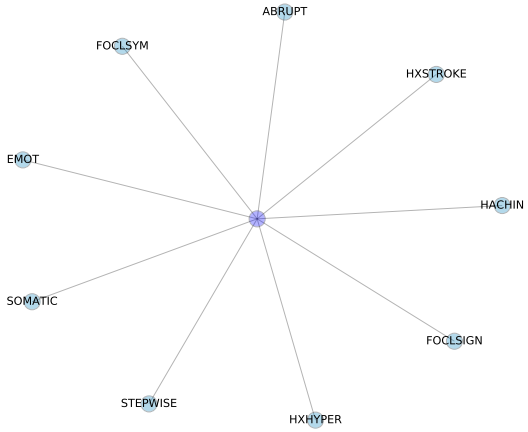
Values		Comments
Original	Replacement	
10.0-100.0	-	-
888.8	M	(10.55%)
-4.0	CM	-4.0 rather than -4 due to data type. Conditionally missing as form not required. (1.42%)

Dependency Trigger -
Relationship NACCBMI = (WEIGHT × 703)/HEIGHT^2 if HEIGHT and WEIGHT not M (NACCBMI = M if any M)

Inspect Parent	Calculate/Derive	Impute
✗	✓	✗

General Comments Values should be updated if HEIGHT and/or WEIGHT imputed. Ensure any calculated values are within the allowed range.

B2 - Hachinski Ischemic Score and Cerebrovascular Disease



Number of Variables	17
Number of Variables Used	9
Form Required?	✗
Form Missingness	2.27%

ABRUPT

Form Versions	Derived (NACC)	Derived (New)
1.2, 2.0	✗	✗

Description Abrupt onset (re: cognitive status)
Data Type Binary

Values		Comments
Original	**Replacement**	
0	-	-
2	-	Binary due to number of options available.
-4	CM	Conditionally missing as form not required. (2.27%)

Dependency Trigger -
Relationship -

Inspect Parent	Calculate/Derive	Impute
✗	✗	✗

General Comments -

STEPWISE

Form Versions	Derived (NACC)	Derived (New)
1.2, 2.0	✗	✗

Description Stepwise deterioration (re: cognitive status)
Data Type Binary

Values		Comments
Original	**Replacement**	
0-1	-	-
-4	CM	Conditionally missing as form not required. (2.27%)

Dependency Trigger -
Relationship -

Inspect Parent	Calculate/Derive	Impute
✗	✗	✗

General Comments -

SOMATIC

Form Versions	Derived (NACC)	Derived (New)
1.2, 2.0	✗	✗

Description Somatic complaints
Data Type Binary

Values		Comments
Original	**Replacement**	
0-1	-	-
-4	CM	Conditionally missing as form not required. (2.27%)

Dependency Trigger -
Relationship -

Inspect Parent	Calculate/Derive	Impute
✗	✗	✗

General Comments -

EMOT

Form Versions	Derived (NACC)	Derived (New)
1.2, 2.0	✗	✗

Description Emotional incontinence
Data Type Binary

Values		Comments
Original	**Replacement**	
0-1	-	-
-4	CM	Conditionally missing as form not required. (2.27%)

Dependency Trigger -
Relationship -

Inspect Parent	Calculate/Derive	Impute
✗	✗	✗

General Comments -

HXHYPER

Form Versions	Derived (NACC)	Derived (New)
1.2, 2.0	✗	✗

Description History or presence of hypertension
Data Type Binary

Values		Comments
Original	Replacement	
0-1	-	-
-4	CM	Conditionally missing as form not required. (2.27%)

Dependency Trigger -
Relationship -

Inspect Parent	Calculate/Derive	Impute
✗	✗	✗

General Comments -

HXSTROKE

Form Versions	Derived (NACC)	Derived (New)
1.2, 2.0	✗	✗

Description History of stroke
Data Type Binary

Values		Comments
Original	Replacement	
0	-	-
2	-	Binary due to number of options available.
-4	CM	Conditionally missing as form not required. (2.27%)

Dependency Trigger -
Relationship -

Inspect Parent	Calculate/Derive	Impute
✗	✗	✗

General Comments -

FOCLSYM

Form Versions	Derived (NACC)	Derived (New)
1.2, 2.0	✗	✗

Description Focal neurological symptoms
Data Type Binary

Values		Comments
Original	**Replacement**	
0	-	-
2	-	Binary due to number of options available.
-4	CM	Conditionally missing as form not required. (2.27%)

Dependency Trigger -
Relationship -

Inspect Parent	Calculate/Derive	Impute
✗	✗	✗

General Comments -

FOCLSIGN

Form Versions	Derived (NACC)	Derived (New)
1.2, 2.0	✗	✗

Description Focal neurological signs
Data Type Binary

Values		Comments
Original	**Replacement**	
0	-	-
2	-	Binary due to number of options available.
-4	CM	Conditionally missing as form not required. (2.27%)

Dependency Trigger -
Relationship -

Inspect Parent	Calculate/Derive	Impute
✗	✗	✗

General Comments -

HACHIN

Form Versions	Derived (NACC)	Derived (New)
1.2, 2.0	✗	✗

Description Hachinski ischemic score
Data Type Continuous

Values		Comments
Original	**Replacement**	
0-12	-	-
-4	CM	Conditionally missing as form not required. (2.27%)

Dependency Trigger Relationship -
HACHIN = sum of ABRUPT, STEPWISE, SOMATIC, EMOT, HXHYPER, HXSTROKE, FOCLSYM and FOCLSIGN

Inspect Parent	Calculate/Derive	Impute
✗	✗	✗

General Comments No need to calculate/derive variable as no missing values in any of the variables involved.

Dropped Variables

Name CVDCOG
Description Cerebrovascular disease contributing to cognitive impairment
Reason Only available in version 2.0 of the UDS.

Name STROKCOG
Description Relationship between stroke and cognitive impairment
Reason Only available in version 2.0 of the UDS.

Name CVDIMAG
Description Imaging evidence
Reason Only available in version 2.0 of the UDS.

Name CVDIMAG1
Description Single strategic infarct
Reason Only available in version 2.0 of the UDS.

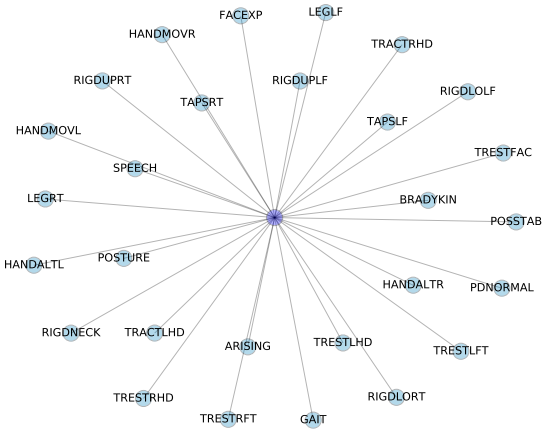
Name CVDIMAG2
Description Multiple infarcts
Reason Only available in version 2.0 of the UDS.

Name CVDIMAG3
Description Extensive white matter hyperintensity
Reason Only available in version 2.0 of the UDS.

Name CVDIMAG4
Description Other imaging evidence
Reason Only available in version 2.0 of the UDS.

Name CVDIMAGX
Description Other imaging evidence - specify
Reason Free-text variable and only available in version 2.0 of the UDS.

B3 - Unified Parkinson's Disease Rating Scale



Number of Variables	55
Number of Variables Used	28
Form Required?	✗
Form Missingness	2.58%

SPEECH

Form Versions	Derived (NACC)	Derived (New)
1.2, 2.0	✗	✗

Description Speech
Data Type Ordinal

Values		Comments
Original	**Replacement**	
0-4	-	-
8	CM	Conditionally missing as value omitted for a reason. (0.30%)
-4	CM	Conditionally missing as form not required. (2.58%)

Dependency Trigger -
Relationship -

Inspect Parent	Calculate/Derive	Impute
✗	✗	✗

General Comments -

FACEXP

Form Versions	Derived (NACC)	Derived (New)
1.2, 2.0	✗	✗

Description Facial expression
Data Type Ordinal

Values		Comments
Original	**Replacement**	
0-4	-	-
8	CM	Conditionally missing as value omitted for a reason. (0.07%)
-4	CM	Conditionally missing as form not required. (2.58%)

Dependency Trigger -
Relationship -

Inspect Parent	Calculate/Derive	Impute
✗	✗	✗

General Comments -

TRESTFAC

Form Versions	Derived (NACC)	Derived (New)
1.2, 2.0	✗	✗

Description Tremor at rest - face, lips, chin
Data Type Ordinal

Values		Comments
Original	**Replacement**	
0-4	-	-
8	CM	Conditionally missing as value omitted for a reason. (0.02%)
-4	CM	Conditionally missing as form not required. (2.58%)

Dependency Trigger -
Relationship -

Inspect Parent	Calculate/Derive	Impute
✗	✗	✗

General Comments -

TRESTRHD

Form Versions	Derived (NACC)	Derived (New)
1.2, 2.0	✗	✗

Description Tremor at rest - right hand
Data Type Ordinal

Values		Comments
Original	**Replacement**	
0-4	-	-
8	CM	Conditionally missing as value omitted for a reason. (0.03%)
-4	CM	Conditionally missing as form not required. (2.58%)

Dependency Trigger -
Relationship -

Inspect Parent	Calculate/Derive	Impute
✗	✗	✗

General Comments -

TRESTLHD

Form Versions	Derived (NACC)	Derived (New)
1.2, 2.0	✗	✗

Description Tremor at rest - left hand
Data Type Ordinal

Values		Comments
Original	**Replacement**	
0-4	-	-
8	CM	Conditionally missing as value omitted for a reason. (0.04%)
-4	CM	Conditionally missing as form not required. (2.58%)

Dependency Trigger -
Relationship -

Inspect Parent	Calculate/Derive	Impute
✗	✗	✗

General Comments -

TRESTRFT

Form Versions	Derived (NACC)	Derived (New)
1.2, 2.0	✗	✗

Description Tremor at rest - right foot
Data Type Ordinal

Values		Comments
Original	**Replacement**	
0-4	-	-
8	CM	Conditionally missing as value omitted for a reason. (0.04%)
-4	CM	Conditionally missing as form not required. (2.58%)

Dependency Trigger -
Relationship -

Inspect Parent	Calculate/Derive	Impute
✗	✗	✗

General Comments -

TRESTLFT

Form Versions	Derived (NACC)	Derived (New)
1.2, 2.0	✗	✗

Description Tremor at rest - left foot
Data Type Ordinal

Values		Comments
Original	**Replacement**	
0-4	-	-
8	CM	Conditionally missing as value omitted for a reason. (0.05%)
-4	CM	Conditionally missing as form not required. (2.58%)

Dependency Trigger -
Relationship -

Inspect Parent	Calculate/Derive	Impute
✗	✗	✗

General Comments -

TRACTRHD

Form Versions	Derived (NACC)	Derived (New)
1.2, 2.0	✗	✗

Description Action or postural tremor - right hand
Data Type Ordinal

Values		Comments
Original	**Replacement**	
0-4	-	-
8	CM	Conditionally missing as value omitted for a reason. (0.23%)
-4	CM	Conditionally missing as form not required. (2.58%)

Dependency Trigger -
Relationship -

Inspect Parent	Calculate/Derive	Impute
✗	✗	✗

General Comments -

TRACTLHD

Form Versions	Derived (NACC)	Derived (New)
1.2, 2.0	✗	✗

Description Action or postural tremor - left hand
Data Type Ordinal

Values		Comments
Original	**Replacement**	
0-4	-	-
8	CM	Conditionally missing as value omitted for a reason. (0.26%)
-4	CM	Conditionally missing as form not required. (2.58%)

Dependency Trigger -
Relationship -

Inspect Parent	Calculate/Derive	Impute
✗	✗	✗

General Comments -

RIGDNECK

Form Versions	Derived (NACC)	Derived (New)
1.2, 2.0	✗	✗

Description Rigidity - neck
Data Type Ordinal

Values		Comments
Original	**Replacement**	
0-4	-	-
8	CM	Conditionally missing as value omitted for a reason. (0.23%)
-4	CM	Conditionally missing as form not required. (2.58%)

Dependency Trigger -
Relationship -

Inspect Parent	Calculate/Derive	Impute
✗	✗	✗

General Comments -

RIGDUPRT

Form Versions	Derived (NACC)	Derived (New)
1.2, 2.0	✗	✗

Description Rigidity - right upper extremity
Data Type Ordinal

Values		Comments
Original	**Replacement**	
0-4	-	-
8	CM	Conditionally missing as value omitted for a reason. (0.19%)
-4	CM	Conditionally missing as form not required. (2.58%)

Dependency Trigger -
Relationship -

Inspect Parent	Calculate/Derive	Impute
✗	✗	✗

General Comments -

RIGDUPLF

Form Versions	Derived (NACC)	Derived (New)
1.2, 2.0	✗	✗

Description Rigidity - left upper extremity
Data Type Ordinal

Values		Comments
Original	**Replacement**	
0-4	-	-
8	CM	Conditionally missing as value omitted for a reason. (0.20%)
-4	CM	Conditionally missing as form not required. (2.58%)

Dependency Trigger -
Relationship -

Inspect Parent	Calculate/Derive	Impute
✗	✗	✗

General Comments -

RIGDLORT

Form Versions	Derived (NACC)	Derived (New)
1.2, 2.0	✗	✗

Description Rigidity - right lower extremity
Data Type Ordinal

Values		Comments
Original	Replacement	
0-4	-	-
8	CM	Conditionally missing as value omitted for a reason. (0.26%)
-4	CM	Conditionally missing as form not required. (2.58%)

Dependency Trigger -
Relationship -

Inspect Parent	Calculate/Derive	Impute
✗	✗	✗

General Comments -

RIGDLOLF

Form Versions	Derived (NACC)	Derived (New)
1.2, 2.0	✗	✗

Description Rigidity - left lower extremity
Data Type Ordinal

Values		Comments
Original	Replacement	
0-4	-	-
8	CM	Conditionally missing as value omitted for a reason. (0.27%)
-4	CM	Conditionally missing as form not required. (2.58%)

Dependency Trigger -
Relationship -

Inspect Parent	Calculate/Derive	Impute
✗	✗	✗

General Comments -

TAPSRT

Form Versions	Derived (NACC)	Derived (New)
1.2, 2.0	✗	✗

Description Finger taps - right hand
Data Type Ordinal

Values		Comments
Original	**Replacement**	
0-4	-	-
8	CM	Conditionally missing as value omitted for a reason. (1.76%)
-4	CM	Conditionally missing as form not required. (2.58%)

Dependency Trigger -
Relationship -

Inspect Parent	Calculate/Derive	Impute
✗	✗	✗

General Comments -

TAPSLF

Form Versions	Derived (NACC)	Derived (New)
1.2, 2.0	✗	✗

Description Finger taps - left hand
Data Type Ordinal

Values		Comments
Original	**Replacement**	
0-4	-	-
8	CM	Conditionally missing as value omitted for a reason. (1.78%)
-4	CM	Conditionally missing as form not required. (2.58%)

Dependency Trigger -
Relationship -

Inspect Parent	Calculate/Derive	Impute
✗	✗	✗

General Comments -

HANDMOVR

Form Versions	Derived (NACC)	Derived (New)
1.2, 2.0	✗	✗

Description Hand movements - right hand
Data Type Ordinal

Values		Comments
Original	**Replacement**	
0-4	-	-
8	CM	Conditionally missing as value omitted for a reason. (1.80%)
-4	CM	Conditionally missing as form not required. (2.58%)

Dependency Trigger -
Relationship -

Inspect Parent	Calculate/Derive	Impute
✗	✗	✗

General Comments -

HANDMOVL

Form Versions	Derived (NACC)	Derived (New)
1.2, 2.0	✗	✗

Description Hand movements - left hand
Data Type Ordinal

Values		Comments
Original	**Replacement**	
0-4	-	-
8	CM	Conditionally missing as value omitted for a reason. (1.80%)
-4	CM	Conditionally missing as form not required. (2.58%)

Dependency Trigger -
Relationship -

Inspect Parent	Calculate/Derive	Impute
✗	✗	✗

General Comments -

HANDALTR

Form Versions	Derived (NACC)	Derived (New)
1.2, 2.0	✗	✗

Description Alternating movement - right hand
Data Type Ordinal

Values		Comments
Original	**Replacement**	
0-4	-	-
8	CM	Conditionally missing as value omitted for a reason. (2.07%)
-4	CM	Conditionally missing as form not required. (2.58%)

Dependency Trigger -
Relationship -

Inspect Parent	Calculate/Derive	Impute
✗	✗	✗

General Comments -

HANDALTL

Form Versions	Derived (NACC)	Derived (New)
1.2, 2.0	✗	✗

Description Alternating movement - left hand
Data Type Ordinal

Values		Comments
Original	**Replacement**	
0-4	-	-
8	CM	Conditionally missing as value omitted for a reason. (2.10%)
-4	CM	Conditionally missing as form not required. (2.58%)

Dependency Trigger -
Relationship -

Inspect Parent	Calculate/Derive	Impute
✗	✗	✗

General Comments -

LEGRT

Form Versions	Derived (NACC)	Derived (New)
1.2, 2.0	✗	✗

Description Leg agility - right leg
Data Type Ordinal

Values		Comments
Original	**Replacement**	
0-4	-	-
8	CM	Conditionally missing as value omitted for a reason. (2.17%)
-4	CM	Conditionally missing as form not required. (2.58%)

Dependency Trigger -
Relationship -

Inspect Parent	Calculate/Derive	Impute
✗	✗	✗

General Comments -

LEGLF

Form Versions	Derived (NACC)	Derived (New)
1.2, 2.0	✗	✗

Description Leg agility - left leg
Data Type Ordinal

Values		Comments
Original	**Replacement**	
0-4	-	-
8	CM	Conditionally missing as value omitted for a reason. (2.20%)
-4	CM	Conditionally missing as form not required. (2.58%)

Dependency Trigger -
Relationship -

Inspect Parent	Calculate/Derive	Impute
✗	✗	✗

General Comments -

ARISING

Form Versions	Derived (NACC)	Derived (New)
1.2, 2.0	✗	✗

Description Arising from chair
Data Type Ordinal

Values		Comments
Original	**Replacement**	
0-4	-	-
8	CM	Conditionally missing as value omitted for a reason. (1.23%)
-4	CM	Conditionally missing as form not required. (2.58%)

Dependency Trigger -
Relationship -

Inspect Parent	Calculate/Derive	Impute
✗	✗	✗

General Comments -

POSTURE

Form Versions	Derived (NACC)	Derived (New)
1.2, 2.0	✗	✗

Description Posture
Data Type Ordinal

Values		Comments
Original	**Replacement**	
0-4	-	-
8	CM	Conditionally missing as value omitted for a reason. (1.07%)
-4	CM	Conditionally missing as form not required. (2.58%)

Dependency Trigger -
Relationship -

Inspect Parent	Calculate/Derive	Impute
✗	✗	✗

General Comments -

GAIT

Form Versions	Derived (NACC)	Derived (New)
1.2, 2.0	✗	✗

Description Gait
Data Type Ordinal

Values		Comments
Original	**Replacement**	
0-4	-	-
8	CM	Conditionally missing as value omitted for a reason. (1.12%)
-4	CM	Conditionally missing as form not required. (2.58%)

Dependency Trigger -
Relationship -

Inspect Parent	Calculate/Derive	Impute
✗	✗	✗

General Comments -

POSSTAB

Form Versions	Derived (NACC)	Derived (New)
1.2, 2.0	✗	✗

Description Posture stability
Data Type Ordinal

Values		Comments
Original	**Replacement**	
0-4	-	-
8	CM	Conditionally missing as value omitted for a reason. (2.68%)
-4	CM	Conditionally missing as form not required. (2.58%)

Dependency Trigger -
Relationship -

Inspect Parent	Calculate/Derive	Impute
✗	✗	✗

General Comments -

BRADYKIN

Form Versions	Derived (NACC)	Derived (New)
1.2, 2.0	✗	✗

Description Body bradykinesia and hypokinesia
Data Type Ordinal

Values		Comments
Original	**Replacement**	
0-4	-	-
8	CM	Conditionally missing as value omitted for a reason. (0.34%)
-4	CM	Conditionally missing as form not required. (2.58%)

Dependency Trigger -
Relationship -

Inspect Parent	Calculate/Derive	Impute
✗	✗	✗

General Comments -

PDNORMAL

Form Versions	Derived (NACC)	Derived (New)
1.2, 2.0	✗	✗

Description Unified Parkinson's Disease Rating Scale (UPDRS) normal
Data Type Binary

Values		Comments
Original	**Replacement**	
0-1	-	-
8	CM	Conditionally missing as cannot be sensibly imputed. (0.30%)
-4	CM	Conditionally missing as form not required. (2.58%)

Dependency Trigger -
Relationship PDNORMAL = 1 if SPEECH, FACEXP, TRESTFAC, TRESTRHD, TRESTLHD, TRESTRFT, TRESTLFT, TRACTRHD, TRACTLHD, RIGDNECK, RIGDUPRT, RIGDUPLF, RIGDLORT, RIGDLOLF, TAPSRT, TAPSLF, HANDMOVR, HANDMOVL, HANDALTR, HANDALTL, LEGRT, LEGLF, ARISING, POSTURE, GAIT, POSSTAB and BRADYKIN = 0 else 0 (PDNORMAL = CM if all CM or all CM and 0)

Inspect Parent	Calculate/Derive	Impute
✗	✗	✗

General Comments NACC's coding guidebooks and data element dictionaries for versions 1.2 and 2.0 state all B3 variables are dependent on PDNORMAL but dependency does not hold. No need to calculate/derive variable as no missing values in any of the variables involved.

Dropped Variables

Name SPEECHX
Description Speech; untestable - specify reason
Reason Free-text variable and only available in version 2.0 of the UDS.

Name FACEXPX
Description Facial expression; untestable - specify reason
Reason Free-text variable and only available in version 2.0 of the UDS.

Name TRESTFAX
Description Tremor at rest - face, lips, chin; untestable - specify reason
Reason Free-text variable and only available in version 2.0 of the UDS.

Name TRESTRHX
Description Tremor at rest - right hand; untestable - specify reason
Reason Free-text variable and only available in version 2.0 of the UDS.

Name TRESTLHX
Description Tremor at rest - left hand; untestable - specify reason
Reason Free-text variable and only available in version 2.0 of the UDS.

Name TRESTRFX
Description Tremor at rest - right foot; untestable - specify reason
Reason Free-text variable and only available in version 2.0 of the UDS.

Name TRESTLFX
Description Tremor at rest - left foot; untestable - specify reason
Reason Free-text variable and only available in version 2.0 of the UDS.

Name TRACTRHX
Description Action or postural tremor - right hand; untestable - specify reason
Reason Free-text variable and only available in version 2.0 of the UDS.

Name TRACTLHX
Description Action or postural tremor - left hand; untestable - specify reason
Reason Free-text variable and only available in version 2.0 of the UDS.

Name RIGDNEX
Description Rigidity - neck; untestable - specify reason
Reason Free-text variable and only available in version 2.0 of the UDS.

Name RIGDUPRX
Description Rigidity - right upper extremity; untestable - specify reason
Reason Free-text variable and only available in version 2.0 of the UDS.

Name RIGDUPLX
Description Rigidity - left upper extremity; untestable - specify reason
Reason Free-text variable and only available in version 2.0 of the UDS.

Name RIGDLORX
Description Rigidity - right lower extremity; untestable - specify reason
Reason Free-text variable and only available in version 2.0 of the UDS.

Name RIGDLOLX
Description Rigidity - left lower extremity; untestable - specify reason
Reason Free-text variable and only available in version 2.0 of the UDS.

Name TAPSRTX
Description Finger taps - right hand; untestable - specify reason
Reason Free-text variable.

Name TAPSLFX
Description Finger taps - left hand; untestable - specify reason
Reason Free-text variable.

Name HANDMVRX
Description Hand movements - right hand; untestable - specify reason
Reason Free-text variable.

Name HANDMVLX
Description Hand movements - left hand; untestable - specify reason
Reason Free-text variable.

Name HANDATRX
Description Alternating movement - right hand; untestable - specify reason
Reason Free-text variable.

Name HANDATLX
Description Alternating movement - left hand; untestable - specify reason
Reason Free-text variable.

Name LEGRTX
Description Leg agility - right leg; untestable - specify reason
Reason Free-text variable.

Name LEGLFX
Description Leg agility - left leg; untestable - specify reason
Reason Free-text variable.

Name ARISINGX
Description Arising from chair; untestable - specify reason
Reason Free-text variable.

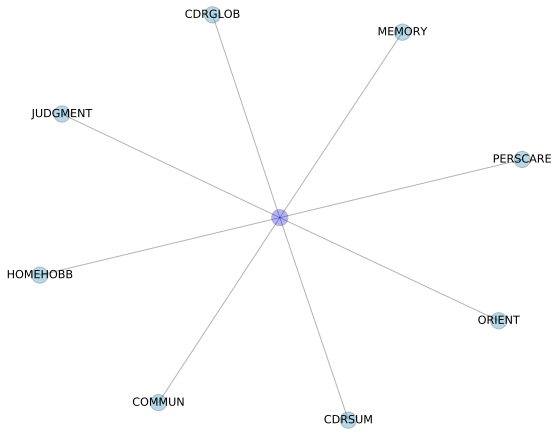
Name POSTUREX
Description Posture; untestable - specify reason
Reason Free-text variable.

Name GAITX
Description Gait; untestable - specify reason
Reason Free-text variable.

Name POSSTABX
Description Posture stability; untestable - specify reason
Reason Free-text variable.

Name BRADYKIX
Description Body bradykinesia and hypokinesia; untestable - specify reason
Reason Free-text variable and only available in version 2.0 of the UDS.

B4 - Clinical Dementia Rating



Number of Variables	10
Number of Variables Used	8
Form Required?	✓
Form Missingness	0.00%

MEMORY

Form Versions	Derived (NACC)	Derived (New)
1.2, 2.0, 3.0	✗	✗

Description Memory
Data Type Ordinal

Values		Comments
Original	**Replacement**	
0.0-1.0 (step=0.5)	-	-
2.0-3.0 (step=1.0)	-	-

Dependency Trigger -
Relationship -

Inspect Parent	Calculate/Derive	Impute
✗	✗	✗

General Comments -

ORIENT

Form Versions	Derived (NACC)	Derived (New)
1.2, 2.0, 3.0	✗	✗

Description Orientation
Data Type Ordinal

Values		Comments
Original	**Replacement**	
0.0-1.0 (step=0.5)	-	-
2.0-3.0 (step=1.0)	-	-

Dependency Trigger -
Relationship -

Inspect Parent	Calculate/Derive	Impute
✗	✗	✗

General Comments -

JUDGMENT

Form Versions	Derived (NACC)	Derived (New)
1.2, 2.0, 3.0	✗	✗

Description Judgment and problem-solving
Data Type Ordinal

Values		Comments
Original	Replacement	
0.0-1.0 (step=0.5)	-	-
2.0-3.0 (step=1.0)	-	-

Dependency Trigger -
Relationship -

Inspect Parent	Calculate/Derive	Impute
✗	✗	✗

General Comments -

COMMUN

Form Versions	Derived (NACC)	Derived (New)
1.2, 2.0, 3.0	✗	✗

Description Community affairs
Data Type Ordinal

Values		Comments
Original	Replacement	
0.0-1.0 (step=0.5)	-	-
2.0-3.0 (step=1.0)	-	-

Dependency Trigger -
Relationship -

Inspect Parent	Calculate/Derive	Impute
✗	✗	✗

General Comments -

HOMEHOBB

Form Versions	Derived (NACC)	Derived (New)
1.2, 2.0, 3.0	✗	✗

Description Home and hobbies
Data Type Ordinal

Values		Comments
Original	Replacement	
0.0-1.0 (step=0.5)	-	-
2.0-3.0 (step=1.0)	-	-

Dependency Trigger -
Relationship -

Inspect Parent	Calculate/Derive	Impute
✗	✗	✗

General Comments -

PERSCARE

Form Versions	Derived (NACC)	Derived (New)
1.2, 2.0, 3.0	✗	✗

Description Personal care
Data Type Ordinal

Values		Comments
Original	Replacement	
0.0-3.0 (step=1.0)	-	-

Dependency Trigger -
Relationship -

Inspect Parent	Calculate/Derive	Impute
✗	✗	✗

General Comments -

CDRSUM

Form Versions	Derived (NACC)	Derived (New)
1.2, 2.0, 3.0	✗	✗

Description Clinical Dementia Rating (CDR) sum of boxes
Data Type Continuous

Values		Comments
Original	Replacement	
0.0-16.0 (step=0.5)	-	-
17.0-18.0 (step=1.0)	-	-

Dependency Trigger -
Relationship CDRSUM = sum of MEMORY, ORIENT, JUDGMENT, COMMUN, HOMEHOBB and PERSCARE

Inspect Parent	Calculate/Derive	Impute
✗	✗	✗

General Comments No need to calculate/derive variable as no missing values in any of the variables involved.

CDRGLOB

Form Versions	Derived (NACC)	Derived (New)
1.2, 2.0, 3.0	✗	✗

Description Global CDR
Data Type Ordinal

Values		Comments
Original	**Replacement**	
0.0-1.0 (step=0.5)	-	-
2.0-3.0 (step=1.0)	-	-

Dependency Trigger -
Relationship CDRGLOB is derived using the Washington University CDR-assignment algorithm.

Inspect Parent	Calculate/Derive	Impute
✗	✗	✗

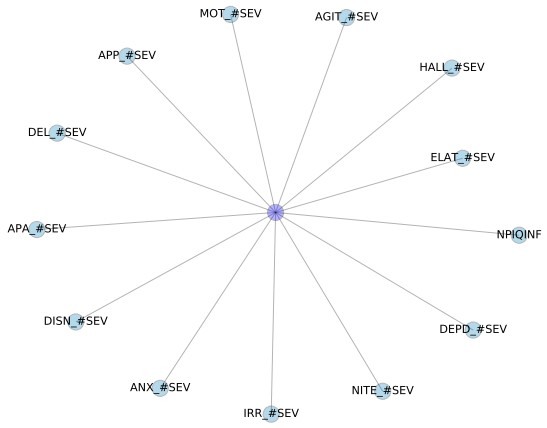
General Comments Relationship not verified as values automatically generated, and no missingness. No need to calculate/derive variable as no missing values in any of the variables involved.

Dropped Variables

Name COMPORT
Description Behaviour, comportment and personality
Reason Not available in version 1.2 of the UDS.

Name CDRLANG
Description Language
Reason Not available in version 1.2 of the UDS.

B5 - Neuropsychiatric Inventory Questionnaire



Number of Variables	26
Number of Variables Used	13
Form Required?	✗
Form Missingness	5.42%

NPIQINF

Form Versions	Derived (NACC)	Derived (New)
1.2, 2.0, 3.0	✗	✗

Description Neuropsychiatric Inventory Questionnaire (NPI-Q) co-participant
Data Type Categorical

Values		Comments
Original	**Replacement**	
1-3	-	-
-4	CM	Conditionally missing as form not required. (5.42%)

Dependency Trigger -
Relationship -

Inspect Parent	Calculate/Derive	Impute
✗	✗	✗

General Comments -

DEL_SEV

Form Versions	Derived (NACC)	Derived (New)
1.2, 2.0, 3.0	✗	✓

Description Delusions and their severity in the last month
Data Type Ordinal

Values		Comments
Original	**Replacement**	
0-3	-	Indicate none, mild, moderate and severe respectively.
M	-	(0.00% (DEL 9) + 0.00% (DELSEV 9))
CM	-	Conditionally missing as form not required. (5.42% (-4))

Dependency Trigger -
Relationship -

Inspect Parent	Calculate/Derive	Impute
✗	✗	✓

General Comments Derived from DEL and DELSEV. A conditionally missing value should not be used as a potential fill value if imputed.

HALL_SEV

Form Versions	Derived (NACC)	Derived (New)
1.2, 2.0, 3.0	✗	✓

Description Hallucinations and their severity in the last month
Data Type Ordinal

Values		Comments
Original	**Replacement**	
0-3	-	Indicate none, mild, moderate and severe respectively.
M	-	(0.00% (HALL 9) + 0.00% (HALLSEV 9))
CM	-	Conditionally missing as form not required. (5.42% (-4))

Dependency Trigger -
Relationship -

Inspect Parent	Calculate/Derive	Impute
✗	✗	✓

General Comments Derived from HALL and HALLSEV. A conditionally missing value should not be used as a potential fill value if imputed.

AGIT_SEV

Form Versions	Derived (NACC)	Derived (New)
1.2, 2.0, 3.0	✗	✓

Description Agitation or aggression, and the severity, in the last month
Data Type Ordinal

Values		Comments
Original	**Replacement**	
0-3	-	Indicate none, mild, moderate and severe respectively.
M	-	(0.00% (AGIT 9) + 0.00% (AGITSEV 9))
CM	-	Conditionally missing as form not required. (5.42% (-4))

Dependency Trigger -
Relationship -

Inspect Parent	Calculate/Derive	Impute
✗	✗	✓

General Comments Derived from AGIT and AGITSEV. A conditionally missing value should not be used as a potential fill value if imputed.

DEPD_SEV

Form Versions	Derived (NACC)	Derived (New)
1.2, 2.0, 3.0	✗	✓

Description Depression or dysphoria, and the severity, in the last month
Data Type Ordinal

Values		Comments
Original	**Replacement**	
0-3	-	Indicate none, mild, moderate and severe respectively.
M	-	(0.00% (DEPD 9) + 0.00% (DEPDSEV 9))
CM	-	Conditionally missing as form not required. (5.42% (-4))

Dependency Trigger -
Relationship -

Inspect Parent	Calculate/Derive	Impute
✗	✗	✓

General Comments Derived from DEPD and DEPDSEV. A conditionally missing value should not be used as a potential fill value if imputed.

ANX_SEV

Form Versions	Derived (NACC)	Derived (New)
1.2, 2.0, 3.0	✗	✓

Description Anxiety and the severity in the last month
Data Type Ordinal

Values		Comments
Original	**Replacement**	
0-3	-	Indicate none, mild, moderate and severe respectively.
M	-	(0.00% (ANX 9) + 0.00% (ANXSEV 9))
CM	-	One ANXSEV value converted from -4 to 8 to indicate dependence on ANX. Conditionally missing as form not required. (5.42% (-4))

Dependency Trigger -
Relationship -

Inspect Parent	Calculate/Derive	Impute
✗	✗	✓

General Comments Derived from ANX and ANXSEV. A conditionally missing value should not be used as a potential fill value if imputed.

ELAT_SEV

Form Versions	Derived (NACC)	Derived (New)
1.2, 2.0, 3.0	✗	✓

Description Elation or euphoria, and the severity, in the last month
Data Type Ordinal

Values		Comments
Original	**Replacement**	
0-3	-	Indicate none, mild, moderate and severe respectively.
M	-	(0.00% (ELAT 9) + 0.00% (ELATSEV 9))
CM	-	Conditionally missing as form not required. (5.42% (-4))

Dependency Trigger -
Relationship -

Inspect Parent	Calculate/Derive	Impute
✗	✗	✓

General Comments Derived from ELAT and ELATSEV. A conditionally missing value should not be used as a potential fill value if imputed.

APA_SEV

Form Versions	Derived (NACC)	Derived (New)
1.2, 2.0, 3.0	✗	✓

Description Apathy or indifference, and the severity, in the last month
Data Type Ordinal

Values		Comments
Original	**Replacement**	
0-3	-	Indicate none, mild, moderate and severe respectively.
M	-	(0.00% (APA 9) + 0.00% (APASEV 9))
CM	-	Conditionally missing as form not required. (5.42% (-4))

Dependency Trigger -
Relationship -

Inspect Parent	Calculate/Derive	Impute
✗	✗	✓

General Comments Derived from APA and APASEV. A conditionally missing value should not be used as a potential fill value if imputed.

DISN_SEV

Form Versions	Derived (NACC)	Derived (New)
1.2, 2.0, 3.0	✗	✓

Description Disinhibition and the severity in the last month
Data Type Ordinal

Values		Comments
Original	**Replacement**	
0-3	-	Indicate none, mild, moderate and severe respectively.
M	-	(0.00% (DISN 9) + 0.00% (DISNSEV 9))
CM	-	Conditionally missing as form not required. (5.42% (-4))

Dependency Trigger -
Relationship -

Inspect Parent	Calculate/Derive	Impute
✗	✗	✓

General Comments Derived from DISN and DISNSEV. A conditionally missing value should not be used as a potential fill value if imputed.

IRR_SEV

Form Versions	Derived (NACC)	Derived (New)
1.2, 2.0, 3.0	✗	✓

Description Irritability or lability, and the severity, in the last month
Data Type Ordinal

Values		Comments
Original	**Replacement**	
0-3	-	Indicate none, mild, moderate and severe respectively.
M	-	(0.00% (IRR 9) + 0.00% (IRRSEV 9))
CM	-	Conditionally missing as form not required. (5.42% (-4))

Dependency Trigger -
Relationship -

Inspect Parent	Calculate/Derive	Impute
✗	✗	✓

General Comments Derived from IRR and IRRSEV. A conditionally missing value should not be used as a potential fill value if imputed.

MOT_SEV

Form Versions	Derived (NACC)	Derived (New)
1.2, 2.0, 3.0	✗	✓

Description Motor disturbance and the severity in the last month
Data Type Ordinal

Values		Comments
Original	**Replacement**	
0-3	-	Indicate none, mild, moderate and severe respectively.
M	-	(0.00% (MOT 9) + 0.00% (MOTSEV 9))
CM	-	Conditionally missing as form not required. (5.42% (-4))

Dependency Trigger -
Relationship -

Inspect Parent	Calculate/Derive	Impute
✗	✗	✓

General Comments Derived from MOT and MOTSEV. A conditionally missing value should not be used as a potential fill value if imputed.

NITE_SEV

Form Versions	Derived (NACC)	Derived (New)
1.2, 2.0, 3.0	✗	✓

Description Nighttime behaviours and their severity in the last month
Data Type Ordinal

Values		Comments
Original	**Replacement**	
0-3	-	Indicate none, mild, moderate and severe respectively.
M	-	(0.00% (NITE 9) + 0.00% (NITESEV 9))
CM	-	Conditionally missing as form not required. (5.42% (-4))

Dependency Trigger -
Relationship -

Inspect Parent	Calculate/Derive	Impute
✗	✗	✓

General Comments Derived from NITE and NITESEV. A conditionally missing value should not be used as a potential fill value if imputed.

APP_SEV

Form Versions	Derived (NACC)	Derived (New)
1.2, 2.0, 3.0	✗	✓

Description Appetite and eating problems, and their severity, in the last month
Data Type Ordinal

Values		Comments
Original	**Replacement**	
0-3	-	Indicate none, mild, moderate and severe respectively.
M	-	One APP value converted from 1 to M as APPSEV value M. (0.00% (APP 9) + 0.00% (APPSEV 9))
CM	-	Conditionally missing as form not required. (5.42% (-4))

Dependency Trigger -
Relationship -

Inspect Parent	Calculate/Derive	Impute
✗	✗	✓

General Comments Derived from APP and APPSEV. A conditionally missing value should not be used as a potential fill value if imputed.

Dropped Variables

Name NPIQINFX
Description NPI-Q co-participant, other - specify
Reason Free-text variable.

Name DEL
Description Delusions in the last month
Reason Replaced by the derived variable DEL_SEV.

Name DELSEV
Description Delusions severity
Reason Replaced by the derived variable DEL_SEV.

Name HALL
Description Hallucinations in the last month
Reason Replaced by the derived variable HALL_SEV.

Name HALLSEV
Description Hallucinations severity
Reason Replaced by the derived variable HALL_SEV.

Name AGIT
Description Agitation or aggression in the last month
Reason Replaced by the derived variable AGIT_SEV.

Name AGITSEV
Description Agitation or aggression severity
Reason Replaced by the derived variable AGIT_SEV.

Name DEPD
Description Depression or dysphoria in the last month
Reason Replaced by the derived variable DEPD_SEV.

Name DEPDSEV
Description Depression or dysphoria severity
Reason Replaced by the derived variable DEPD_SEV.

Name ANX
Description Anxiety in the last month
Reason Replaced by the derived variable ANX_SEV.

Name ANXSEV
Description Anxiety severity
Reason Replaced by the derived variable ANX_SEV.

Name ELAT
Description Elation or euphoria in the last month
Reason Replaced by the derived variable ELAT_SEV.

Name ELATSEV
Description Elation or euphoria severity
Reason Replaced by the derived variable ELAT_SEV.

Name APA
Description Apathy or indifference in the last month
Reason Replaced by the derived variable APA_SEV.

Name APASEV
Description Apathy or indifference severity
Reason Replaced by the derived variable APA_SEV.

Name DISN
Description Disinhibition in the last month
Reason Replaced by the derived variable DISN_SEV.

Name DISNSEV
Description Disinhibition severity
Reason Replaced by the derived variable DISN_SEV.

Name IRR
Description Irritability or lability in the last month
Reason Replaced by the derived variable IRR_SEV.

Name IRRSEV
Description Irritability or lability severity
Reason Replaced by the derived variable IRR_SEV.

Name MOT
Description Motor disturbance in the last month
Reason Replaced by the derived variable MOT_SEV.

Name MOTSEV
Description Motor disturbance severity
Reason Replaced by the derived variable MOT_SEV.

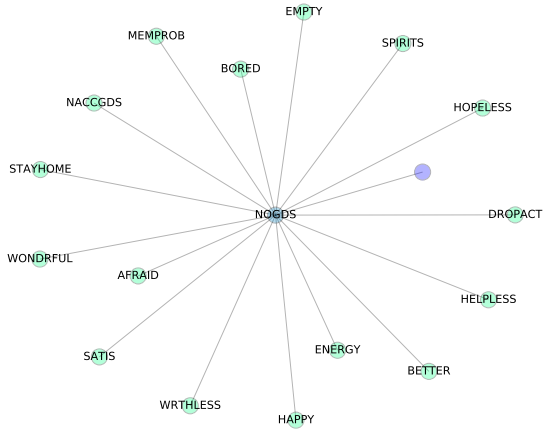
Name NITE
Description Nighttime behaviours in the last month
Reason Replaced by the derived variable NITE_SEV.

Name NITESEV
Description Nighttime behaviours severity
Reason Replaced by the derived variable NITE_SEV.

Name APP
Description Appetite and eating problems in the last month
Reason Replaced by the derived variable APP_SEV.

Name APPSEV
Description Appetite and eating problems severity
Reason Replaced by the derived variable APP_SEV.

B6 - Geriatric Depression Scale



Number of Variables	17
Number of Variables Used	17
Form Required?	✗
Form Missingness	2.98%

NOGDS

Form Versions	Derived (NACC)	Derived (New)
1.2, 2.0, 3.0	✗	✗

Description Is the subject able to complete the Geriatric Depression Scale (GDS), based on the clinician's best judgment?
Data Type Binary

Values		Comments
Original	Replacement	
0-1	-	-
-4	CM	Conditionally missing as form not required. (2.98%)

Dependency Trigger 1
Relationship -

Inspect Parent	Calculate/Derive	Impute
✗	✗	✗

General Comments -

SATIS

Form Versions	Derived (NACC)	Derived (New)
1.2, 2.0, 3.0	✗	✗

Description Are you basically satisfied with your life?
Data Type Binary

Values		Comments
Original	Replacement	
0-1	-	-
CM	-	Replaced 9 values to indicate dependence on NOGDS. (5.24%)
9	M	(0.02%)
-4	CM	Conditionally missing as form not required. (2.98%)

Dependency Trigger -
Relationship -

Inspect Parent	Calculate/Derive	Impute
✗	✗	✓

General Comments No need to inspect parent as no missing values in parent. A conditionally missing value should not be used as a potential fill value if imputed.

DROPACT

Form Versions	Derived (NACC)	Derived (New)
1.2, 2.0, 3.0	✗	✗

Description Have you dropped many of your activities and interests?
Data Type Binary

Values		Comments
Original	**Replacement**	
0-1	-	-
CM	-	Replaced 9 values to indicate dependence on NOGDS. (5.24%)
9	M	(0.02%)
-4	CM	Conditionally missing as form not required. (2.98%)

Dependency Trigger -
Relationship -

Inspect Parent	Calculate/Derive	Impute
✗	✗	✓

General Comments No need to inspect parent as no missing values in parent. A conditionally missing value should not be used as a potential fill value if imputed.

EMPTY

Form Versions	Derived (NACC)	Derived (New)
1.2, 2.0, 3.0	✗	✗

Description Do you feel that your life is empty?
Data Type Binary

Values		Comments
Original	**Replacement**	
0-1	-	-
CM	-	Replaced 9 values to indicate dependence on NOGDS. (5.24%)
9	M	(0.02%)
-4	CM	Conditionally missing as form not required. (2.98%)

Dependency Trigger -
Relationship -

Inspect Parent	Calculate/Derive	Impute
✗	✗	✓

General Comments No need to inspect parent as no missing values in parent. A conditionally missing value should not be used as a potential fill value if imputed.

BORED

Form Versions	Derived (NACC)	Derived (New)
1.2, 2.0, 3.0	✗	✗

Description Do you often get bored?
Data Type Binary

Values		Comments
Original	Replacement	
0-1	-	-
CM	-	Replaced 9 values to indicate dependence on NOGDS. (5.24%)
9	M	(0.02%)
-4	CM	Conditionally missing as form not required. (2.98%)

Dependency Trigger -
Relationship -

Inspect Parent	Calculate/Derive	Impute
✗	✗	✓

General Comments No need to inspect parent as no missing values in parent. A conditionally missing value should not be used as a potential fill value if imputed.

SPIRITS

Form Versions	Derived (NACC)	Derived (New)
1.2, 2.0, 3.0	✗	✗

Description Are you in good spirits most of the time?
Data Type Binary

Values		Comments
Original	Replacement	
0-1	-	-
CM	-	Replaced 9 values to indicate dependence on NOGDS. (5.24%)
9	M	(0.03%)
-4	CM	Conditionally missing as form not required. (2.98%)

Dependency Trigger -
Relationship -

Inspect Parent	Calculate/Derive	Impute
✗	✗	✓

General Comments No need to inspect parent as no missing values in parent. A conditionally missing value should not be used as a potential fill value if imputed.

AFRAID

Form Versions	Derived (NACC)	Derived (New)
1.2, 2.0, 3.0	✗	✗

Description Are you afraid that something bad is going to happen to you?
Data Type Binary

Values		Comments
Original	Replacement	
0-1	-	-
CM	-	Replaced 9 values to indicate dependence on NOGDS. (5.24%)
9	M	(0.04%)
-4	CM	Conditionally missing as form not required. (2.98%)

Dependency Trigger -
Relationship -

Inspect Parent	Calculate/Derive	Impute
✗	✗	✓

General Comments No need to inspect parent as no missing values in parent. A conditionally missing value should not be used as a potential fill value if imputed.

HAPPY

Form Versions	Derived (NACC)	Derived (New)
1.2, 2.0, 3.0	✗	✗

Description Do you feel happy most of the time?
Data Type Binary

Values		Comments
Original	Replacement	
0-1	-	-
CM	-	Replaced 9 values to indicate dependence on NOGDS. (5.24%)
9	M	(0.05%)
-4	CM	Conditionally missing as form not required. (2.98%)

Dependency Trigger -
Relationship -

Inspect Parent	Calculate/Derive	Impute
✗	✗	✓

General Comments No need to inspect parent as no missing values in parent. A conditionally missing value should not be used as a potential fill value if imputed.

HELPLESS

Form Versions	Derived (NACC)	Derived (New)
1.2, 2.0, 3.0	✗	✗

Description Do you often feel helpless?
Data Type Binary

Values		Comments
Original	Replacement	
0-1	-	-
CM	-	Replaced 9 values to indicate dependence on NOGDS. (5.24%)
9	M	(0.03%)
-4	CM	Conditionally missing as form not required. (2.98%)

Dependency Trigger -
Relationship -

Inspect Parent	Calculate/Derive	Impute
✗	✗	✓

General Comments No need to inspect parent as no missing values in parent. A conditionally missing value should not be used as a potential fill value if imputed.

STAYHOME

Form Versions	Derived (NACC)	Derived (New)
1.2, 2.0, 3.0	✗	✗

Description Do you prefer to stay at home, rather than going out and doing new things?
Data Type Binary

Values		Comments
Original	Replacement	
0-1	-	-
CM	-	Replaced 9 values to indicate dependence on NOGDS. (5.24%)
9	M	(0.05%)
-4	CM	Conditionally missing as form not required. (2.98%)

Dependency Trigger -
Relationship -

Inspect Parent	Calculate/Derive	Impute
✗	✗	✓

General Comments No need to inspect parent as no missing values in parent. A conditionally missing value should not be used as a potential fill value if imputed.

MEMPROB

Form Versions	Derived (NACC)	Derived (New)
1.2, 2.0, 3.0	✗	✗

Description Do you feel you have more problems with memory than most?
Data Type Binary

Values		Comments
Original	Replacement	
0-1	-	-
CM	-	Replaced 9 values to indicate dependence on NOGDS. (5.24%)
9	M	(0.09%)
-4	CM	Conditionally missing as form not required. (2.98%)

Dependency Trigger -
Relationship -

Inspect Parent	Calculate/Derive	Impute
✗	✗	✓

General Comments No need to inspect parent as no missing values in parent. A conditionally missing value should not be used as a potential fill value if imputed.

WONDRFUL

Form Versions	Derived (NACC)	Derived (New)
1.2, 2.0, 3.0	✗	✗

Description Do you think it is wonderful to be alive now?
Data Type Binary

Values		Comments
Original	Replacement	
0-1	-	-
CM	-	Replaced 9 values to indicate dependence on NOGDS. (5.24%)
9	M	(0.07%)
-4	CM	Conditionally missing as form not required. (2.98%)

Dependency Trigger -
Relationship -

Inspect Parent	Calculate/Derive	Impute
✗	✗	✓

General Comments No need to inspect parent as no missing values in parent. A conditionally missing value should not be used as a potential fill value if imputed.

WRTHLESS

Form Versions	Derived (NACC)	Derived (New)
1.2, 2.0, 3.0	✗	✗

Description Do you feel pretty worthless the way you are now?
Data Type Binary

Values		Comments
Original	Replacement	
0-1	-	-
CM	-	Replaced 9 values to indicate dependence on NOGDS. (5.24%)
9	M	(0.06%)
-4	CM	Conditionally missing as form not required. (2.98%)

Dependency Trigger -
Relationship -

Inspect Parent	Calculate/Derive	Impute
✗	✗	✓

General Comments No need to inspect parent as no missing values in parent. A conditionally missing value should not be used as a potential fill value if imputed.

ENERGY

Form Versions	Derived (NACC)	Derived (New)
1.2, 2.0, 3.0	✗	✗

Description Do you feel full of energy?
Data Type Binary

Values		Comments
Original	Replacement	
0-1	-	-
CM	-	Replaced 9 values to indicate dependence on NOGDS. (5.24%)
9	M	(0.06%)
-4	CM	Conditionally missing as form not required. (2.98%)

Dependency Trigger -
Relationship -

Inspect Parent	Calculate/Derive	Impute
✗	✗	✓

General Comments No need to inspect parent as no missing values in parent. A conditionally missing value should not be used as a potential fill value if imputed.

HOPELESS

Form Versions	Derived (NACC)	Derived (New)
1.2, 2.0, 3.0	✗	✗

Description Do you feel that your situation is hopeless?
Data Type Binary

Values		Comments
Original	Replacement	
0-1	-	-
CM	-	Replaced 9 values to indicate dependence on NOGDS. (5.24%)
9	M	(0.03%)
-4	CM	Conditionally missing as form not required. (2.98%)

Dependency Trigger -
Relationship -

Inspect Parent	Calculate/Derive	Impute
✗	✗	✓

General Comments No need to inspect parent as no missing values in parent. A conditionally missing value should not be used as a potential fill value if imputed.

BETTER

Form Versions	Derived (NACC)	Derived (New)
1.2, 2.0, 3.0	✗	✗

Description Do you think that most people are better off than you are?
Data Type Binary

Values		Comments
Original	Replacement	
0-1	-	-
CM	-	Replaced 9 values to indicate dependence on NOGDS. (5.24%)
9	M	(0.07%)
-4	CM	Conditionally missing as form not required. (2.98%)

Dependency Trigger -
Relationship -

Inspect Parent	Calculate/Derive	Impute
✗	✗	✓

General Comments No need to inspect parent as no missing values in parent. A conditionally missing value should not be used as a potential fill value if imputed.

NACCGDS

Form Versions	Derived (NACC)	Derived (New)
1.2, 2.0, 3.0	✗	✗

Description Total GDS score
Data Type Continuous

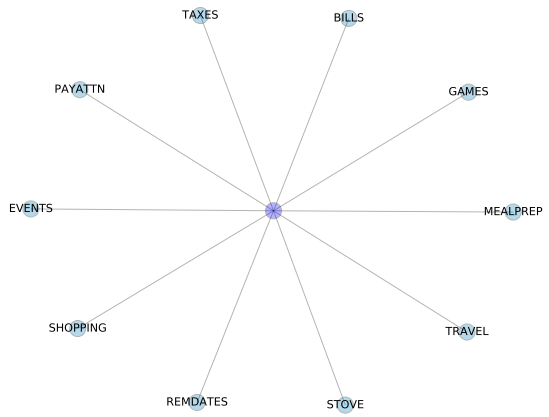
Values		Comments
Original	**Replacement**	
0-15	-	-
88	CM	Indicates subject not able to complete the GDS or more than three GDS items missing. (5.24%)
-4	CM	Conditionally missing as form not required. (2.98%)

Dependency Trigger -
Relationship NACCGDS = sum of SATIS, DROPACT, EMPTY, BORED, SPIRITS, AFRAID, HAPPY, HELPLESS, STAYHOME, MEMPROB, WONDRFUL, WRTHLESS, ENERGY, HOPELESS and BETTER (GDS items) where <= 3 missing (sum of scores + sum of scores / number of scores × number missing, rounded) (NACCGDS = CM if > 3 missing)

Inspect Parent	Calculate/Derive	Impute
✗	✗	✗

General Comments No need to inspect parent, as no missing values in parent; or calculate/derive variable, as no missing values in variable. Values not updated when GDS items imputed.

B7 - Functional Activities Questionnaire



Number of Variables	10
Number of Variables Used	10
Form Required?	✗
Form Missingness	3.22%

BILLS

Form Versions	Derived (NACC)	Derived (New)
1.2, 2.0, 3.0	✗	✗

Description In the past four weeks, did the subject have difficulty or need help with: writing checks, paying bills or balancing a checkbook

Data Type Ordinal

Values		Comments
Original	Replacement	
0-3	-	-
8	CM	(7.95%)
9	M	(0.00%)
-4	CM	Conditionally missing as form not required. (3.22%)

Dependency Trigger -

Relationship -

Inspect Parent	Calculate/Derive	Impute
✗	✗	✓

General Comments A conditionally missing value should be used as a potential fill value if imputed.

TAXES

Form Versions	Derived (NACC)	Derived (New)
1.2, 2.0, 3.0	✗	✗

Description In the past four weeks, did the subject have difficulty or need help with: assembling tax records, business affairs or other papers

Data Type Ordinal

Values		Comments
Original	Replacement	
0-3	-	-
8	CM	(11.74%)
9	M	(0.00%)
-4	CM	Conditionally missing as form not required. (3.22%)

Dependency Trigger -

Relationship -

Inspect Parent	Calculate/Derive	Impute
✗	✗	✓

General Comments A conditionally missing value should be used as a potential fill value if imputed.

SHOPPING

Form Versions	Derived (NACC)	Derived (New)
1.2, 2.0, 3.0	✗	✗

Description In the past four weeks, did the subject have difficulty or need help with: shopping alone for clothes, household necessities or groceries

Data Type Ordinal

Values		Comments
Original	**Replacement**	
0-3	-	-
8	CM	(2.35%)
9	M	(0.00%)
-4	CM	Conditionally missing as form not required. (3.22%)

Dependency Trigger -

Relationship -

Inspect Parent	Calculate/Derive	Impute
✗	✗	✓

General Comments A conditionally missing value should be used as a potential fill value if imputed.

GAMES

Form Versions	Derived (NACC)	Derived (New)
1.2, 2.0, 3.0	✗	✗

Description In the past four weeks, did the subject have difficulty or need help with: playing a game of skill such as bridge or chess, or working on a hobby

Data Type Ordinal

Values		Comments
Original	**Replacement**	
0-3	-	-
8	CM	(9.55%)
9	M	(0.00%)
-4	CM	Conditionally missing as form not required. (3.22%)

Dependency Trigger -

Relationship -

Inspect Parent	Calculate/Derive	Impute
✗	✗	✓

General Comments A conditionally missing value should be used as a potential fill value if imputed.

STOVE

Form Versions	Derived (NACC)	Derived (New)
1.2, 2.0, 3.0	✗	✗

Description In the past four weeks, did the subject have difficulty or need help with: heating water, making a cup of coffee, or turning off the stove

Data Type Ordinal

Values		Comments
Original	Replacement	
0-3	-	-
8	CM	(1.81%)
9	M	(0.00%)
-4	CM	Conditionally missing as form not required. (3.22%)

Dependency Trigger -

Relationship -

Inspect Parent	Calculate/Derive	Impute
✗	✗	✓

General Comments A conditionally missing value should be used as a potential fill value if imputed.

MEALPREP

Form Versions	Derived (NACC)	Derived (New)
1.2, 2.0, 3.0	✗	✗

Description In the past four weeks, did the subject have difficulty or need help with: preparing a balanced meal

Data Type Ordinal

Values		Comments
Original	Replacement	
0-3	-	-
8	CM	(9.76%)
9	M	(0.00%)
-4	CM	Conditionally missing as form not required. (3.22%)

Dependency Trigger -

Relationship -

Inspect Parent	Calculate/Derive	Impute
✗	✗	✓

General Comments A conditionally missing value should be used as a potential fill value if imputed.

EVENTS

Form Versions	Derived (NACC)	Derived (New)
1.2, 2.0, 3.0	✗	✗

Description In the past four weeks, did the subject have difficulty or need help with: keeping track of current events

Data Type Ordinal

Values		Comments
Original	**Replacement**	
0-3	-	-
8	CM	(1.10%)
9	M	(0.00%)
-4	CM	Conditionally missing as form not required. (3.22%)

Dependency Trigger -

Relationship -

Inspect Parent	Calculate/Derive	Impute
✗	✗	✓

General Comments A conditionally missing value should be used as a potential fill value if imputed.

PAYATTN

Form Versions	Derived (NACC)	Derived (New)
1.2, 2.0, 3.0	✗	✗

Description In the past four weeks, did the subject have difficulty or need help with: paying attention to and understanding a TV programme, book or magazine

Data Type Ordinal

Values		Comments
Original	**Replacement**	
0-3	-	-
8	CM	(0.49%)
9	M	(0.00%)
-4	CM	Conditionally missing as form not required. (3.22%)

Dependency Trigger -

Relationship -

Inspect Parent	Calculate/Derive	Impute
✗	✗	✓

General Comments A conditionally missing value should be used as a potential fill value if imputed.

REMDATES

Form Versions	Derived (NACC)	Derived (New)
1.2, 2.0, 3.0	✗	✗

Description In the past four weeks, did the subject have difficulty or need help with: remembering appointments, family occasions, holidays or medications

Data Type Ordinal

Values		Comments
Original	**Replacement**	
0-3	-	-
8	CM	(0.50%)
9	M	(0.00%)
-4	CM	Conditionally missing as form not required. (3.22%)

Dependency Trigger -

Relationship -

Inspect Parent	Calculate/Derive	Impute
✗	✗	✓

General Comments A conditionally missing value should be used as a potential fill value if imputed.

TRAVEL

Form Versions	Derived (NACC)	Derived (New)
1.2, 2.0, 3.0	✗	✗

Description In the past four weeks, did the subject have difficulty or need help with: travelling out of the neighbourhood, driving, or arranging to take public transportation

Data Type Ordinal

Values		Comments
Original	**Replacement**	
0-3	-	-
8	CM	(1.21%)
9	M	(0.00%)
-4	CM	Conditionally missing as form not required. (3.22%)

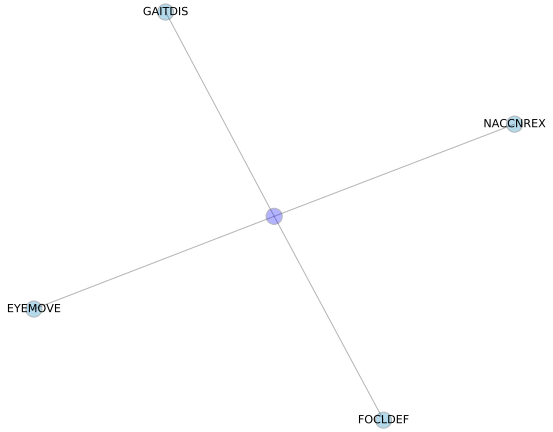
Dependency Trigger -

Relationship -

Inspect Parent	Calculate/Derive	Impute
✗	✗	✓

General Comments A conditionally missing value should be used as a potential fill value if imputed.

B8 - Physical/Neurological Exam Findings



Number of Variables	47
Number of Variables Used	4
Form Required?	✗
Form Missingness	2.30%

NACCNREX

Form Versions	Derived (NACC)	Derived (New)
1.2, 2.0, 3.0	✗	✗

Description Were all findings unremarkable?
Data Type Binary

Values		Comments
Original	**Replacement**	
0-1	-	-
9	M	(0.23%)
-4	CM	Conditionally missing as form not required. (2.30%)

Dependency Trigger -
Relationship -

Inspect Parent	Calculate/Derive	Impute
✗	✗	✓

General Comments A conditionally missing value should not be used as a potential fill value if imputed.

FOCLDEF

Form Versions	Derived (NACC)	Derived (New)
1.2, 2.0	✗	✗

Description Are focal deficits present indicative of central nervous system disorder?
Data Type Binary

Values		Comments
Original	**Replacement**	
0-1	-	-
9	M	(0.80%)
-4	CM	Conditionally missing as form not required. (2.30%)

Dependency Trigger -
Relationship -

Inspect Parent	Calculate/Derive	Impute
✗	✗	✓

General Comments A conditionally missing value should not be used as a potential fill value if imputed.

GAITDIS

Form Versions	Derived (NACC)	Derived (New)
1.2, 2.0	✗	✗

Description Is gait disorder present indicative of central nervous system disorder?
Data Type Binary

Values		Comments
Original	**Replacement**	
0-1	-	-
9	M	(1.13%)
-4	CM	Conditionally missing as form not required. (2.30%)

Dependency Trigger -
Relationship -

Inspect Parent	Calculate/Derive	Impute
✗	✗	✓

General Comments A conditionally missing value should not be used as a potential fill value if imputed.

EYEMOVE

Form Versions	Derived (NACC)	Derived (New)
1.2, 2.0	✗	✗

Description Are there eye movement abnormalities present indicative of central nervous system disorder?
Data Type Binary

Values		Comments
Original	**Replacement**	
0-1	-	-
9	M	(0.77%)
-4	CM	Conditionally missing as form not required. (2.30%)

Dependency Trigger -
Relationship -

Inspect Parent	Calculate/Derive	Impute
✗	✗	✓

General Comments A conditionally missing value should not be used as a potential fill value if imputed.

Dropped Variables

Name NORMEXAM
Description Were there abnormal neurological exam findings?
Reason Only available in version 3.0 of the UDS.

Name PARKSIGN
Description Parkinsonian signs
Reason Only available in version 3.0 of the UDS.

Name RESTTRL
Description Resting tremor - left arm
Reason Only available in version 3.0 of the UDS.

Name RESTTRR
Description Resting tremor - right arm
Reason Only available in version 3.0 of the UDS.

Name SLOWINGL
Description Slowing of fine motor movements - left side
Reason Only available in version 3.0 of the UDS.

Name SLOWINGR
Description Slowing of fine motor movements - right side
Reason Only available in version 3.0 of the UDS.

Name RIGIDL
Description Rigidity - left arm
Reason Only available in version 3.0 of the UDS.

Name RIGIDR
Description Rigidity - right arm
Reason Only available in version 3.0 of the UDS.

Name BRADY
Description Bradykinesia
Reason Only available in version 3.0 of the UDS.

Name PARKGAIT
Description Parkinsonian gait disorder
Reason Only available in version 3.0 of the UDS.

Name POSTINST
Description Postural instability
Reason Only available in version 3.0 of the UDS.

Name CVDSIGNS
Description Neurological sign considered by examiner to be most likely consistent with cerebrovascular disease
Reason Only available in version 3.0 of the UDS.

Name CORTDEF
Description Cortical cognitive deficit (e.g. aphasia, apraxia, neglect)
Reason Only available in version 3.0 of the UDS.

Name SIVDFIND
Description Focal or other neurological findings consistent with subcortical ischemic vascular dementia (SIVD)

Reason Only available in version 3.0 of the UDS.

Name CVDMOTL
Description Motor (may include weakness of combination of face, arm and leg; reflex changes, etc.) - left side
Reason Only available in version 3.0 of the UDS.

Name CVDMOTR
Description Motor (may include weakness of combination of face, arm and leg; reflex changes, etc.) - right side
Reason Only available in version 3.0 of the UDS.

Name CORTVISL
Description Cortical visual field loss - left side
Reason Only available in version 3.0 of the UDS.

Name CORTVISR
Description Cortical visual field loss - right side
Reason Only available in version 3.0 of the UDS.

Name SOMATL
Description Somatosensory loss - left side
Reason Only available in version 3.0 of the UDS.

Name SOMATR
Description Somatosensory loss - right side
Reason Only available in version 3.0 of the UDS.

Name POSTCORT
Description Higher cortical visual problem suggesting posterior cortical atrophy (e.g. prosopagnosia, simultagnosia, Balint's syndrome) or apraxia of gaze
Reason Only available in version 3.0 of the UDS.

Name PSPCBS
Description Findings suggestive of progressive supranuclear palsy (PSP), corticobasal syndrome (CBS) or other related disorders
Reason Only available in version 3.0 of the UDS.

Name EYEPSP
Description Eye movement changes consistent with PSP
Reason Only available in version 3.0 of the UDS.

Name DYSPSP
Description Dysarthria consistent with PSP
Reason Only available in version 3.0 of the UDS.

Name AXIALPSP
Description Axial rigidity consistent with PSP
Reason Only available in version 3.0 of the UDS.

Name GAITPSP
Description Gait disorder consistent with PSP
Reason Only available in version 3.0 of the UDS.

Name APRAXSP
Description Apraxia of speech
Reason Only available in version 3.0 of the UDS.

Name APRAXL
Description Apraxia consistent with CBS - left side

Reason Only available in version 3.0 of the UDS.

Name APRAXR
Description Apraxia consistent with CBS - right side
Reason Only available in version 3.0 of the UDS.

Name CORTSENL
Description Cortical sensory deficits consistent with CBS - left side
Reason Only available in version 3.0 of the UDS.

Name CORTSENR
Description Cortical sensory deficits consistent with CBS - right side
Reason Only available in version 3.0 of the UDS.

Name ATAXL
Description Ataxia consistent with CBS - left side
Reason Only available in version 3.0 of the UDS.

Name ATAXR
Description Ataxia consistent with CBS - right side
Reason Only available in version 3.0 of the UDS.

Name ALIENLML
Description Alien limb consistent with CBS - left side
Reason Only available in version 3.0 of the UDS.

Name ALIENLMR
Description Alien limb consistent with CBS - right side
Reason Only available in version 3.0 of the UDS.

Name DYSTONL
Description Dystonia consistent with CBS, PSP or related disorder - left side
Reason Only available in version 3.0 of the UDS.

Name DYSTONR
Description Dystonia consistent with CBS, PSP or related disorder - right side
Reason Only available in version 3.0 of the UDS.

Name MYOCLLT
Description Myoclonus consistent with CBS - left side
Reason Only available in version 3.0 of the UDS.

Name MYOCLRT
Description Myoclonus consistent with CBS - right side
Reason Only available in version 3.0 of the UDS.

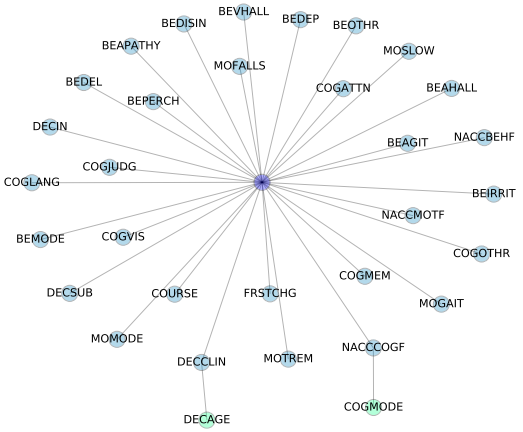
Name ALSFIND
Description Findings suggest amyotrophic lateral sclerosis (ALS) (e.g. muscle wasting, fasciculations, upper motor and/or lower motor neuron signs)
Reason Only available in version 3.0 of the UDS.

Name GAITNPH
Description Normal pressure hydrocephalus - gait apraxia
Reason Only available in version 3.0 of the UDS.

Name OTHNEUR
Description Other findings (e.g. cerebella ataxia, chorea, myoclonus)
Reason Only available in version 3.0 of the UDS.

Name OTHNEURX
Description Other findings - specify
Reason Free-text variable and only available in version 3.0 of the UDS.

B9 - Clinician Judgment of Symptoms



Number of Variables	59
Number of Variables Used	32
Form Required?	✓
Form Missingness	0.00%

DECSUB

Form Versions	Derived (NACC)	Derived (New)
1.2, 2.0, 3.0	✗	✗

Description Does the subject report a decline in memory (relative to previously attained abilities)?

Data Type Binary

Values		Comments
Original	**Replacement**	
0-1	-	-
8	CM	Conditionally missing as value omitted for a reason. (0.00%)
9	M	(0.00%)

Dependency Trigger -

Relationship -

Inspect Parent	Calculate/Derive	Impute
✗	✗	✓

General Comments A conditionally missing value should be used as a potential fill value if imputed.

DECIN

Form Versions	Derived (NACC)	Derived (New)
1.2, 2.0, 3.0	✗	✗

Description Does the co-participant report a decline in subject's memory (relative to previously attained abilities)?

Data Type Binary

Values		Comments
Original	**Replacement**	
0-1	-	-
8	CM	(0.00%)
9	M	(2.90%)

Dependency Trigger -

Relationship -

Inspect Parent	Calculate/Derive	Impute
✗	✗	✓

General Comments A conditionally missing value should be used as a potential fill value if imputed.

DECCLIN

Form Versions	Derived (NACC)	Derived (New)
1.2, 2.0	✗	✗

Description Clinician believes there is a meaningful decline in memory, non-memory cognitive abilities, behaviour, ability to manage his/her affairs, or there are motor/movement changes

Data Type Binary

Values		Comments
Original	**Replacement**	
0-1	-	-
-4	M	Missing as form required and variable independent. (0.00%)

Dependency Trigger 0

Relationship -

Inspect Parent	Calculate/Derive	Impute
✗	✗	✓

General Comments Original forms and NACC's data element dictionaries for versions 1.2 and 2.0 state all following B9 variables are dependent on DECCLIN, but NACC's coding guidebooks for versions 1.2 and 2.0 and NACC's researchers data dictionary contradict this.

DECAGE

Form Versions	Derived (NACC)	Derived (New)
1.2, 2.0, 3.0	✗	✗

Description Based on clinician's assessment, at what age did the cognitive decline begin?

Data Type Continuous

Values		Comments
Original	**Replacement**	
15-110	-	-
888	CM	(40.55%)
999	M	(1.98%)

Dependency Trigger -

Relationship -

Inspect Parent	Calculate/Derive	Impute
✓	✗	✓

General Comments Unable to determine whether values should be updated if parent imputed, as no missingness for parent in data set. A conditionally missing value should be used as a potential fill value if child imputed, as conditionally missing values do not exclusively result from DECCLIN 0 values.

COGMEM

Form Versions	Derived (NACC)	Derived (New)
1.2, 2.0, 3.0	✗	✗

Description Indicate whether the subject currently is meaningfully impaired, relative to previously attained abilities, in memory

Data Type Binary

Values		Comments
Original	**Replacement**	
0-1	-	-
9	M	(0.07%)

Dependency Trigger -

Relationship -

Inspect Parent	Calculate/Derive	Impute
✗	✗	✓

General Comments -

COGJUDG

Form Versions	Derived (NACC)	Derived (New)
1.2, 2.0, 3.0	✗	✗

Description Indicate whether the subject currently is meaningfully impaired, relative to previously attained abilities, in executive function - judgment, planning or problem-solving

Data Type Binary

Values		Comments
Original	**Replacement**	
0-1	-	-
9	M	(0.16%)

Dependency Trigger -

Relationship -

Inspect Parent	Calculate/Derive	Impute
✗	✗	✓

General Comments -

COGLANG

Form Versions	Derived (NACC)	Derived (New)
1.2, 2.0, 3.0	✗	✗

Description Indicate whether the subject currently is meaningfully impaired, relative to previously attained abilities, in language

Data Type Binary

Values		Comments
Original	Replacement	
0-1	-	-
9	M	(0.15%)

Dependency Trigger -
Relationship -

Inspect Parent	Calculate/Derive	Impute
✗	✗	✓

General Comments -

COGVIS

Form Versions	Derived (NACC)	Derived (New)
1.2, 2.0, 3.0	✗	✗

Description Indicate whether the subject currently is meaningfully impaired, relative to previously attained abilities, in visuospatial function
Data Type Binary

Values		Comments
Original	Replacement	
0-1	-	-
9	M	(0.50%)

Dependency Trigger -
Relationship -

Inspect Parent	Calculate/Derive	Impute
✗	✗	✓

General Comments -

COGATTN

Form Versions	Derived (NACC)	Derived (New)
1.2, 2.0, 3.0	✗	✗

Description Indicate whether the subject currently is meaningfully impaired, relative to previously attained abilities, in attention or concentration
Data Type Binary

Values		Comments
Original	Replacement	
0-1	-	-
9	M	(0.42%)

Dependency Trigger -
Relationship -

Inspect Parent	Calculate/Derive	Impute
✗	✗	✓

General Comments -

COGOTHR

Form Versions	Derived (NACC)	Derived (New)
1.2, 2.0, 3.0	✗	✗

Description Indicate whether the subject currently is meaningfully impaired, relative to previously attained abilities, in other cognitive domains
Data Type Binary

Values		Comments
Original	Replacement	
0-1	-	-
9	M	(1.66%)

Dependency Trigger -
Relationship -

Inspect Parent	Calculate/Derive	Impute
✗	✗	✓

General Comments -

NACCCOGF

Form Versions	Derived (NACC)	Derived (New)
1.2, 2.0, 3.0	✗	✗

Description Indicate the predominant symptom that was first recognised as a decline in the subject's cognition
Data Type Categorical

Values		Comments
Original	Replacement	
0	CM	88 in original forms and NACC's coding guidebooks. (40.83%)
1-8	-	-
99	M	(0.31%)

Dependency Trigger CM
Relationship -

Inspect Parent	Calculate/Derive	Impute
✗	✗	✓

General Comments Unable to determine whether a relationship exists due to dropped variables. A conditionally missing value should be used as a potential fill value if imputed.

COGMODE

Form Versions	Derived (NACC)	Derived (New)
1.2, 2.0, 3.0	✗	✗

Description Mode of onset of cognitive symptoms
Data Type Categorical

Values		Comments
Original	**Replacement**	
0	CM	88 in original forms and NACC's coding guidebooks. (40.92%)
1-4	-	-
99	M	(0.80%)

Dependency Trigger -
Relationship -

Inspect Parent	Calculate/Derive	Impute
✓	✗	✓

General Comments Only missing and conditionally missing values should be updated if parent imputed. A conditionally missing value should be used as a potential fill value if child imputed, as conditionally missing values do not exclusively result from NACCCOGF CM (or M) values.

BEAPATHY

Form Versions	Derived (NACC)	Derived (New)
1.2, 2.0, 3.0	✗	✗

Description Subject currently manifests meaningful change in behaviour - Apathy, withdrawal
Data Type Binary

Values		Comments
Original	**Replacement**	
0-1	-	-
9	M	(0.26%)

Dependency Trigger -
Relationship -

Inspect Parent	Calculate/Derive	Impute
✗	✗	✓

General Comments -

BEDEP

Form Versions	Derived (NACC)	Derived (New)
1.2, 2.0, 3.0	✗	✗

Description Subject currently manifests meaningful change in behaviour - Depressed mood
Data Type Binary

Values		Comments
Original	**Replacement**	
0-1	-	-
9	M	(0.35%)

Dependency Trigger -
Relationship -

Inspect Parent	Calculate/Derive	Impute
✗	✗	✓

General Comments -

BEVHALL

Form Versions	Derived (NACC)	Derived (New)
1.2, 2.0, 3.0	✗	✗

Description Subject currently manifests meaningful change in behaviour - Psychosis - Visual hallucinations
Data Type Binary

Values		Comments
Original	**Replacement**	
0-1	-	-
9	M	(0.52%)

Dependency Trigger -
Relationship -

Inspect Parent	Calculate/Derive	Impute
✗	✗	✓

General Comments -

BEAHALL

Form Versions	Derived (NACC)	Derived (New)
1.2, 2.0, 3.0	✗	✗

Description Subject currently manifests meaningful change in behaviour - Psychosis - Auditory hallucinations
Data Type Binary

Values		Comments
Original	Replacement	
0-1	-	-
9	M	(0.57%)

Dependency Trigger -
Relationship -

Inspect Parent	Calculate/Derive	Impute
✗	✗	✓

General Comments -

BEDEL

Form Versions	Derived (NACC)	Derived (New)
1.2, 2.0, 3.0	✗	✗

Description Subject currently manifests meaningful change in behaviour - Psychosis - Abnormal, false or delusional beliefs
Data Type Binary

Values		Comments
Original	Replacement	
0-1	-	-
9	M	(0.50%)

Dependency Trigger -
Relationship -

Inspect Parent	Calculate/Derive	Impute
✗	✗	✓

General Comments -

BEDISIN

Form Versions	Derived (NACC)	Derived (New)
1.2, 2.0, 3.0	✗	✗

Description Subject currently manifests meaningful change in behaviour - Disinhibition
Data Type Binary

Values		Comments
Original	Replacement	
0-1	-	-
9	M	(0.30%)

Dependency Trigger -

Relationship -

Inspect Parent	Calculate/Derive	Impute
✗	✗	✓

General Comments -

BEIRRIT

Form Versions	Derived (NACC)	Derived (New)
1.2, 2.0, 3.0	✗	✗

Description Subject currently manifests meaningful change in behaviour - Irritability
Data Type Binary

Values		Comments
Original	Replacement	
0-1	-	-
9	M	(0.25%)

Dependency Trigger -
Relationship -

Inspect Parent	Calculate/Derive	Impute
✗	✗	✓

General Comments -

BEAGIT

Form Versions	Derived (NACC)	Derived (New)
1.2, 2.0, 3.0	✗	✗

Description Subject currently manifests meaningful change in behaviour - Agitation
Data Type Binary

Values		Comments
Original	Replacement	
0-1	-	-
9	M	(0.24%)

Dependency Trigger -
Relationship -

Inspect Parent	Calculate/Derive	Impute
✗	✗	✓

General Comments -

BEPERCH

Form Versions	Derived (NACC)	Derived (New)
1.2, 2.0, 3.0	✗	✗

Description Subject currently manifests meaningful change in behaviour - Personality change
Data Type Binary

Values		Comments
Original	Replacement	
0-1	-	-
9	M	(0.45%)

Dependency Trigger -
Relationship -

Inspect Parent	Calculate/Derive	Impute
✗	✗	✓

General Comments -

BEOTHR

Form Versions	Derived (NACC)	Derived (New)
1.2, 2.0, 3.0	✗	✗

Description Subject currently manifests meaningful change in behaviour - Other
Data Type Binary

Values		Comments
Original	Replacement	
0-1	-	0 not replaced as 'no/unknown'

Dependency Trigger -
Relationship -

Inspect Parent	Calculate/Derive	Impute
✗	✗	✗

General Comments -

NACCBEHF

Form Versions	Derived (NACC)	Derived (New)
1.2, 2.0, 3.0	✗	✗

Description Indicate the predominant symptom that was first recognised as a decline in the subject's behaviour
Data Type Categorical

Values		Comments
Original	**Replacement**	
0	CM	88 in original forms and NACC's coding guidebooks. (57.96%)
1-10	-	-
99	M	(0.83%)

Dependency Trigger -
Relationship -

Inspect Parent	Calculate/Derive	Impute
✗	✗	✓

General Comments Unable to determine whether a relationship exists due to dropped variables. A conditionally missing value should be used as a potential fill value if imputed.

BEMODE

Form Versions	Derived (NACC)	Derived (New)
1.2, 2.0, 3.0	✗	✗

Description Mode of onset of behavioural symptoms
Data Type Categorical

Values		Comments
Original	**Replacement**	
0	CM	88 in original forms and NACC's coding guidebooks. (58.15%)
1-4	-	-
99	M	(1.24%)

Dependency Trigger -
Relationship -

Inspect Parent	Calculate/Derive	Impute
✗	✗	✓

General Comments NACC's documentation suggests variable is dependent on NACCBEHF but dependency does not hold. A conditionally missing value should be used as a potential fill value if imputed.

MOGAIT

Form Versions	Derived (NACC)	Derived (New)
1.2, 2.0, 3.0	✗	✗

Description Indicate whether the subject currently has meaningful changes in motor function - Gait disorder
Data Type Binary

Values		Comments
Original	Replacement	
0-1	-	-
9	M	(0.49%)

Dependency Trigger -
Relationship -

Inspect Parent	Calculate/Derive	Impute
✗	✗	✓

General Comments -

MOFALLS

Form Versions	Derived (NACC)	Derived (New)
1.2, 2.0, 3.0	✗	✗

Description Indicate whether the subject currently has meaningful changes in motor function - Falls
Data Type Binary

Values		Comments
Original	Replacement	
0-1	-	-
9	M	(0.49%)

Dependency Trigger -
Relationship -

Inspect Parent	Calculate/Derive	Impute
✗	✗	✓

General Comments -

MOTREM

Form Versions	Derived (NACC)	Derived (New)
1.2, 2.0, 3.0	✗	✗

Description Indicate whether the subject currently has meaningful changes in motor function - Tremor
Data Type Binary

Values		Comments
Original	Replacement	
0-1	-	-
9	M	(0.29%)

Dependency Trigger -
Relationship -

Inspect Parent	Calculate/Derive	Impute
✗	✗	✓

General Comments -

MOSLOW

Form Versions	Derived (NACC)	Derived (New)
1.2, 2.0, 3.0	✗	✗

Description Indicate whether the subject currently has meaningful changes in motor function - Slowness
Data Type Binary

Values		Comments
Original	Replacement	
0-1	-	-
9	M	(0.32%)

Dependency Trigger -
Relationship -

Inspect Parent	Calculate/Derive	Impute
✗	✗	✓

General Comments -

NACCMOTF

Form Versions	Derived (NACC)	Derived (New)
1.2, 2.0, 3.0	✗	✗

Description Indicate the predominant symptom that was first recognised as a decline in the subject's motor function
Data Type Categorical

Values		Comments
Original	Replacement	
0	CM	88 in original forms and NACC's coding guidebooks. (78.34%)
1-4	-	-
99	M	(0.70%)

Dependency Trigger -
Relationship -

Inspect Parent	Calculate/Derive	Impute
✗	✗	✓

General Comments NACC's documentation suggests a relationship exists with MOGAIT, MOFALLS, MOTREM and MOSLOW but relationship does not hold. A conditionally missing value should be used as a potential fill value if imputed.

MOMODE

Form Versions	Derived (NACC)	Derived (New)
1.2, 2.0, 3.0	✗	✗

Description Mode of onset of motor symptoms
Data Type Categorical

Values		Comments
Original	**Replacement**	
0	CM	88 in original forms and NACC's coding guidebooks. (78.45%)
1-4	-	-
99	M	(1.10%)

Dependency Trigger -
Relationship -

Inspect Parent	Calculate/Derive	Impute
✗	✗	✓

General Comments NACC's documentation suggests variable is dependent on NACCMOTF but dependency does not hold. A conditionally missing value should be used as a potential fill value if imputed.

COURSE

Form Versions	Derived (NACC)	Derived (New)
1.2, 2.0, 3.0	✗	✗

Description Overall course of decline of cognitive/behavioural/motor syndrome
Data Type Categorical

Values		Comments
Original	**Replacement**	
1-5	-	-
8	CM	(40.42%)
9	M	(1.32%)

Dependency Trigger -
Relationship COURSE = CM if NACCCOGF, NACCBEHF and NACCMOTF = CM

Inspect Parent	Calculate/Derive	Impute
✗	✓	✓

General Comments Only missing values should be updated if NACCCOGF, NACCBEHF and/or NACCMOTF imputed. A conditionally missing value should not be used as a potential fill value if variable imputed.

FRSTCHG

Form Versions	Derived (NACC)	Derived (New)
1.2, 2.0, 3.0	✗	✗

Description Indicate the predominant domain that was first recognised as changed in the subject

Data Type Categorical

Values		Comments
Original	**Replacement**	
1-3	-	-
8	CM	(40.42%)
9	M	(0.84%)

Dependency Trigger -

Relationship FRSTCHG = CM if NACCCOGF, NACCBEHF and NACCMOTF = CM

Inspect Parent	Calculate/Derive	Impute
✗	✓	✓

General Comments NACC's documentation suggests a more complex relationship exists with NACCCOGF, NACCBEHF and NACCMOTF but only relationship stated holds. Only missing values should be updated if NACCCOGF, NACCBEHF and/or NACCMOTF imputed. A conditionally missing value should not be used as a potential fill value if variable imputed.

Dropped Variables

Name COGFLUC
Description Indicate whether the subject currently has fluctuating cognition
Reason Not available in version 1.2 of the UDS.

Name COGOTHRX
Description Other cognitive impairment - specify
Reason Free-text variable.

Name NACCCGFX
Description Other predominant symptom first recognised as a decline in the subject's cognition - specify
Reason Free-text variable.

Name COGMODEX
Description Other mode of onset of cognitive symptoms - specify
Reason Free-text variable.

Name BEVWELL
Description If yes, are the (visual) hallucinations well-formed and detailed?
Reason Not available in version 1.2 of the UDS.

Name BEREM
Description Subject currently manifests meaningful change in behaviour - REM sleep behaviour disorder (RBD)
Reason Not available in version 1.2 of the UDS.

Name BEOTHRX
Description Subject currently manifests meaningful change in behaviour, other - specify
Reason Free-text variable.

Name NACCBEFX
Description Other predominant symptom first recognised as a decline in the subject's behaviour - specify
Reason Free-text variable.

Name BEMODEX
Description Other mode of onset of behavioural symptoms - specify
Reason Free-text variable.

Name MOMODEX
Description Other mode of onset of motor symptoms - specify
Reason Free-text variable.

Name MOMOPARK
Description Were changes in motor function suggestive of Parkinsonism?
Reason Not available in version 1.2 of the UDS.

Name DECCLCOG
Description Based on the clinician's judgment, is the subject currently experiencing meaningful impairment in cognition?
Reason Only available in version 3.0 of the UDS.

Name COGORI
Description Indicate whether the subject currently is meaningfully impaired, relative to previously attained abilities, in orientation

Reason Only available in version 3.0 of the UDS.

Name COGFLAGO
Description At what age did the fluctuating cognition begin?
Reason Only available in version 3.0 of the UDS.

Name DECCLBE
Description Based on the clinician's judgment, is the subject currently experiencing any kind of behavioural symptoms?
Reason Only available in version 3.0 of the UDS.

Name BEVHAGO
Description If well-formed, clear-cut visual hallucinations, at what age did these hallucinations begin?
Reason Only available in version 3.0 of the UDS.

Name BEREMAGO
Description If yes, at what age did the RBD begin?
Reason Only available in version 3.0 of the UDS.

Name BEANX
Description Subject currently manifests meaningful change in behaviour - Anxiety
Reason Only available in version 3.0 of the UDS.

Name BEAGE
Description Based on the clinician's assessment, at what age did the behavioural symptoms begin?
Reason Only available in version 3.0 of the UDS.

Name DECCLMOT
Description Based on the clinician's judgment, is the subject currently experiencing any motor symptoms?
Reason Only available in version 3.0 of the UDS.

Name PARKAGE
Description If yes, at what age did the motor symptoms suggestive of Parkinsonism begin?
Reason Only available in version 3.0 of the UDS.

Name MOMOALS
Description Were changes in motor function suggestive of amyotrophic lateral sclerosis (ALS)?
Reason Only available in version 3.0 of the UDS.

Name ALSAGE
Description If yes, at what age did the motor symptoms suggestive of ALS begin?
Reason Only available in version 3.0 of the UDS.

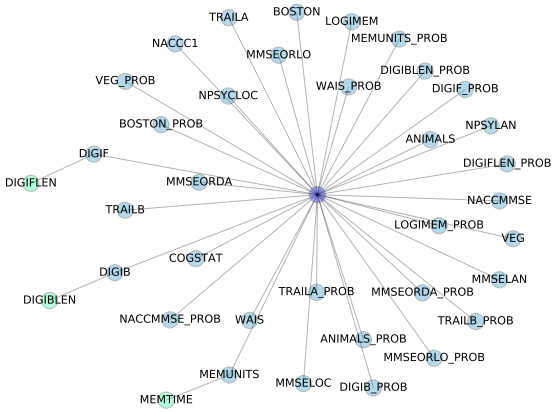
Name MOAGE
Description Based on the clinician's assessment, at what age did the motor changes begin?
Reason Only available in version 3.0 of the UDS.

Name LBDEVAL
Description Is the subject a potential candidate for further evaluation for Lewy body disease?
Reason Only available in version 3.0 of the UDS.

Name FTLDEVAL
Description Is the subject a potential candidate for further evaluation for frontotemporal lobar degeneration (FTLD)?
Reason Only available in version 3.0 of the UDS.

Name B9CHG
Description Indicates changes in information reported at previous visit
Reason Information provided would not be available at an initial visit and only available in version 1.2 of the UDS.

C1 - Neuropsychological Battery



Number of Variables	48
Number of Variables Used	37
Form Required?	✓
Form Missingness	0.00%

MMSELOC

Form Versions	Derived (NACC)	Derived (New)
1.2, 2.0, 3.0	✗	✗

Description Administration of the Mini-Mental State Examination (MMSE) was:
Data Type Categorical

Values		Comments
Original	Replacement	
1-3	-	-
-4	M	Missing as form required and variable independent. (1.87%)

Dependency Trigger -
Relationship -

Inspect Parent	Calculate/Derive	Impute
✗	✗	✓

General Comments -

MMSELAN

Form Versions	Derived (NACC)	Derived (New)
1.2, 2.0, 3.0	✗	✗

Description Language of MMSE administration
Data Type Categorical

Values		Comments
Original	Replacement	
1-3	-	-
-4	M	Missing as form required and variable independent. (1.87%)

Dependency Trigger -
Relationship -

Inspect Parent	Calculate/Derive	Impute
✗	✗	✓

General Comments -

MMSEORDA

Form Versions	Derived (NACC)	Derived (New)
1.2, 2.0, 3.0	✗	✗

Description Orientation subscale score - Time
Data Type Continuous

Values		Comments
Original	**Replacement**	
0-5	-	-
95-98	CM	Conditionally missing as MMSEORDA_PROB generated. (3.15%)
-4	M	Missing as form required and variable independent. (1.87%)

Dependency Trigger -
Relationship -

Inspect Parent	Calculate/Derive	Impute
✗	✗	✓

General Comments A conditionally missing value should be used as a potential fill value if imputed.

MMSEORDA_PROB

Form Versions	Derived (NACC)	Derived (New)
1.2, 2.0, 3.0	✗	✓

Description Reason an answer was not provided for MMSEORDA
Data Type Categorical

Values		Comments
Original	**Replacement**	
95-98	-	-
CM	-	Placeholder for 0-5 and M values in MMSEORDA. (96.85%)

Dependency Trigger -
Relationship MMSEORDA_PROB = 95-98 if MMSEORDA = CM else CM

Inspect Parent	Calculate/Derive	Impute
✗	✓	✓

General Comments Derived from MMSEORDA. Values should be updated if MMSEORDA imputed. A conditionally missing value should not be used as a potential fill value if variable imputed.

MMSEORLO

Form Versions	Derived (NACC)	Derived (New)
1.2, 2.0, 3.0	✗	✗

Description Orientation subscale score - Place
Data Type Continuous

Values		Comments
Original	**Replacement**	
0-5	-	-
95-98	CM	Conditionally missing as MMSEORLO_PROB generated. (3.16%)
-4	M	Missing as form required and variable independent. (1.87%)

Dependency Trigger -
Relationship -

Inspect Parent	Calculate/Derive	Impute
✗	✗	✓

General Comments A conditionally missing value should be used as a potential fill value if imputed.

MMSEORLO_PROB

Form Versions	Derived (NACC)	Derived (New)
1.2, 2.0, 3.0	✗	✓

Description Reason an answer was not provided for MMSEORLO
Data Type Categorical

Values		Comments
Original	**Replacement**	
95-98	-	-
CM	-	Placeholder for 0-5 and M values in MMSEORLO. (96.84%)

Dependency Trigger -
Relationship MMSEORLO_PROB = 95-98 if MMSEORLO = CM else CM

Inspect Parent	Calculate/Derive	Impute
✗	✓	✓

General Comments Derived from MMSEORLO. Values should be updated if MMSEORLO imputed. A conditionally missing value should not be used as a potential fill value if variable imputed.

NACCMMSE

Form Versions	Derived (NACC)	Derived (New)
1.2, 2.0, 3.0	✗	✗

Description Total MMSE score (using D-L-R-O-W)
Data Type Continuous

Values		Comments
Original	**Replacement**	
0-30	-	-
88	Dropped	Dropped to avoid multiple types of conditionally missing but irrelevant for the data set used as introduced in version 3.0.
95-98	CM	Conditionally missing as NACCMMSE_PROB generated. (3.45%)
-4	M	Missing as form required and variable independent. (1.87%)

Dependency Trigger -
Relationship -

Inspect Parent	Calculate/Derive	Impute
✗	✗	✓

General Comments A conditionally missing value should be used as a potential fill value if imputed.

NACCMMSE_PROB

Form Versions	Derived (NACC)	Derived (New)
1.2, 2.0, 3.0	✗	✓

Description Reason an answer was not provided for NACCMMSE
Data Type Categorical

Values		Comments
Original	**Replacement**	
95-98	-	-
CM	-	Placeholder for 0-30 and M values in NACCMMSE. (96.55%)

Dependency Trigger -
Relationship NACCMMSE_PROB = 95-98 if NACCMMSE = CM else CM

Inspect Parent	Calculate/Derive	Impute
✗	✓	✓

General Comments Derived from NACCMMSE. Values should be updated if NACCMMSE imputed. A conditionally missing value should not be used as a potential fill value if variable imputed.

NPSYCLOC

Form Versions	Derived (NACC)	Derived (New)
1.2, 2.0, 3.0	✗	✗

Description The remainder of the battery was administered:
Data Type Categorical

Values		Comments
Original	**Replacement**	
1-3	-	-
-4	M	Missing as form required and variable independent. (1.87%)

Dependency Trigger -
Relationship -

Inspect Parent	Calculate/Derive	Impute
✗	✗	✓

General Comments -

NPSYLAN

Form Versions	Derived (NACC)	Derived (New)
1.2, 2.0, 3.0	✗	✗

Description Language of test administration
Data Type Categorical

Values		Comments
Original	**Replacement**	
1-3	-	-
-4	M	Missing as form required and variable independent. (1.87%)

Dependency Trigger -
Relationship -

Inspect Parent	Calculate/Derive	Impute
✗	✗	✓

General Comments -

LOGIMEM

Form Versions	Derived (NACC)	Derived (New)
1.2, 2.0, 3.0	✗	✗

Description Total number of story units recalled from this current test administration
Data Type Continuous

Values		Comments
Original	**Replacement**	
0-25	-	-
95-98	CM	Conditionally missing as LOGIMEM_PROB generated. (9.13%)
-4	M	Missing as form required and variable independent. (1.87%)

Dependency Trigger -
Relationship -

Inspect Parent	Calculate/Derive	Impute
✗	✗	✓

General Comments A conditionally missing value should be used as a potential fill value if imputed.

LOGIMEM_PROB

Form Versions	Derived (NACC)	Derived (New)
1.2, 2.0, 3.0	✗	✓

Description Reason an answer was not provided for LOGIMEM
Data Type Categorical

Values		Comments
Original	Replacement	
95-98	-	-
CM	-	Placeholder for 0-25 and M values in LOGIMEM. (90.87%)

Dependency Trigger -
Relationship LOGIMEM_PROB = 95-98 if LOGIMEM = CM else CM

Inspect Parent	Calculate/Derive	Impute
✗	✓	✓

General Comments Derived from LOGIMEM. Values should be updated if LOGIMEM imputed. A conditionally missing value should not be used as a potential fill value if variable imputed.

DIGIF

Form Versions	Derived (NACC)	Derived (New)
1.2, 2.0, 3.0	✗	✗

Description Digit span forward trials correct
Data Type Continuous

Values		Comments
Original	Replacement	
0-12	-	-
95-98	CM	Conditionally missing as DIGIF_PROB generated. (7.84%)
-4	M	Missing as form required and variable independent. (1.87%)

Dependency Trigger CM

Relationship -

Inspect Parent	Calculate/Derive	Impute
✗	✗	✓

General Comments A conditionally missing value should be used as a potential fill value if imputed.

DIGIFLEN

Form Versions	Derived (NACC)	Derived (New)
1.2, 2.0, 3.0	✗	✗

Description Digit span forward length
Data Type Continuous

Values		Comments
Original	Replacement	
0-8	-	-
95-98	CM	Conditionally missing as DIGIFLEN_PROB generated. (7.88%)
-4	CM	Conditionally missing even though form required as variable dependent. (1.87%)

Dependency Trigger -
Relationship -

Inspect Parent	Calculate/Derive	Impute
✓	✗	✓

General Comments Values should be updated if parent imputed. A conditionally missing value should be used as a potential fill value if child imputed, as conditionally missing values do not exclusively result from DIGIF CM (or M) values.

DIGIF_PROB

Form Versions	Derived (NACC)	Derived (New)
1.2, 2.0, 3.0	✗	✓

Description Reason an answer was not provided for DIGIF
Data Type Categorical

Values		Comments
Original	Replacement	
95-98	-	-
CM	-	Placeholder for 0-12 and M values in DIGIF. (92.16%)

Dependency Trigger -
Relationship DIGIF_PROB = 95-98 if DIGIF = CM else CM

Inspect Parent	Calculate/Derive	Impute
✗	✓	✓

General Comments Derived from DIGIF. Values should be updated if DIGIF imputed. A conditionally missing value should not be used as a potential fill value if variable imputed.

DIGIFLEN_PROB

Form Versions	Derived (NACC)	Derived (New)
1.2, 2.0, 3.0	✗	✓

Description Reason an answer was not provided for DIGIFLEN
Data Type Categorical

Values		Comments
Original	Replacement	
95-98	-	-
CM	-	Placeholder for 0-8 and CM values resulting from dependency in DIGIFLEN. (92.12%)

Dependency Trigger -
Relationship DIGIFLEN_PROB = 95-98 if DIGIFLEN = CM else CM

Inspect Parent	Calculate/Derive	Impute
✗	✓	✓

General Comments Derived from DIGIFLEN. Values should be updated if DIGIFLEN imputed. A conditionally missing value should not be used as a potential fill value if variable imputed.

DIGIB

Form Versions	Derived (NACC)	Derived (New)
1.2, 2.0, 3.0	✗	✗

Description Digit span backward trials correct
Data Type Continuous

Values		Comments
Original	Replacement	
0-12	-	-
95-98	CM	Conditionally missing as DIGIB_PROB generated. (8.18%)
-4	M	Missing as form required and variable independent. (1.87%)

Dependency Trigger CM
Relationship -

Inspect Parent	Calculate/Derive	Impute
✗	✗	✓

General Comments A conditionally missing value should be used as a potential fill value if imputed.

DIGIBLEN

Form Versions	Derived (NACC)	Derived (New)
1.2, 2.0, 3.0	✗	✗

Description Digit span backward length
Data Type Continuous

Values		Comments
Original	Replacement	
0-8	-	0-7 in original forms and NACC's coding guidebooks for versions 1.2 and 2.0.
95-98	CM	Conditionally missing as DIGIBLEN_PROB generated. (8.18%)
-4	CM	Conditionally missing even though form required as variable dependent. (1.87%)

Dependency Trigger -
Relationship -

Inspect Parent	Calculate/Derive	Impute
✓	✗	✓

General Comments Values should be updated if parent imputed. A conditionally missing value should be used as a potential fill value if child imputed, as conditionally missing values do not exclusively result from DIGIB CM (or M) values.

DIGIB_PROB

Form Versions	Derived (NACC)	Derived (New)
1.2, 2.0, 3.0	✗	✓

Description Reason an answer was not provided for DIGIB
Data Type Categorical

Values		Comments
Original	Replacement	
95-98	-	-
CM	-	Placeholder for 0-12 and M values in DIGIB. (91.82%)

Dependency Trigger -
Relationship DIGIB_PROB = 95-98 if DIGIB = CM else CM

Inspect Parent	Calculate/Derive	Impute
✗	✓	✓

General Comments Derived from DIGIB. Values should be updated if DIGIB imputed. A conditionally missing value should not be used as a potential fill value if variable imputed.

DIGIBLEN_PROB

Form Versions	Derived (NACC)	Derived (New)
1.2, 2.0, 3.0	✗	✓

Description Reason an answer was not provided for DIGIBLEN
Data Type Categorical

Values		Comments
Original	**Replacement**	
95-98	-	-
CM	-	Placeholder for 0-8 and CM values resulting from dependency in DIGIBLEN. (91.82%)

Dependency Trigger -
Relationship DIGIBLEN_PROB = 95-98 if DIGIBLEN = CM else CM

Inspect Parent	Calculate/Derive	Impute
✗	✓	✓

General Comments Derived from DIGIBLEN. Values should be updated if DIGIBLEN imputed. A conditionally missing value should not be used as a potential fill value if variable imputed.

ANIMALS

Form Versions	Derived (NACC)	Derived (New)
1.2, 2.0, 3.0	✗	✗

Description Animals - Total number of animals named in 60 seconds
Data Type Continuous

Values		Comments
Original	**Replacement**	
0-77	-	-
95-98	CM	Conditionally missing as ANIMALS_PROB generated. (7.04%)
-4	M	Missing as form required and variable independent. (1.87%)

Dependency Trigger -
Relationship -

Inspect Parent	Calculate/Derive	Impute
✗	✗	✓

General Comments A conditionally missing value should be used as a potential fill value if imputed.

ANIMALS_PROB

Form Versions	Derived (NACC)	Derived (New)
1.2, 2.0, 3.0	✗	✓

Description Reason an answer was not provided for ANIMALS
Data Type Categorical

Values		Comments
Original	**Replacement**	
95-98	-	-
CM	-	Placeholder for 0-77 and M values in ANIMALS. (92.96%)

Dependency Trigger Relationship -
ANIMALS_PROB = 95-98 if ANIMALS = CM else CM

Inspect Parent	Calculate/Derive	Impute
✗	✓	✓

General Comments Derived from ANIMALS. Values should be updated if ANIMALS imputed. A conditionally missing value should not be used as a potential fill value if variable imputed.

VEG

Form Versions	Derived (NACC)	Derived (New)
1.2, 2.0, 3.0	✗	✗

Description Vegetables - Total number of vegetables named in 60 seconds
Data Type Continuous

Values		Comments
Original	**Replacement**	
0-77	-	-
95-98	CM	Conditionally missing as VEG_PROB generated. (8.63%)
-4	M	Missing as form required and variable independent. (1.87%)

Dependency Trigger Relationship -
-

Inspect Parent	Calculate/Derive	Impute
✗	✗	✓

General Comments A conditionally missing value should be used as a potential fill value if imputed.

VEG_PROB

Form Versions	Derived (NACC)	Derived (New)
1.2, 2.0, 3.0	✗	✓

Description Reason an answer was not provided for VEG
Data Type Categorical

Values		Comments
Original	**Replacement**	
95-98	-	-
CM	-	Placeholder for 0-77 and M values in VEG. (91.37%)

Dependency Trigger -
Relationship VEG_PROB = 95-98 if VEG = CM else CM

Inspect Parent	Calculate/Derive	Impute
✗	✓	✓

General Comments Derived from VEG. Values should be updated if VEG imputed. A conditionally missing value should not be used as a potential fill value if variable imputed.

TRAILA

Form Versions	Derived (NACC)	Derived (New)
1.2, 2.0, 3.0	✗	✗

Description Trail Making Test Part A - Total number of seconds to complete
Data Type Continuous

Values		Comments
Original	**Replacement**	
0-150	-	-
995-998	CM	Conditionally missing as TRAILA_PROB generated. (11.04%)
-4	M	Missing as form required and variable independent. (1.87%)

Dependency Trigger -
Relationship -

Inspect Parent	Calculate/Derive	Impute
✗	✗	✓

General Comments A conditionally missing value should be used as a potential fill value if imputed.

TRAILA_PROB

Form Versions	Derived (NACC)	Derived (New)
1.2, 2.0, 3.0	✗	✓

Description Reason an answer was not provided for TRAILA
Data Type Categorical

Values		Comments
Original	**Replacement**	
995-998	-	-
CM	-	Placeholder for 0-150 and M values in TRAILA. (88.96%)

Dependency Trigger Relationship -
TRAILA_PROB = 995-998 if TRAILA = CM else CM

Inspect Parent	Calculate/Derive	Impute
✗	✓	✓

General Comments Derived from TRAILA. Values should be updated if TRAILA imputed. A conditionally missing value should not be used as a potential fill value if variable imputed.

TRAILB

Form Versions	Derived (NACC)	Derived (New)
1.2, 2.0, 3.0	✗	✗

Description Trail Making Test Part B - Total number of seconds to complete
Data Type Continuous

Values		Comments
Original	**Replacement**	
0-300	-	-
995-998	CM	Conditionally missing as TRAILB_PROB generated. (19.59%)
-4	M	Missing as form required and variable independent. (1.87%)

Dependency Trigger Relationship -
-

Inspect Parent	Calculate/Derive	Impute
✗	✗	✓

General Comments A conditionally missing value should be used as a potential fill value if imputed.

TRAILB_PROB

Form Versions	Derived (NACC)	Derived (New)
1.2, 2.0, 3.0	✗	✓

Description Reason an answer was not provided for TRAILB
Data Type Categorical

Values		Comments
Original	**Replacement**	
995-998	-	-
CM	-	Placeholder for 0-300 and M values in TRAILB. (80.41%)

Dependency Trigger Relationship -
TRAILB_PROB = 995-998 if TRAILB = CM else CM

Inspect Parent	Calculate/Derive	Impute
✗	✓	✓

General Comments Derived from TRAILB. Values should be updated if TRAILB imputed. A conditionally missing value should not be used as a potential fill value if variable imputed.

WAIS

Form Versions	Derived (NACC)	Derived (New)
1.2, 2.0	✗	✗

Description Wechsler Adult Intelligence Scale (Revised) (WAIS-R) Digit Symbol
Data Type Continuous

Values		Comments
Original	**Replacement**	
0-93	-	-
95-98	CM	Conditionally missing as WAIS_PROB generated. (15.12%)
-4	M	Missing as form required and variable independent. (1.87%)

Dependency Trigger Relationship -
-

Inspect Parent	Calculate/Derive	Impute
✗	✗	✓

General Comments A conditionally missing value should be used as a potential fill value if imputed.

WAIS_PROB

Form Versions	Derived (NACC)	Derived (New)
1.2, 2.0	✗	✓

Description Reason an answer was not provided for WAIS
Data Type Categorical

Values		Comments
Original	**Replacement**	
95-98	-	-
CM	-	Placeholder for 0-93 and M values in WAIS. (84.88%)

Dependency Trigger Relationship -
WAIS_PROB = 95-98 if WAIS = CM else CM

Inspect Parent	Calculate/Derive	Impute
✗	✓	✓

General Comments Derived from WAIS. Values should be updated if WAIS imputed. A conditionally missing value should not be used as a potential fill value if variable imputed.

MEMUNITS

Form Versions	Derived (NACC)	Derived (New)
1.2, 2.0, 3.0	✗	✗

Description Logical Memory IIA - Delayed - Total number of story units recalled
Data Type Continuous

Values		Comments
Original	**Replacement**	
0-25	-	-
95-98	CM	Conditionally missing as MEMUNITS_PROB generated. (9.38%)
-4	M	Missing as form required and variable independent. (1.87%)

Dependency Trigger Relationship CM
-

Inspect Parent	Calculate/Derive	Impute
✗	✗	✓

General Comments NACC's documentation suggests variable is dependent on LOGIMEM but dependency does not hold. A conditionally missing value should be used as a potential fill value if imputed.

MEMTIME

Form Versions	Derived (NACC)	Derived (New)
1.2, 2.0, 3.0	✗	✗

Description Logical Memory IIA - Delayed - Time elapsed since Logical Memory IA - Immediate
Data Type Continuous

Values		Comments
Original	**Replacement**	
0-85	-	-
99	M	(1.57%)
-4	CM	Conditionally missing even though form required as variable dependent. (11.25%)

Dependency Trigger -
Relationship -

Inspect Parent	Calculate/Derive	Impute
✓	✗	✓

General Comments 88 (CM) value in original forms and NACC's coding guidebooks for versions 1.2 and 2.0 but omitted from NACC's researchers data dictionary and data set. Values should be updated if parent imputed. A conditionally missing value should not be used as a potential fill value if child imputed.

MEMUNITS_PROB

Form Versions	Derived (NACC)	Derived (New)
1.2, 2.0, 3.0	✗	✓

Description Reason an answer was not provided for MEMUNITS
Data Type Categorical

Values		Comments
Original	**Replacement**	
95-98	-	-
CM	-	Placeholder for 0-25 and M values in MEMUNITS. (90.62%)

Dependency Trigger -
Relationship MEMUNITS_PROB = 95-98 if MEMUNITS = CM else CM

Inspect Parent	Calculate/Derive	Impute
✗	✓	✓

General Comments Derived from MEMUNITS. Values should be updated if MEMUNITS imputed. A conditionally missing value should not be used as a potential fill value if variable imputed.

BOSTON

Form Versions	Derived (NACC)	Derived (New)
1.2, 2.0, 3.0	✗	✗

Description Boston Naming Test (30) - Total score
Data Type Continuous

Values		Comments
Original	**Replacement**	
0-30	-	-
95-98	CM	Conditionally missing as BOSTON_PROB generated. (8.91%)
-4	M	Missing as form required and variable independent. (1.87%)

Dependency Trigger -
Relationship -

Inspect Parent	Calculate/Derive	Impute
✗	✗	✓

General Comments A conditionally missing value should be used as a potential fill value if imputed.

BOSTON_PROB

Form Versions	Derived (NACC)	Derived (New)
1.2, 2.0, 3.0	✗	✓

Description Reason an answer was not provided for BOSTON
Data Type Categorical

Values		Comments
Original	**Replacement**	
95-98	-	-
CM	-	Placeholder for 0-30 and M values in BOSTON. (91.09%)

Dependency Trigger -
Relationship BOSTON_PROB = 95-98 if BOSTON = CM else CM

Inspect Parent	Calculate/Derive	Impute
✗	✓	✓

General Comments Derived from BOSTON. Values should be updated if BOSTON imputed. A conditionally missing value should not be used as a potential fill value if variable imputed.

COGSTAT

Form Versions	Derived (NACC)	Derived (New)
1.2, 2.0, 3.0	✗	✗

Description Per clinician, based on the neuropsychological examination, the subject's cognitive status is deemed
Data Type Categorical

Values		Comments
Original	**Replacement**	
0-4	-	-
9	M	(0.00%)
-4	M	Missing as form required and variable independent. (1.87%)

Dependency Trigger -
Relationship -

Inspect Parent	Calculate/Derive	Impute
✗	✗	✓

General Comments -

NACCC1

Form Versions	Derived (NACC)	Derived (New)
1.2, 2.0, 3.0	✓	✗

Description Form date discrepancy between UDS Form A1 and Form C1
Data Type Binary

Values		Comments
Original	**Replacement**	
0-1	-	-
-4	CM	Conditionally missing as cannot be sensibly imputed. (0.00%)

Dependency Trigger -
Relationship -

Inspect Parent	Calculate/Derive	Impute
✗	✗	✗

General Comments -

Dropped Variables

Name MMSELANX
Description Language of MMSE administration, other - specify
Reason Free-text variable.

Name PENTAGON
Description Intersecting pentagon subscale score
Reason Not available in version 1.2 of the UDS.

Name NPSYLANX
Description Language of test administration, other - specify
Reason Free-text variable.

Name LOGIMO
Description If this test has been administered to the subject within the past 3 months, specify the date previously administered (month)
Reason Information provided would not be available at an initial visit.

Name LOGIDAY
Description If this test has been administered to the subject within the past 3 months, specify the date previously administered (day)
Reason Information provided would not be available at an initial visit.

Name LOGIYR
Description If this test has been administered to the subject within the past 3 months, specify the date previously administered (year)
Reason Information provided would not be available at an initial visit.

Name LOGIPREV
Description Total score from the previous test administration
Reason Information provided would not be available at an initial visit.

Name TRAILARR
Description Part A - Number of commission errors
Reason Not available in version 1.2 of the UDS.

Name TRAILALI
Description Part A - Number of correct lines
Reason Not available in version 1.2 of the UDS.

Name TRAILBRR
Description Part B - Number of commission errors
Reason Not available in version 1.2 of the UDS.

Name TRAILBLI
Description Part B - Number of correct lines
Reason Not available in version 1.2 of the UDS.

Name MMSECOMP
Description Was any part of the MMSE completed?
Reason Only available in version 3.0 of the UDS.

Name MMSEVIS
Description Subject was unable to complete one or more sections due to visual impairment
Reason Only available in version 3.0 of the UDS.

Name MMSEHEAR

Description Subject was unable to complete one or more sections due to hearing impairment
Reason Only available in version 3.0 of the UDS.

Name UDSBENTC
Description Total score for copy of Benson figure
Reason Only available in version 3.0 of the UDS.

Name UDSBENTD
Description Total score for 10 to 15 minute delayed drawing of Benson figure
Reason Only available in version 3.0 of the UDS.

Name UDSBENRS
Description Recognised original stimulus from among four options
Reason Only available in version 3.0 of the UDS.

Name UDSVERFC
Description Number of correct F-words generated in 1 minute
Reason Only available in version 3.0 of the UDS.

Name UDSVERFN
Description Number of F-words repeated in 1 minute
Reason Only available in version 3.0 of the UDS.

Name UDSVERNF
Description Number of non-F-words and rule violation errors in 1 minute
Reason Only available in version 3.0 of the UDS.

Name UDSVERLC
Description Number of correct L-words generated in 1 minute
Reason Only available in version 3.0 of the UDS.

Name UDSVERLR
Description Number of L-words repeated in 1 minute
Reason Only available in version 3.0 of the UDS.

Name UDSVERLN
Description Number of non-L-words and rule violation errors in 1 minute
Reason Only available in version 3.0 of the UDS.

Name UDSVERTN
Description Total number of correct F-words and L-words
Reason Only available in version 3.0 of the UDS.

Name UDSVERTE
Description Total number of F-word and L-word repetition errors
Reason Only available in version 3.0 of the UDS.

Name UDSVERTI
Description Total number of non-F/L-words and rule violation errors
Reason Only available in version 3.0 of the UDS.

C2 - Neuropsychological Battery

Number of Variables	47
Number of Variables Used	0

Dropped Variables

Name MOCACOMP
Description Was any part of the Montreal Cognitive Assessment (MoCA) administered?
Reason Only available in version 3.0 of the UDS.

Name MOCAREAS
Description If no part of MoCA administered, reason code
Reason Only available in version 3.0 of the UDS.

Name MOCALOC
Description Where was MoCA administered?
Reason Only available in version 3.0 of the UDS.

Name MOCALAN
Description Language of MoCA administration
Reason Only available in version 3.0 of the UDS.

Name MOCALANX
Description Language of MoCA administration, other - specify
Reason Free-text variable and only available in version 3.0 of the UDS.

Name MOCAVIS
Description Subject was unable to complete one or more sections due to visual impairment
Reason Only available in version 3.0 of the UDS.

Name MOCAHEAR
Description Subject was unable to complete one or more sections due to hearing impairment
Reason Only available in version 3.0 of the UDS.

Name MOCATOTS
Description MoCA Total Raw Score - Uncorrected
Reason Only available in version 3.0 of the UDS.

Name MOCATRAI
Description MoCA: Visuospatial/executive - Trails
Reason Only available in version 3.0 of the UDS.

Name MOCACUBE
Description MoCA: Visuospatial/executive - Cube
Reason Only available in version 3.0 of the UDS.

Name MOCACLOC
Description MoCA: Visuospatial/executive - Clock contour
Reason Only available in version 3.0 of the UDS.

Name MOCACLON
Description MoCA: Visuospatial/executive - Clock numbers
Reason Only available in version 3.0 of the UDS.

Name MOCACLOH
Description MoCA: Visuospatial/executive - Clock hands
Reason Only available in version 3.0 of the UDS.

Name MOCANAMI
Description MoCA: Language - Naming
Reason Only available in version 3.0 of the UDS.

Name MOCAREGI
Description MoCA: Memory - Registration (two trials)
Reason Only available in version 3.0 of the UDS.

Name MOCADIGI
Description MoCA: Attention - Digits
Reason Only available in version 3.0 of the UDS.

Name MOCALETT
Description MoCA: Attention - Letter A
Reason Only available in version 3.0 of the UDS.

Name MOCASER7
Description MoCA: Attention - Serial 7s
Reason Only available in version 3.0 of the UDS.

Name MOCAREPE
Description MoCA: Language - Repetition
Reason Only available in version 3.0 of the UDS.

Name MOCAFLUE
Description MoCA: Language - Fluency
Reason Only available in version 3.0 of the UDS.

Name MOCAABST
Description MoCA: Abstraction
Reason Only available in version 3.0 of the UDS.

Name MOCARECN
Description MoCA: Delayed recall - No cue
Reason Only available in version 3.0 of the UDS.

Name MOCARECC
Description MoCA: Delayed recall - Category clue
Reason Only available in version 3.0 of the UDS.

Name MOCARECR
Description MoCA: Delayed recall - Recognition
Reason Only available in version 3.0 of the UDS.

Name MOCAORDT
Description MoCA: Orientation - Date
Reason Only available in version 3.0 of the UDS.

Name MOCAORMO
Description MoCA: Orientation - Month
Reason Only available in version 3.0 of the UDS.

Name MOCAORYR
Description MoCA: Orientation - Year
Reason Only available in version 3.0 of the UDS.

Name MOCAORDY
Description MoCA: Orientation - Day
Reason Only available in version 3.0 of the UDS.

Name MOCAORPL
Description MoCA: Orientation - Place
Reason Only available in version 3.0 of the UDS.

Name MOCAORCT
Description MoCA: Orientation - City
Reason Only available in version 3.0 of the UDS.

Name CRAFTVRS
Description Craft Story 21 Recall (Immediate) - Total story units recalled, verbatim scoring
Reason Only available in version 3.0 of the UDS.

Name CRAFTURS
Description Craft Story 21 Recall (Immediate) - Total story units recalled, paraphrase scoring
Reason Only available in version 3.0 of the UDS.

Name DIGFORCT
Description Number Span Test: Forward - Number of correct trials
Reason Only available in version 3.0 of the UDS.

Name DIGFORSL
Description Number Span Test: Forward - Longest span forward
Reason Only available in version 3.0 of the UDS.

Name DIGBACCT
Description Number Span Test: Backward - Number of correct trials
Reason Only available in version 3.0 of the UDS.

Name DIGBACLS
Description Number Span Test: Backward - Longest span backward
Reason Only available in version 3.0 of the UDS.

Name CRAFTDVR
Description Craft Story 21 Recall (Delayed) - Total story units recalled, verbatim scoring
Reason Only available in version 3.0 of the UDS.

Name CRAFTDRE
Description Craft Story 21 Recall (Delayed) - Total story units recalled, paraphrase scoring
Reason Only available in version 3.0 of the UDS.

Name CRAFTDTI
Description Craft Story 21 Recall (Delayed) - Delay time
Reason Only available in version 3.0 of the UDS.

Name CRAFTCUE
Description Craft Story 21 Recall (Delayed) - Cue (boy) needed
Reason Only available in version 3.0 of the UDS.

Name MINTTOTS
Description Multilingual Naming Test (MINT): Total score
Reason Only available in version 3.0 of the UDS.

Name MINTTOTW
Description Multilingual Naming Test (MINT): Total correct without semantic cue
Reason Only available in version 3.0 of the UDS.

Name MINTSCNG
Description Multilingual Naming Test (MINT): Semantic cues - Number given
Reason Only available in version 3.0 of the UDS.

Name MINTSCNC
Description Multilingual Naming Test (MINT): Semantic cues - Number correct with cue
Reason Only available in version 3.0 of the UDS.

Name MINTPCNG
Description Multilingual Naming Test (MINT): Phonemic cues - Number given
Reason Only available in version 3.0 of the UDS.

Name MINTPCNC
Description Multilingual Naming Test (MINT): Phonemic cues - Number correct with cue
Reason Only available in version 3.0 of the UDS.

Name NACCC2
Description Form date discrepancy between UDS Form A1 and Form C2
Reason Only available in version 3.0 of the UDS.

D2 - Clinician-assessed Medical Conditions

Number of Variables	33
Number of Variables Used	0

Dropped Variables

Name CANCER
Description Cancer present in the last 12 months (excluding non-melanoma skin cancer), primary or metastatic
Reason Only available in version 3.0 of the UDS.

Name CANCSITE
Description Cancer primary site - specify
Reason Free-text variable and only available in version 3.0 of the UDS.

Name DIABET
Description Diabetes present at visit
Reason Only available in version 3.0 of the UDS.

Name MYOINF
Description Myocardial infarct present within the past 12 months
Reason Only available in version 3.0 of the UDS.

Name CONGHRT
Description Congestive heart failure present
Reason Only available in version 3.0 of the UDS.

Name AFIBRILL
Description Atrial fibrillation present
Reason Only available in version 3.0 of the UDS.

Name HYPERT
Description Hypertension present
Reason Only available in version 3.0 of the UDS.

Name ANGINA
Description Angina present
Reason Only available in version 3.0 of the UDS.

Name HYPCHOL
Description Hypercholesterolemia present
Reason Only available in version 3.0 of the UDS.

Name VB12DEF
Description B12 deficiency present
Reason Only available in version 3.0 of the UDS.

Name THYDIS
Description Thyroid disease present
Reason Only available in version 3.0 of the UDS.

Name ARTH
Description Arthritis present
Reason Only available in version 3.0 of the UDS.

Name ARTYPE
Description Arthritis type
Reason Only available in version 3.0 of the UDS.

Name ARTYPEX
Description Other arthritis type - specify
Reason Free-text variable and only available in version 3.0 of the UDS.

Name ARTUPEX
Description Arthritis region affected - upper extremity
Reason Only available in version 3.0 of the UDS.

Name ARTLOEX
Description Arthritis region affected - lower extremity
Reason Only available in version 3.0 of the UDS.

Name ARTSPIN
Description Arthritis region affected - spine
Reason Only available in version 3.0 of the UDS.

Name ARTUNKN
Description Arthritis region affected - unknown
Reason Only available in version 3.0 of the UDS.

Name URINEINC
Description Incontinence present - urinary
Reason Only available in version 3.0 of the UDS.

Name BOWLINC
Description Incontinence present - bowel
Reason Only available in version 3.0 of the UDS.

Name SLEEPAP
Description Sleep apnea present
Reason Only available in version 3.0 of the UDS.

Name REMDIS
Description REM sleep behaviour disorder (RBD) present
Reason Only available in version 3.0 of the UDS.

Name HYPOSOM
Description Hyposomnia/insomnia present
Reason Only available in version 3.0 of the UDS.

Name SLEEPOTH
Description Other sleep disorder present
Reason Only available in version 3.0 of the UDS.

Name SLEEPOTX
Description Other sleep disorder - specify
Reason Free-text variable and only available in version 3.0 of the UDS.

Name ANGIOCP
Description Carotid procedure - angioplasty, endarterectomy or stent within the past 12 months
Reason Only available in version 3.0 of the UDS.

Name ANGIOPCI
Description Percutaneous coronary intervention - angioplasty and/or stent within the past 12 months
Reason Only available in version 3.0 of the UDS.

Name PACEMAKE
Description Procedure - pacemaker and/or defibrillator within the past 12 months
Reason Only available in version 3.0 of the UDS.

Name HVALVE
Description Procedure - heart valve replacement or repair within the past 12 months

Reason Only available in version 3.0 of the UDS.

Name ANTIENC
Description Antibody-mediated encephalopathy within the past 12 months
Reason Only available in version 3.0 of the UDS.

Name ANTIENCX
Description Antibody-mediated encephalopathy - specify
Reason Free-text variable and only available in version 3.0 of the UDS.

Name OTHCOND
Description Other medical conditions or procedures within the past 12 months not listed above
Reason Only available in version 3.0 of the UDS.

Name OTHCONDX
Description Other medical conditions - specify
Reason Free-text variable and only available in version 3.0 of the UDS.

Milestones

Number of Variables	16
Number of Variables Used	0

Dropped Variables

Name NACCDIED
Description Subject is known to be deceased
Reason Information provided would not be available at an initial visit.

Name NACCMOD
Description Month of death
Reason Information provided would not be available at an initial visit.

Name NACCYOD
Description Year of death
Reason Information provided would not be available at an initial visit.

Name NACCAUTP
Description Neuropathology data from an autopsy is available
Reason Information provided would not be available at an initial visit.

Name NACCACTV
Description Follow-up status at the Alzheimer's Disease Center (ADC)
Reason Information provided would not be available at an initial visit.

Name NACCNOVS
Description No longer followed annually in person or by telephone
Reason Information provided would not be available at an initial visit.

Name NACCDSMO
Description Month of discontinuation from annual follow-up
Reason Information provided would not be available at an initial visit.

Name NACCDSDY
Description Day of discontinuation from annual follow-up
Reason Information provided would not be available at an initial visit.

Name NACCDSYR
Description Year of discontinuation from annual follow-up
Reason Information provided would not be available at an initial visit.

Name NACCNURP
Description Permanently moved to a nursing home
Reason Information provided would not be available at an initial visit.

Name NACCNRMO
Description Month permanently moved to a nursing home
Reason Information provided would not be available at an initial visit.

Name NACCNRDY
Description Day permanently moved to a nursing home
Reason Information provided would not be available at an initial visit.

Name NACCNRYR
Description Year permanently moved to a nursing home
Reason Information provided would not be available at an initial visit.

Name NACCFTD
Description One or more FTLD (frontotemporal lobar degeneration) Module visits completed
Reason Irrelevant as data from the FTLD Module not utilised.

Name NACCMDSS
Description Subject's status in the Minimal Data Set (MDS) and Uniform Data Set (UDS)
Reason Irrelevant as only data available at an initial visit considered.

Name NACCPAFF
Description Previously affiliated subject
Reason Irrelevant as only data available at an initial visit considered.

Appendix B

Diagnostic and Differential Variable Importances

Table B.1 provides the diagnostic and differential importances for the 260 variables utilised. The diagnostic importance is the importance of a variable for diagnosing dementia, according to the dementia classifier; and the differential importance is the importance of a variable for the differential diagnosis of dementia, according to the pairwise dementia subtype classifiers. The latter was calculated using all 10 pairwise subtype classifiers, as explained in section 4.2.3. The variables are ordered with regards to their diagnostic importance; and a description of each of them is given, based on those provided by the National Alzheimer's Coordinating Center (2017). The short descriptions used in figures 4.6 and 4.11, as well as table 4.5, are also provided for the relevant variables (in square brackets following the description).

The table shows that the variables found to be important for diagnosing dementia are different to those found to be important for the differential diagnosis of dementia. This is demonstrated by DECCLIN, which indicates whether the clinician believed there was a meaningful decline in one or more of a variety of domains, or there were motor/movement changes. It is in the top 20 important variables for diagnosing dementia, but is of almost no importance for differentiating between subtypes. Another example is HXSTROKE, which provides the subject's stroke history. It is the most important variable for the differential diagnosis of dementia, but is of very little importance for diagnosing dementia.

Variable	Description [Short Description]	Diagnostic Importance	Differential Importance
COGJUDG	Indicate whether the subject currently is meaningfully impaired, relative to previously attained abilities, in executive function - judgment, planning or problem-solving [Impaired in judgment, planning or problem-solving]	100.00	2.52
HOMEHOBB	Home and hobbies [CDR - Home and hobbies]	96.48	2.43
COMMUN	Community affairs [CDR - Community affairs]	60.14	2.05
ORIENT	Orientation [CDR - Orientation]	54.76	3.70
NACCCOGF	Indicate the predominant symptom that was first recognised as a decline in the subject's cognition [Predominant symptom for decline in cognition]	54.67	30.26
COGMODE	Mode of onset of cognitive symptoms [Mode of onset of cognitive symptoms]	54.10	19.59
BILLS	In the past four weeks, did the subject have difficulty or need help with: writing checks, paying bills or balancing a checkbook [Recent difficulty with bills]	53.59	4.45
FRSTCHG	Indicate the predominant domain that was first recognised as changed in the subject [Predominant domain first recognised as changed]	49.17	28.74
CDRSUM	Clinical Dementia Rating (CDR) sum of boxes [CDR sum of boxes]	49.07	3.13
JUDGMENT	Judgment and problem-solving [CDR - Judgment and problem-solving]	42.95	3.08
MEMORY	Memory [CDR - Memory]	39.22	5.41
DECAGE	Based on clinician's assessment, at what age did the cognitive decline begin? [Age cognitive decline began]	37.74	12.48
INDEPEND	Level of independence [Level of independence]	36.48	2.16
CDRGLOB	Global CDR [Global CDR]	34.36	2.51
SHOPPING	In the past four weeks, did the subject have difficulty or need help with: shopping alone for clothes, household necessities or groceries [Recent difficulty with shopping]	33.31	4.52

TRAVEL	In the past four weeks, did the subject have difficulty or need help with: travelling out of the neighbourhood, driving, or arranging to take public transportation [Recent difficulty with travel]	32.36	4.80
DECCLIN	Clinician believes there is a meaningful decline in memory, non-memory cognitive abilities, behaviour, ability to manage his/her affairs, or there are motor/movement changes [Clinician believes meaningful decline]	28.41	0.22
TAXES	In the past four weeks, did the subject have difficulty or need help with: assembling tax records, business affairs or other papers [Recent difficulty with taxes]	28.11	4.77
COGSTAT	Per clinician, based on the neuropsychological examination, the subject's cognitive status is deemed [Subject's cognitive status]	27.85	3.65
COGMEM	Indicate whether the subject currently is meaningfully impaired, relative to previously attained abilities, in memory [Meaningfully impaired in memory]	27.28	20.95
MMSEORDA	Orientation subscale score - Time [MMSE - Orientation subscale score - Time]	25.47	3.47
EVENTS	In the past four weeks, did the subject have difficulty or need help with: keeping track of current events [Recent difficulty keeping track of events]	24.01	3.84
MEALPREP	In the past four weeks, did the subject have difficulty or need help with: preparing a balanced meal [Recent difficulty preparing a meal]	21.71	4.53
PERSCARE	Personal care [CDR - Personal care]	20.05	1.88
DECIN	Does the co-participant report a decline in subject's memory (relative to previously attained abilities)? [Co-participant reports a decline in memory]	18.68	9.69
STOVE	In the past four weeks, did the subject have difficulty or need help with: heating water, making a cup of coffee, or turning off the stove [Recent difficulty with stove]	18.26	3.93

REMDATES	In the past four weeks, did the subject have difficulty or need help with: remembering appointments, family occasions, holidays or medications [Recent difficulty remembering dates and medications]	18.20	5.85
COURSE	Overall course of decline of cognitive/behavioural/motor syndrome [Overall course of decline]	18.07	14.92
COGLANG	Indicate whether the subject currently is meaningfully impaired, relative to previously attained abilities, in language [Meaningfully impaired in language]	17.75	7.79
NACCMMSE	Total MMSE score (using D-L-R-O-W) [Total MMSE score]	17.47	3.49
LOGIMEM	Total number of story units recalled from this current test administration [Number of story units recalled]	17.05	6.17
NACCBEHF	Indicate the predominant symptom that was first recognised as a decline in the subject's behaviour [Predominant symptom for decline in behaviour]	15.82	12.88
PAYATTN	In the past four weeks, did the subject have difficulty or need help with: paying attention to and understanding a TV programme, book or magazine [Recent difficulty with paying attention]	15.13	3.18
TRAILB	Trail Making Test Part B - Total number of seconds to complete [Trail Making Test Part B - Seconds]	14.99	5.59
MEMUNITS	Logical Memory IIA - Delayed - Total number of story units recalled [Number of story units recalled - Delayed]	13.52	7.44
GAMES	In the past four weeks, did the subject have difficulty or need help with: playing a game of skill such as bridge or chess, or working on a hobby [Recent difficulty playing games of skill]	12.08	4.53
ANIMALS	Animals - Total number of animals named in 60 seconds [Number of animals in 60 seconds]	10.92	3.98
MMSEORLO	Orientation subscale score - Place [MMSE - Orientation subscale score - Place]	10.77	3.25
TRAILB_PROB	Reason an answer was not provided for TRAILB [Reason Trail Making Test Part B not completed]	10.33	4.02

BEMODE	Mode of onset of behavioural symptoms [Mode of onset of behavioural symptoms]	9.67	11.40
BEAPATHY	Subject currently manifests meaningful change in behaviour - Apathy, withdrawal [Meaningful change in behaviour - Apathy, withdrawal]	9.48	2.48
DIGIFLEN	Digit span forward length [Digit span forward length]	9.08	4.03
VEG	Vegetables - Total number of vegetables named in 60 seconds [Number of vegetables in 60 seconds]	8.87	4.62
BOSTON	Boston Naming Test (30) - Total score [Boston Naming Test (30) - Total score]	8.50	5.78
COGVIS	Indicate whether the subject currently is meaningfully impaired, relative to previously attained abilities, in visuospatial function [Meaningfully impaired in visuospatial function]	7.32	7.13
HANDALTR	Alternating movement - right hand [Alternating movement - Right hand]	7.04	5.67
DIGIF	Digit span forward trials correct [Digit span forward trials correct]	6.71	4.31
NACCMOTF	Indicate the predominant symptom that was first recognised as a decline in the subject's motor function [Predominant symptom for decline in motor function]	6.55	14.83
MOMODE	Mode of onset of motor symptoms [Mode of onset of motor symptoms]	6.55	22.14
TRAILA	Trail Making Test Part A - Total number of seconds to complete [Trail Making Test Part A - Seconds]	6.31	5.52
BEIRRIT	Subject currently manifests meaningful change in behaviour - Irritability [Meaningful change in behaviour - Irritability]	5.78	2.04
COGATTN	Indicate whether the subject currently is meaningfully impaired, relative to previously attained abilities, in attention or concentration [Meaningfully impaired in attention or concentration]	5.52	3.18
APA_SEV	Apathy or indifference, and the severity, in the last month [Recent apathy or indifference and severity]	5.45	3.90
TAPSRT	Finger taps - right hand [Finger taps - Right hand]	5.26	7.77
RIGDUPRT	Rigidity - right upper extremity [Rigidity - Right upper extremity]	5.26	6.16

TRACTRHD	Action or postural tremor - right hand [Action or postural tremor - Right hand]	5.15	3.32
DIGIB	Digit span backward trials correct [Digit span backward trials correct]	5.07	4.38
WAIS	Wechsler Adult Intelligence Scale (Revised) (WAIS-R) Digit Symbol [WAIS-R Digit Symbol]	5.00	5.41
HANDMOVR	Hand movements - right hand [Hand movements - Right hand]	4.57	6.93
TRAILA_PROB	Reason an answer was not provided for TRAILA [Reason Trail Making Test Part A not completed]	4.55	3.46
BEPERCH	Subject currently manifests meaningful change in behaviour - Personality change [Meaningful change in behaviour - Personality change]	4.54	15.76
AGIT_SEV	Agitation or aggression, and the severity, in the last month	4.47	2.92
QUITSMOK	If the subject quit smoking, age at which he/she last smoked (i.e. quit) [Age subject quit smoking]	4.39	6.73
SPEECH	Speech [Speech]	4.35	8.07
DECSUB	Does the subject report a decline in memory (relative to previously attained abilities)?	4.25	4.37
MOFALLS	Indicate whether the subject currently has meaningful changes in motor function - Falls [Meaningful changes in motor function - Falls]	4.12	6.72
MOSLOW	Indicate whether the subject currently has meaningful changes in motor function - Slowness [Meaningful changes in motor function - Slowness]	4.00	15.37
PACKSPER	Average number of packs smoked per day	3.98	5.29
HEARWAID	If the subject usually wears a hearing aid(s), is the subject's hearing functionally normal with a hearing aid(s)?	3.94	4.88
INBIR_#MOS	Months from co-participant's month/year of birth to month/year of visit	3.84	6.22
MEMTIME	Logical Memory IIA - Delayed - Time elapsed since Logical Memory IA - Immediate [Time elapsed since immediate story unit recall]	3.63	7.11
RAND_BVAR	Synthetic binary variable generated by randomly permuting INSEX	3.55	4.15
NACCAMD	Total number of medications reported at each visit	3.55	4.43

RAND_CVAR	Synthetic categorical variable generated by randomly permuting TRAILB_PROB	3.51	5.47
BEDEL	Subject currently manifests meaningful change in behaviour - Psychosis - Abnormal, false or delusional beliefs	3.50	3.56
BPSYS	Subject blood pressure (sitting), systolic	3.50	4.13
HRATE	Subject resting heart rate (pulse)	3.46	4.27
NACCDAD	Indicator of father with cognitive impairment	3.44	4.46
INCALLS	If no, approximate frequency of telephone contact?	3.41	5.73
NACCGDS	Total GDS score	3.41	4.07
BEDEP	Subject currently manifests meaningful change in behaviour - Depressed mood	3.34	1.82
DIGIBLEN	Digit span backward length	3.32	4.08
WEIGHT	Subject's weight (lbs)	3.29	4.52
INVISITS	If no, approximate frequency of in-person visits?	3.27	5.67
NACCBMI	Body mass index (BMI)	3.23	3.69
HISPOR	Hispanic origins	3.22	4.78
BEDISIN	Subject currently manifests meaningful change in behaviour - Disinhibition [Meaningful change in behaviour - Disinhibition]	3.22	18.03
NACCSTYR_#YRS	Years from most recently reported year of stroke as of the initial visit to year of visit [Years from last stroke]	3.20	72.34
BPDIAS	Subject blood pressure (sitting), diastolic	3.12	3.89
NACCTIYR_#YRS	Years from most recently reported year of TIA as of the initial visit to year of visit [Years from last transient ischemic attack]	3.11	14.47
HOPELESS	Do you feel that your situation is hopeless?	3.08	3.38
HEIGHT	Subject's height (inches)	3.08	4.55
HALL_SEV	Hallucinations and their severity in the last month [Recent hallucinations and severity]	3.02	7.80
MEMPROB	Do you feel you have more problems with memory than most?	2.97	3.88
INRELTO	Co-participant's relationship to subject	2.94	5.80

MOT_SEV	Motor disturbance and the severity in the last month	2.93	5.25
WAIS_PROB	Reason an answer was not provided for WAIS	2.85	3.64
DEL_SEV	Delusions and their severity in the last month	2.85	3.97
NACCMOM	Indicator of mother with cognitive impairment	2.83	4.03
DISN_SEV	Disinhibition and the severity in the last month [Recent disinhibition and severity]	2.76	6.80
INRELY	Is there a question about the co-participant's reliability?	2.68	2.97
IRR_SEV	Irritability or lability, and the severity, in the last month	2.67	2.73
VISWCORR	If the subject usually wears corrective lenses, is the subject's vision functionally normal with corrective lenses?	2.65	4.02
BIRTH_#MOS	Months from subject's month/year of birth to month/year of visit [Months from subject's birth]	2.57	12.54
POSSTAB	Posture stability [Posture stability]	2.54	8.34
EDUC	Years of education	2.53	3.04
NACCNIHR	Derived National Institutes of Health (NIH) race definitions	2.53	5.92
ANX_SEV	Anxiety and the severity in the last month	2.52	2.87
SPIRITS	Are you in good spirits most of the time?	2.52	3.13
PDOTHRYR_#YRS	Years from year of parkinsonian disorder diagnosis to year of visit [Years from parkinsonian disorder diagnosis]	2.50	28.88
ELAT_SEV	Elation or euphoria, and the severity, in the last month	2.50	4.65
HAPPY	Do you feel happy most of the time?	2.47	2.63
INSEX	Co-participant's sex	2.45	4.21
CVOTHR	Other cardiovascular disease	2.44	2.54
DEPD_SEV	Depression or dysphoria, and the severity, in the last month	2.42	3.57
MMSELOC	Administration of the Mini-Mental State Examination (MMSE) was:	2.40	3.23
NACCAANX	Reported current use of an anxiolytic, sedative or hypnotic agent	2.40	4.24
MOGAIT	Indicate whether the subject currently has meaningful changes in motor function - Gait disorder [Meaningful changes in motor function - Gait disorder]	2.40	14.37

TAPSLF	Finger taps - left hand [Finger taps - Left hand]	2.39	6.70
NACCACEI	Reported current use of an angiotensin converting enzyme (ACE) inhibitor	2.39	2.59
NACCAGE	Subject's age at visit [Subject's age at visit]	2.38	14.78
ENERGY	Do you feel full of energy?	2.36	4.49
HACHIN	Hachinski ischemic score [Hachinski ischemic score]	2.36	32.32
NACCDBMD	Reported current use of a diabetes medication	2.35	2.80
BETTER	Do you think that most people are better off than you are?	2.34	4.25
HANDMOVL	Hand movements - left hand [Hand movements - Left hand]	2.34	7.03
VISION	Without corrective lenses, is the subject's vision functionally normal?	2.33	2.70
NACCFAM	Indicator of first-degree family member with cognitive impairment	2.31	3.15
NITE_SEV	Nighttime behaviours and their severity in the last month	2.29	3.67
ABRUPT	Abrupt onset (re: cognitive status) [Abrupt onset (cognitive status)]	2.28	20.45
RAND_VAR	Synthetic variable generated by randomly sampling from a Normal distribution	2.26	2.78
APP_SEV	Appetite and eating problems, and their severity, in the last month	2.25	3.50
STAYHOME	Do you prefer to stay at home, rather than going out and doing new things?	2.24	2.69
INLIVWTH	Does the co-participant live with the subject?	2.23	5.17
VISCORR	Does the subject usually wear corrective lenses?	2.23	3.73
NACCBETA	Reported current use of a beta-adrenergic blocking agent (Beta-Blocker)	2.23	5.01
NACCEMD	Reported current use of estrogen hormone therapy	2.20	3.56
SMOKYRS	Total years smoked cigarettes	2.20	3.44
NPSYCLOC	The remainder of the battery was administered:	2.20	3.09
HEARING	Without a hearing aid(s), is the subject's hearing functionally normal?	2.20	3.83
NACCDIUR	Reported current use of a diuretic	2.19	4.44
HEARAID	Does the subject usually wear a hearing aid(s)?	2.18	4.87
NACCCCBS	Reported current use of a calcium channel blocking agent	2.18	3.62
NACCREFR	Principal referral source	2.15	3.06

RAND_DOCVAR	Synthetic ordinal/continuous variable generated by randomly permuting CDRSUM	2.15	2.62
FACEXP	Facial expression [Facial expression]	2.15	12.38
NACCAAAS	Reported current use of an antiadrenergic agent	2.13	3.21
PDYR_#YRS	Years from year of PD diagnosis to year of visit [Years from Parkinson's disease diagnosis]	2.13	31.46
NPIQINF	Neuropsychiatric Inventory Questionnaire (NPI-Q) co-participant	2.12	4.17
NACCREAS	Primary reason for coming to an Alzheimer's Disease Center (ADC)	2.12	1.94
B12DEF	Vitamin B12 deficiency	2.12	3.06
HANDED	Is the subject left- or right-handed?	2.11	3.02
AFRAID	Are you afraid that something bad is going to happen to you?	2.10	4.32
NACCANGI	Reported current use of an angiotensin II inhibitor	2.09	2.70
NACCHTNC	Reported current use of an antihypertensive combination therapy	2.07	2.74
HELPLESS	Do you often feel helpless?	2.05	3.75
NACCAHTN	Reported current use of any type of antihypertensive or blood pressure medication	2.05	3.65
NACCAC	Reported current use of an anticoagulant or antiplatelet agent	2.05	4.12
WONDRFUL	Do you think it is wonderful to be alive now?	2.02	2.44
LEGLF	Leg agility - left leg [Leg agility - Left leg]	2.02	7.73
NACCNSD	Reported current use of nonsteroidal anti-inflammatory medication	2.01	2.23
NACCADEP	Reported current use of an antidepressant	2.01	3.07
TRAUMBRF	Brain trauma - brief unconsciousness	2.01	3.36
WRTHLESS	Do you feel pretty worthless the way you are now?	2.01	3.27
HANDALTL	Alternating movement - left hand [Alternating movement - Left hand]	2.00	6.68
PRIMLANG	Primary language	2.00	3.50
HYPERCHO	Hypercholesterolemia	1.97	2.43
SATIS	Are you basically satisfied with your life?	1.97	2.59
BEOTHR	Subject currently manifests meaningful change in behaviour - Other	1.96	1.54
GAIT	Gait [Gait]	1.94	10.64

RIGDUPLF	Rigidity - left upper extremity	1.93	5.44
FOCLSYM	Focal neurological symptoms [Focal neurological symptoms]	1.93	15.12
PSYCDIS	Other psychiatric disorder	1.93	3.59
ARISING	Arising from chair [Arising from chair]	1.93	8.49
DROPACT	Have you dropped many of your activities and interests?	1.92	2.99
THYROID	Thyroid disease	1.92	2.20
DEPOTHR	Depression episodes more than two years ago	1.91	2.27
CVANGIO	Angioplasty/endarterectomy/stent	1.90	3.86
EMPTY	Do you feel that your life is empty?	1.90	2.76
NACCLIVS	Living situation	1.89	3.72
MARISTAT	Marital status	1.88	4.87
NACCLIPL	Reported current use of lipid lowering medication	1.88	2.65
BORED	Do you often get bored?	1.88	2.88
LEGRT	Leg agility - right leg [Leg agility - Right leg]	1.88	7.44
CVBYPASS	Cardiac bypass procedure	1.88	2.94
HYPERTEN	Hypertension	1.87	3.59
MEMUNITS_PROB	Reason an answer was not provided for MEMUNITS	1.87	2.97
FOCLSIGN	Focal neurological signs [Focal neurological signs]	1.87	17.73
NCOTHR	Other neurological condition	1.87	2.26
RESIDENC	Type of residence	1.85	2.84
NACCVASD	Reported current use of a vasodilator	1.85	4.29
HXHYPER	History or presence of hypertension	1.82	5.11
POSTURE	Posture [Posture]	1.79	6.94
NACCPDMD	Reported current use of an antiparkinson agent [Current use of an anti-parkinson agent]	1.77	15.28
PDOTHR	Other parkinsonian disorder [Other parkinsonian disorder]	1.76	24.50
ALCOHOL	Alcohol abuse - clinically significant impairment occurring over a 12-month period manifested in one of the following areas: work, driving, legal or social	1.75	3.86

TRAUMEXT	Brain trauma - extended unconsciousness	1.75	3.92
EYEMOVE	Are there eye movement abnormalities present indicative of central nervous system disorder? [Eye movement abnormalities indicative of CNS disorder]	1.75	18.21
RIGDNECK	Rigidity - neck	1.74	6.58
TRACTLHD	Action or postural tremor - left hand	1.73	4.12
NACCNREX	Were all findings unremarkable? [All physical/neurological exam findings unremarkable]	1.71	9.38
STEPWISE	Stepwise deterioration (re: cognitive status) [Stepwise deterioration (cognitive status)]	1.70	17.05
SOMATIC	Somatic complaints	1.66	3.36
FOCLDEF	Are focal deficits present indicative of central nervous system disorder? [Focal deficits indicative of CNS disorder]	1.65	11.82
CVCHF	Congestive heart failure	1.63	3.52
PDNORMAL	Unified Parkinson's Disease Rating Scale (UPDRS) normal	1.63	5.28
INCONTF	Incontinence - bowel	1.63	2.43
HISPANIC	Hispanic/Latino ethnicity	1.61	2.23
NACCTBI	History of traumatic brain injury (TBI)	1.61	2.29
GAITDIS	Is gait disorder present indicative of central nervous system disorder? [Gait disorder indicative of CNS disorder]	1.60	20.58
CVAFIB	Atrial fibrillation	1.60	6.53
DIABETES	Diabetes	1.59	2.02
BRADYKIN	Body bradykinesia and hypokinesia [Body bradykinesia and hypokinesia]	1.55	6.88
NPSYLAN	Language of test administration	1.53	2.54
LOGIMEM_PROB	Reason an answer was not provided for LOGIMEM	1.52	2.61
TOBAC30	Smoked cigarettes in last 30 days	1.51	2.12
MMSELAN	Language of MMSE administration	1.50	2.71
FORMVER	Form version number	1.50	2.35
ANYMEDS	Subject taking any medications	1.49	2.73
EMOT	Emotional incontinence	1.49	3.43

HXSTROKE	History of stroke [History of stroke]	1.48	100.00
INCONTU	Incontinence - urinary	1.47	1.75
TRESTRHD	Tremor at rest - right hand	1.46	5.48
BEAGIT	Subject currently manifests meaningful change in behaviour - Agitation	1.46	2.64
BOSTON_PROB	Reason an answer was not provided for BOSTON	1.45	3.28
TRESTFAC	Tremor at rest - face, lips, chin	1.45	2.46
VEG_PROB	Reason an answer was not provided for VEG	1.44	3.74
CBSTROKE	Stroke [Stroke]	1.43	64.00
CVPACE	Pacemaker	1.41	1.82
CVHATT	Heart attack/cardiac arrest	1.40	3.50
RIGDLOLF	Rigidity - left lower extremity	1.39	4.43
CBTIA	Transient ischemic attack (TIA) [Transient ischemic attack]	1.38	9.96
SEIZURES	Seizures	1.37	3.53
RIGDLORT	Rigidity - right lower extremity	1.37	4.62
SEX	Subject's sex	1.35	3.81
TOBAC100	Smoked more than 100 cigarettes in life	1.34	1.51
MOTREM	Indicate whether the subject currently has meaningful changes in motor function - Tremor [Meaningful changes in motor function - Tremor]	1.29	6.68
NACCAPSY	Reported current use of an antipsychotic agent	1.29	5.63
DIGIB_PROB	Reason an answer was not provided for DIGIB	1.26	2.20
DEP2YRS	Active depression in the last two years	1.24	1.88
DIGIFLEN_PROB	Reason an answer was not provided for DIGIFLEN	1.19	1.95
TRESTLHD	Tremor at rest - left hand	1.18	5.70
PD	Parkinson's disease (PD) [Parkinson's disease]	1.18	32.22
ANIMALS_PROB	Reason an answer was not provided for ANIMALS	1.17	2.86
BEVHALL	Subject currently manifests meaningful change in behaviour - Psychosis - Visual hallucinations [Meaningful change in behaviour - Visual hallucinations]	1.13	35.44

NOGDS	Is the subject able to complete the Geriatric Depression Scale (GDS), based on the clinician's best judgment?	1.08	2.21
TRAUMCHR	Brain trauma - chronic deficit	1.06	1.51
DIGIF_PROB	Reason an answer was not provided for DIGIF	0.94	2.42
MMSEORDA_PROB	Reason an answer was not provided for MMSEORDA	0.91	1.74
COGOTHR	Indicate whether the subject currently is meaningfully impaired, relative to previously attained abilities, in other cognitive domains	0.89	2.07
NACCMMSE_PROB	Reason an answer was not provided for NACCMMSE	0.82	2.28
ABUSOTHR	Other abused substances - clinically significant impairment occurring over a 12-month period manifested in one of the following areas: work, driving, legal or social	0.81	3.87
DIGIBLEN_PROB	Reason an answer was not provided for DIGIBLEN	0.80	2.21
TRESTLFT	Tremor at rest - left foot	0.75	2.09
BEAHALL	Subject currently manifests meaningful change in behaviour - Psychosis - Auditory hallucinations	0.66	6.39
TRESTRFT	Tremor at rest - right foot	0.59	4.20
NACCEPMD	Reported current use of estrogen + progestin hormone therapy	0.57	3.68
MMSEORLO_PROB	Reason an answer was not provided for MMSEORLO	0.46	1.89
NACCC1	Form date discrepancy between UDS Form A1 and Form C1	0.25	0.14
NACCVNUM	UDS visit number (order)	0.00	0.00

Table B.1: The 260 variables utilised from the NACC UDS, ordered with regards to their importance for diagnosing dementia (diagnostic importance) according to the dementia classifier. The diagnostic importance of each variable is provided, along with its importance for the differential diagnosis of dementia (differential importance) according to the pairwise dementia subtype classifiers.

References

H. Abdi (2007). 'Metric Multidimensional Scaling (MDS): Analyzing Distance Matrices'. In: *Encyclopedia of Measurement and Statistics*. Ed. by N. Salkind. Thousand Oaks, California: SAGE Publications, Inc. (cited on page 158).

ADC Clinical Task Force and National Alzheimer's Coordinating Center (2006a). *NACC Uniform Data Set (UDS): Initial Visit Packet Coding Guidebook*. Version 1.2 (cited on page 175).

ADC Clinical Task Force and National Alzheimer's Coordinating Center (2006b). *NACC Uniform Data Set (UDS): Initial Visit Packet Data Element Dictionary*. Version 1.2 (cited on page 175).

ADC Clinical Task Force and National Alzheimer's Coordinating Center (2006c). *NACC Uniform Data Set (UDS): Initial Visit Packet Forms*. Version 1.2 (cited on page 175).

ADC Clinical Task Force and National Alzheimer's Coordinating Center (2014a). *NACC Uniform Data Set (UDS): Initial Visit Packet Coding Guidebook*. Version 2.0 (cited on pages 21, 73, 110 and 175).

ADC Clinical Task Force and National Alzheimer's Coordinating Center (2014b). *NACC Uniform Data Set (UDS): Initial Visit Packet Data Element Dictionary*. Version 2.0 (cited on page 175).

ADC Clinical Task Force and National Alzheimer's Coordinating Center (2014c). *NACC Uniform Data Set (UDS): Initial Visit Packet Forms*. Version 2.0 (cited on page 175).

ADC Clinical Task Force and National Alzheimer's Coordinating Center (2017a). *NACC Uniform Data Set (UDS): Initial Visit Packet Coding Guidebook*. Version 3.0 (cited on pages 21 and 175).

ADC Clinical Task Force and National Alzheimer's Coordinating Center (2017b). *NACC Uniform Data Set (UDS): Initial Visit Packet Data Element Dictionary*. Version 3.0 (cited on page 175).

ADC Clinical Task Force and National Alzheimer's Coordinating Center (2017c). *NACC Uniform Data Set (UDS): Initial Visit Packet Forms*. Version 3.0 (cited on pages 4, 19, 23, 35, 69 and 175).

ADC Clinical Task Force and National Alzheimer's Coordinating Center (2019). *NACC Uniform Data Set (UDS): Initial Visit Packet Coding Guidebook*. Version 3.0 (cited on page 19).

E. Ahlqvist, P. Storm, A. Kőrájămăki, M. Martinell, M. Dorkhan, A. Carlsson, P. Vikman, R. Prasad, D. Aly, P. Almgren et al. (2018). 'Novel subgroups of adult-onset diabetes and their association with outcomes: a data-driven cluster analysis of six variables'. In: *The Lancet - Diabetes & Endocrinology* 6.5, pp. 361–369 (cited on page 143).

A. Ahmad and S. Khan (2019). 'Survey of State-of-the-Art Mixed Data Clustering Algorithms'. In: *IEEE Access* 7, pp. 31883–31902 (cited on pages 140 and 141).

M. Ahmed, Y. Zhang, Z. Feng, B. Lo, O. Inan and H. Liao (2019). 'Neuroimaging and Machine Learning for Dementia Diagnosis: Recent Advancements and Future Prospects'. In: *IEEE Reviews in Biomedical Engineering* 12, pp. 19–33 (cited on page 130).

K. Aho, P. Harmsen, S. Hatano, J. Marquardsen, V. Smirnov and T. Strasser (1980). 'Cerebrovascular disease in the community: results of a WHO Collaborative Study'. In: *Bulletin of the World Health Organisation* 58.1, pp. 113–130 (cited on page 40).

M. Alamuri, B. Surampudi and A. Negi (2014). 'A survey of distance/similarity measures for categorical data'. In: *2014 International Joint Conference on Neural Networks (IJCNN)*, pp. 1907–1914 (cited on page 142).

American Psychiatric Association (1994). *DSM-IV: Diagnostic and Statistical Manual of Mental Disorders*. Fourth Edition. Washington, DC: American Psychiatric Association (cited on page 39).

R. Andridge and R. Little (2010). 'A Review of Hot Deck Imputation for Survey Non-response'. In: *International Statistical Review* 78.1, pp. 40–64 (cited on page 54).

P. Andritsos, P. Tsaparas, R. Miller and K. Sevcik (2004). 'LIMBO: Scalable Clustering of Categorical Data'. In: *Advances in Database Technology - EDBT 2004*, pp. 123–146 (cited on page 141).

J. Atkins, E. Boman and B. Hendrickson (1998). 'A Spectral Algorithm for Seriation and the Consecutive Ones Problem'. In: *Society for Industrial and Applied Mathematics (SIAM) Journal on Computing* 28.1, pp. 297–310 (cited on page 100).

R. Baire (1909). 'Sur la représentation des fonctions discontinues'. In: *Acta Mathematica* 32.1, pp. 97–176 (cited on page 146).

D. Barbará, Y. Li and J. Couto (2002). 'COOLCAT: an entropy-based algorithm for categorical clustering'. In: *Proceedings of the eleventh international conference on Information and knowledge management*, pp. 582–589 (cited on page 141).

J. Bartlett, S. Seaman, I. White, J. Carpenter and Alzheimer's Disease Neuroimaging Initiative (2015). 'Multiple imputation of covariates by fully conditional specification: Accommodating the substantive model'. In: *Statistical Methods in Medical Research* 24.4, pp. 462–487 (cited on pages 56 and 59).

R. Beard and T. Neary (2013). 'Making sense of nonsense: experiences of mild cognitive impairment'. In: *Sociology of Health & Illness* 35.1, pp. 130–146 (cited on page 103).

D. Beekly, E. Ramos, W. Lee, W. Deitrich, M. Jacka, J. Wu, J. Hubbard, T. Koepsell, J. Morris, W. Kukull et al. (2007). 'The National Alzheimer's Coordinating Center (NACC) Database: The Uniform Data Set'. In: *Alzheimer Disease & Associated Disorders* 21.3, pp. 249–258 (cited on page 16).

J. Bezdek (1981). *Pattern Recognition with Fuzzy Objective Function Algorithms*. New York, NY: Plenum Press (cited on page 165).

S. Boriah, V. Chandola and V. Kumar (2008). 'Similarity Measures for Categorical Data: A Comparative Evaluation'. In: *Proceedings of the 2008 SIAM International Conference on Data Mining*, pp. 243–254 (cited on page 142).

L. Breiman (1996). 'Bagging Predictors'. In: *Machine Learning* 24.2, pp. 123–140 (cited on page 8).

L. Breiman (2001). 'Random Forests'. In: *Machine Learning* 45, pp. 5–32 (cited on pages 5, 8 and 9).

L. Breiman and A. Cutler (2004). *Random Forests*. Online. Accessed: February 2020. URL: `https://www.stat.berkeley.edu/~breiman/RandomForests/cc_home.htm` (cited on pages 58, 59, 66, 69, 71, 72, 79, 84, 85, 86 and 172).

L. Breiman, J. Friedman, R. Olshen and C. Stone (1984). *Classification And Regression Trees*. Boca Raton, FL: Chapman & Hall (cited on page 4).

E. Cesario, G. Manco and R. Ortale (2007). 'Top-Down Parameter-Free Clustering of High-Dimensional Categorical Data'. In: *IEEE Transactions on Knowledge and Data Engineering* 19.12, pp. 1607–1624 (cited on page 141).

A. Chaturvedi, P. Green and J. Carroll (2001). 'K-modes Clustering'. In: *Journal of Classification* 18.1, pp. 35–55 (cited on page 140).

P.-Y. Chiu, H. Tang, C.-Y. Wei, C. Zhang, G.-U. Hung and W. Zhou (2019). 'NMD-12: A new machine-learning derived screening instrument to detect mild cognitive impairment and dementia'. In: *PLoS ONE* 14.3 (cited on page 133).

F. Chung (1997). *Spectral Graph Theory*. Conference Board of the Mathematical Sciences: Regional Conference Series in Mathematics 92. Providence, Rhode Island: American Mathematical Society (cited on page 149).

L. Cleret de Langavant, E. Bayen and K. Yaffe (2018). 'Unsupervised Machine Learning to Identify High Likelihood of Dementia in Population-Based Surveys: Development and Validation Study'. In: *Journal of Medical Internet Research* 20.7, e10493 (cited on page 143).

P. Contreras and F. Murtagh (2010). 'Fast Hierarchical Clustering from the Baire Distance'. In: *Classification as a Tool for Research*. Ed. by H. Locarek-Junge and C. Weihs. Berlin, Heidelberg: Springer-Verlag. Chap. 25, pp. 235–243 (cited on page 148).

D. Cortes (Nov. 2019). 'Distance approximation using Isolation Forests'. In: *arXiv:1910.12362v2 [stat.ML]* (cited on page 142).

A. Criminisi, J. Shotton and E. Konukoglu (2011). *Decision Forests for Classification, Regression, Density Estimation, Manifold Learning and Semi-Supervised Learning*. Tech. rep. Cambridge, UK: Microsoft Research Ltd (cited on page 5).

R. Croninger and K. Douglas (2005). 'Missing Data and Institutional Research'. In: *New Directions for Institutional Research* 2005.127, pp. 33–49 (cited on page 53).

A. Cutler, D. Cutler and J. Stevens (2012). 'Random Forests'. In: *Ensemble Machine Learning: Methods and Applications*. Ed. by C. Zhang and Y. Ma. Boston, MA: Springer. Chap. 5, pp. 157–175 (cited on pages 58, 71 and 85).

A. Dallora, S. Eivazzadeh, E. Mendes, J. Berglund and P. Anderberg (2017). 'Machine learning and microsimulation techniques on the prognosis of dementia: A systematic literature review'. In: *PLoS ONE* 12.6 (cited on page 133).

M. Dauwan, J. van der Zande, E. van Dellen, I. Sommer, P. Scheltens, A. Lemstra and C. Stam (2016). 'Random forest to differentiate dementia with Lewy bodies

from Alzheimer's disease'. In: *Alzheimer's & Dementia: Diagnosis, Assessment & Disease Monitoring* 4, pp. 99–106 (cited on pages 132 and 133).

J. de Bakker and E. de Vink (1998). 'Denotational models for programming languages: applications of Banach's Fixed Point Theorem'. In: *Topology and its Applications* 85.1–3, pp. 35–52 (cited on pages 146 and 148).

A. Dempster, N. Laird and D. Rubin (1977). 'Maximum Likelihood from Incomplete Data Via the EM Algorithm'. In: *Journal of the Royal Statistical Society: Series B (Methodological)* 39.1, pp. 1–22 (cited on page 55).

M. Desai, A. Mitani, S. Bryson and T. Robinson (2016). 'Multiple Imputation When Rate of Change is the Outcome of Interest'. In: *Journal of Modern Applied Statistical Methods* 15.1, pp. 160–192 (cited on page 56).

Y. Ding and J. Simonoff (2010). 'An Investigation of Missing Data Methods for Classification Trees Applied to Binary Response Data'. In: *Journal of Machine Learning Research* 11.Jan, pp. 131–170 (cited on page 57).

D. Dua and C. Graff (2019). *UCI Machine Learning Repository*. Online by University of California, Irvine, School of Information & Computer Sciences. Accessed: November 2020. URL: `https://archive.ics.uci.edu/ml/index.php` (cited on page 155).

Duke University (2002). *Old Faithful Geyser Eruption Data*. Online. Accessed: November 2020. URL: `https://www2.stat.duke.edu/courses/Fall02/sta290/datasets/geyser` (cited on page 153).

S. Elavarasi and J. Akilandeswari (2014). 'Survey on Clustering Algorithm and Similarity Measure for Categorical Data'. In: *ICTACT Journal on Soft Computing* 4.2, pp. 715–722 (cited on pages 141 and 142).

C. Enders (2010). *Applied Missing Data Analysis*. New York, NY: The Guilford Press (cited on page 54).

B. Everitt, S. Landau, M. Leese and D. Stahl (2011). *Cluster Analysis*. Fifth Edition. Wiley Series in Probability and Statistics 848. Chichester, UK: John Wiley & Sons, Ltd. (cited on page 165).

S. Fahn, R. Elton and UPDRS Development Committee (1987). 'The Unified Parkinson's Disease Rating Scale'. In: *Recent Developments in Parkinson's Disease*. Ed. by S. Fahn, C. Marsden, D. Calne and M. Goldstein. Vol. 2. Florham Park, NJ: Macmillan Healthcare Information (cited on page 18).

T. Fawcett (2006). 'An introduction to ROC analysis'. In: *Pattern Recognition Letters* 27.8, pp. 861–874 (cited on pages 88, 98 and 99).

M. Folstein, S. Folstein and P. McHugh (2001). *Mini-Mental State Examination*. PAR, Inc. Lutz, FL (cited on page 20).

E. Forgy (1965). 'Cluster Analysis of Multivariate Data: Efficiency vs. Interpretability of Classifications'. In: *Biometrics*. Vol. 21. 3, pp. 768–769 (cited on page 140).

A. Foss, M. Markatou and B. Ray (2019). 'Distance Metrics and Clustering Methods for Mixed-type Data'. In: *International Statistical Review* 87.1, pp. 80–109 (cited on pages 141 and 142).

D. Gamberger, N. Lavrač, S. Srivatsa, R. Tanzi and P. Doraiswamy (2017). 'Identification of clusters of rapid and slow decliners among subjects at risk for Alzheimer's disease'. In: *Scientific Reports* 7.6763 (cited on page 143).

V. Ganti, J. Gehrke and R. Ramakrishnan (1999). 'CACTUS - Clustering Categorical Data Using Summaries'. In: *Proceedings of the fifth ACM SIGKDD international conference on Knowledge discovery and data mining*, pp. 73–83 (cited on page 141).

P. Geurts, D. Ernst and L. Wehenkel (2006). 'Extremely randomized trees'. In: *Machine Learning* 63.1, pp. 3–42 (cited on pages 5, 9, 11, 12, 69, 85, 87 and 172).

J. Gower (1971). 'A general coefficient of similarity and some of its properties'. In: *Biometrics* 27.4, pp. 857–871 (cited on page 142).

S. Guha, R. Rastogi and K. Shim (2000). 'Rock: A robust clustering algorithm for categorical attributes'. In: *Information Systems* 25.5, pp. 345–366 (cited on page 141).

D. Hand and R. Till (2001). 'A Simple Generalisation of the Area Under the ROC Curve for Multiple Class Classification Problems'. In: *Machine Learning* 45.2, pp. 171–186 (cited on page 88).

D. Hand and K. Yu (2001). 'Idiot's Bayes - Not So Stupid After All?' In: *International Statistical Review* 69.3, pp. 385–398 (cited on page 46).

S. Hariri, M. Kind and R. Brunner (Nov. 2018). 'Extended Isolation Forest'. In: *arXiv:1811.02141v1 [cs.LG]* (cited on pages 145 and 174).

Z. He, X. Xu and S. Deng (2002). 'Squeezer: An Efficient Algorithm for Clustering Categorical Data'. In: *Journal of Computer Science and Technology* 17.5, pp. 611–624 (cited on pages 141 and 142).

J. Hendrickson (2014). 'Methods for Clustering Mixed Data'. PhD thesis. University of South Carolina (cited on page 141).

T. Ho (1995). 'Random Decision Forests'. In: *Proceedings of 3rd International Conference on Document Analysis and Recognition*, pp. 278–282 (cited on page 5).

Z. Huang (1998). 'Extensions to the k-Means Algorithm for Clustering Large Data Sets with Categorical Values'. In: *Data Mining and Knowledge Discovery* 2.3, pp. 283–304 (cited on pages 140 and 141).

Z. Huang and M. Ng (2003). 'A Note on K-modes Clustering'. In: *Journal of Classification* 20.2, pp. 257–261 (cited on page 141).

L. Hubert and P. Arabie (1985). 'Comparing Partitions'. In: *Journal of Classification* 2.1, pp. 193–218 (cited on page 164).

G. James, D. Witten, T. Hastie and R. Tibshirani (2017). *An Introduction to Statistical Learning with Applications in R*. Eighth Edition. New York, NY: Springer (cited on pages 5, 8, 12 and 140).

E. Jammeh, C. Carroll, S. Pearson, J. Escudero, A. Anastasiou, P. Zhao, T. Chenore, J. Zajicek and E. Ifeachor (2018). 'Machine-learning based identification of undiagnosed dementia in primary care: a feasibility study'. In: *British Journal of General Practice (BJGP) Open* 2.2 (cited on page 132).

W. Jarrold, B. Peintner, D. Wilkins, D. Vergryi, C. Richey, M. Gorno-Tempini and J. Ogar (2014). 'Aided Diagnosis of Dementia Type through Computer-Based Analysis of Spontaneous Speech'. In: *Workshop on Computational Linguistics and Clinical Psychology*, pp. 27–37 (cited on page 132).

H. Jia, Y. Cheung and J. Liu (2016). 'A New Distance Metric for Unsupervised Learning of Categorical Data'. In: *IEEE Transactions on Neural Networks and Learning Systems* 27.5, pp. 1065–1079 (cited on page 142).

J. Josse, N. Prost, E. Scornet and G. Varoquaux (Mar. 2019). 'On the consistency of supervised learning with missing values'. In: *arXiv:1902.06931v2 [stat.ML]* (cited on pages 58 and 60).

A. Kapelner and J. Bleich (2015). 'Prediction with missing data via Bayesian Additive Regression Trees'. In: *The Canadian Journal of Statistics* 43.2, pp. 224–239 (cited on page 60).

D. Kaufer, J. Cummings, P. Ketchel, V. Smith, A. MacMillan, T. Shelley, O. Lopez and S. DeKosky (2000). 'Validation of the NPI-Q, a brief clinical form of the Neuropsychiatric Inventory'. In: *The Journal of Neuropsychiatry and Clinical Neurosciences* 12.2, pp. 233–239 (cited on page 19).

E. Keogh and A. Mueen (2017). 'Curse of Dimensionality'. In: *Encyclopedia of Machine Learning and Data Mining*. Ed. by C. Sammut and G. Webb. Boston, MA: Springer (cited on pages 5 and 141).

King's College London (2020). *About SSNAP*. Online. Accessed: October 2020. URL: `https://www.strokeaudit.org/About-SSNAP.aspx` (cited on page 52).

S. Kotsiantis (2013). 'Decision trees: a recent overview'. In: *Artificial Intelligence Review* 39.4, pp. 261–283 (cited on page 4).

K. Langa, B. Plassman, R. Wallace, A. Herzog, S. Heeringa, M. Ofstedal, J. Burke, G. Fisher, N. Fultz, M. Hurd et al. (2005). 'The Aging, Demographics, and Memory Study: Study Design and Methods'. In: *Neuroepidemiology* 25.4, pp. 181–191 (cited on page 143).

I. Liiv (2010). 'Seriation and matrix reordering methods: An historical overview'. In: *Statistical Analysis and Data Mining* 3.2, pp. 70–91 (cited on page 100).

M. Lin, P. Gong, T. Yang, J. Ye, R. Albin and H. Dodge (2018). 'Big Data Analytical Approaches to the NACC Dataset: Aiding Preclinical Trial Enrichment'. In: *Alzheimer Disease and Associated Disorders* 32.1, pp. 18–27 (cited on page 134).

R. Little and D. Rubin (2002). *Statistical Analysis with Missing Data*. Second Edition. Hoboken, New Jersey: John Wiley & Sons, Inc. (cited on pages 53, 54 and 55).

I. Litvan, K. Bhatia, D. Burn, C. Goetz, A. Lang, I. McKeith, N. Quinn, K. Sethi, C. Shults, G. Wenning et al. (2003). 'Movement Disorders Society Scientific Issues Committee Report: SIC Task Force appraisal of clinical diagnostic criteria for Parkinsonian disorders'. In: *Movement Disorders* 18.5, pp. 467–486 (cited on page 40).

F. Liu, K. Ting and Z.-H. Zhou (2008). 'Isolation Forest'. In: *Proceedings of 2008 Eighth IEEE International Conference on Data Mining*, pp. 413–422 (cited on pages 6, 144 and 145).

S. Lloyd (1982). 'Least squares quantization in PCM'. In: *IEEE Transactions on Information Theory* 28.2, pp. 129–137 (cited on page 140).

G. Louppe, L. Wehenkel, A. Sutera and P. Geurts (2013). 'Understanding variable importances in forests of randomized trees'. In: *Advances in Neural Information Processing Systems 26 (NIPS 2013)*, pp. 431–439 (cited on page 107).

J. Maroco, D. Silva, A. Rodrigues, M. Guerreiro, I. Santana and A. de Mendonça (2011). 'Data mining methods in the prediction of Dementia: A real-data comparison of the accuracy, sensitivity and specificity of linear discriminant analysis, logistic regression, neural networks, support vector machines, classification trees and random forests'. In: *BMC Research Notes* 4.299 (cited on page 133).

Mayo Clinic Staff (2019). *Dementia*. Online. Accessed: September 2020. URL: `https://www.mayoclinic.org/diseases-conditions/dementia/symptoms-causes/syc-20352013` (cited on page 104).

I. McKeith, D. Dickson, J. Lowe, M. Emre, J. O'Brien, H. Feldman, J. Cummings, J. Duda, C. Lippa, E. Perry et al. (2005). 'Diagnosis and management of dementia with Lewy bodies: Third report of the DLB consortium'. In: *Neurology* 65.12, pp. 1863–1872 (cited on pages 2 and 40).

G. McKhann, D. Drachman, M. Folstein, R. Katzman, D. Price and E. Stadlan (1984). 'Clinical diagnosis of Alzheimer's disease: Report of the NINCDS-ADRDA Work Group under the auspices of Department of Health and Human Services Task Force on Alzheimer's Disease'. In: *Neurology* 34.7, pp. 939–944 (cited on pages 2 and 40).

M.-M. Mesulam (2001). 'Primary progressive aphasia'. In: *Annals of Neurology* 49.4, pp. 425–432 (cited on page 40).

M.-M. Mesulam (2003). 'Primary Progressive Aphasia - A Language-Based Dementia'. In: *The New England Journal of Medicine* 349.16, pp. 1535–1542 (cited on page 40).

A. Mitelpunkt, T. Galili, T. Kozlovski, N. Bregman, N. Shachar, M. Markus-Kalish and Y. Benjamini (2020). 'Novel Alzheimer's disease subtypes identified using a data and knowledge driven strategy'. In: *Scientific Reports* 10.1327 (cited on page 143).

J. Morris, S. Weintraub, H. Chui, J. Cummings, C. DeCarli, S. Ferris, N. Foster, D. Galasko, N. Graff-Radford, E. Peskind et al. (2006). 'The Uniform Data Set (UDS): Clinical and Cognitive Variables and Descriptive Data From Alzheimer Disease

Centers'. In: *Alzheimer Disease & Associated Disorders* 20.4, pp. 210–216 (cited on page 16).

J. Morris (1993). 'The Clinical Dementia Rating (CDR): Current version and scoring rules'. In: *Neurology* 43.11, pp. 2412–2414 (cited on page 19).

S. Mueller, M. Weiner, L. Thal, R. Petersen, C. Jack, W. Jagust, J. Trojanowski and A. Toga (2005). 'The Alzheimer's Disease Neuroimaging Initiative'. In: *Neuroimaging Clinics* 15.4, pp. 869–877 (cited on page 143).

National Alzheimer's Coordinating Center (2017). *NACC Uniform Data Set (UDS): Researchers Data Dictionary*. Version 3.0 (cited on pages 16, 34, 37, 39, 84, 108, 124, 175, 176 and 363).

National Alzheimer's Coordinating Center (2019). *The NACC database*. Online. Accessed: December 2019. URL: `https://www.alz.washington.edu/WEB/about_about.html` (cited on pages 3 and 16).

National Alzheimer's Coordinating Center (2020). *The UDS study population*. Online. Accessed: March 2020. URL: `https://www.alz.washington.edu/WEB/study_pop.html` (cited on page 40).

National Institute on Aging (2017). *What is Mixed Dementia? Causes and Diagnosis*. Online. Accessed: September 2020. URL: `https://www.nia.nih.gov/health/what-mixed-dementia-causes-and-diagnosis` (cited on pages 1 and 135).

National Institute on Aging (2019). *Alzheimer's Disease Research Centers*. Online. Accessed: December 2019. URL: `https://www.nia.nih.gov/health/alzheimers-disease-research-centers` (cited on page 15).

D. Neary, J. Snowden, L. Gustafson, U. Passant, D. Stuss, S. Black, M. Freedman, A. Kertesz, P. Robert, M. Albert et al. (1998). 'Frontotemporal lobar degeneration: A consensus on clinical diagnostic criteria'. In: *Neurology* 51.6, pp. 1546–1554 (cited on pages 2 and 40).

NeurologyToday (2004). *Dementia Experts Debate Diagnosis of Mild Cognitive Impairment*. Online. Accessed: September 2020. URL: https://journals.lww.com/neurotodayonline/Fulltext/2004/07000/DEMENTIA_EXPERTS_DEBATE_DIAGNOSIS_OF_MILD.4.aspx (cited on page 103).

S. Pan and Q. Yang (2010). 'A Survey on Transfer Learning'. In: *IEEE Transactions on Knowledge and Data Engineering* 22.10, pp. 1345–1359 (cited on page 134).

F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg et al. (2011). 'Scikit-learn: Machine Learning in Python'. In: *Journal of Machine Learning Research* 12, pp. 2825–2830 (cited on pages 107 and 153).

E. Pellegrini, L. Ballerini, M. Valdes Hernandez, F. Chappell, V. González-Castro, D. Anblagan, S. Danso, S. Muñoz-Maniega, D. Job, C. Pernet et al. (2018). 'Machine learning of neuroimaging for assisted diagnosis of cognitive impairment and dementia: A systematic review'. In: *Alzheimer's & Dementia: Diagnosis, Assessment & Disease Monitoring* 10, pp. 519–535 (cited on pages 130 and 132).

R. Petersen and J. Morris (2005). 'Mild cognitive impairment as a clinical entity and treatment target'. In: *Archives of Neurology* 62.7, pp. 1160–1163 (cited on page 39).

R. Pfeffer, T. Kurosaki, C. Harrah, J. Chance and S. Filos (1982). 'Measurement of Functional Activities in Older Adults in the Community'. In: *Journal of Gerontology* 37.3, pp. 323–329 (cited on page 20).

C. Pinto and A. Subramanyam (2009). 'Mild cognitive impairment: The dilemma'. In: *Indian Journal of Psychiatry* 51.Suppl1, S44–S51 (cited on page 103).

M. Prince, G.-C. Ali, M. Guerchet, M. Prina, E. Albanese and Y.-T. Wu (2016). 'Recent global trends in the prevalence and incidence of dementia, and survival with dementia'. In: *Alzheimer's Research & Therapy* 8.23 (cited on page 1).

M. Prince, A. Wimo, M. Guerchet, G.-C. Ali, Y.-T. Wu, M. Prina and Alzheimer's Disease International (2015). *World Alzheimer Report 2015 - The Global Impact*

of Dementia: An Analysis of Prevalence, Incidence, Cost and Trends. Alzheimer's Disease International, London (cited on page 1).

Y. Qiu, D. Jacobs, K. Messer, D. Salmon and H. Feldman (2018). 'Identifying Cognitive Subtypes of Mild-to-Moderate Alzheimer's Disease Using the National Alzheimer's Coordinating Center Uniform Dataset Neuropsychological Test Battery'. In: *Neurology* 90.15 Supplement P5.180 (cited on page 143).

Y. Qiu, D. Jacobs, K. Messer, D. Salmon and H. Feldman (2019). 'Cognitive heterogeneity in probable Alzheimer disease: Clinical and neuropathologic features'. In: *Neurology* 93.8, e778–e790 (cited on page 143).

W. Rand (1971). 'Objective Criteria for the Evaluation of Clustering Methods'. In: *Journal of the American Statistical Association* 66.336, pp. 846–850 (cited on page 164).

L. Ritchie and H. Tuokko (2011). 'Clinical Decision Trees for Predicting Conversion from Cognitive Impairment No Dementia (CIND) to Dementia in a Longitudinal Population-Based Study'. In: *Archives of Clinical Neuropsychology* 26.1, pp. 16–25 (cited on page 133).

G. Román, T. Tatemichi, T. Erkinjuntti, J. Cummings, J. Masdeu, J. Garcia, L. Amaducci, J.-M. Orgogozo, A. Brun, A. Hofman et al. (1993). 'Vascular dementia - Diagnostic criteria for research studies: Report of the NINDS-AIREN International Workshop'. In: *Neurology* 43.2, pp. 250–260 (cited on pages 2 and 40).

W. Rosen, R. Terry, P. Fuld, R. Katzman and A. Peck (1980). 'Pathological Verification of Ischemic Score in Differentiation of Dementias'. In: *Annals of Neurology* 7.5, pp. 486–488 (cited on page 18).

A. Sarica, A. Cerasa and A. Quattrone (2017). 'Random Forest Algorithm for the Classification of Neuroimaging Data in Alzheimer's Disease: A Systematic Review'. In: *Frontiers in Aging Neuroscience* 9.329 (cited on pages 130 and 133).

A. Saxena, M. Prasad, A. Gupta, N. Bharill, O. Patel, A. Tiwari, M. Er, W. Ding and C.-T. Lin (2017). 'A review of clustering techniques and developments'. In: *Neurocomputing* 267, pp. 664–681 (cited on page 139).

J. Schafer (1997). *Analysis of Incomplete Multivariate Data*. First Edition. London, UK: Chapman & Hall (cited on page 55).

J. Schafer and J. Graham (2002). 'Missing Data: Our View of the State of the Art'. In: *Psychological Methods* 7.2, pp. 147–177 (cited on pages 54 and 55).

scikit-learn developers (2020a). *Clustering: Mutual Information based scores*. Online. Accessed: December 2020. URL: `https://scikit-learn.org/stable/modules/clustering.html#mutual-info-score` (cited on page 158).

scikit-learn developers (2020b). *Clustering: Rand index*. Online. Accessed: May 2021. URL: `https://scikit-learn.org/stable/modules/clustering.html#adjusted-rand-score` (cited on page 164).

scikit-learn developers (2020c). *Naive Bayes*. Online. Accessed: April 2021. URL: `https://scikit-learn.org/stable/modules/naive_bayes.html` (cited on page 46).

J. Sheikh and J. Yesavage (1986). 'Geriatric Depression Scale (GDS): Recent Evidence and Development of a Shorter Version'. In: *Clinical Gerontologist* 5.1–2, pp. 165–173 (cited on page 19).

J. Shi and J. Malik (2000). 'Normalized cuts and image segmentation'. In: *IEEE Transactions on Pattern Analysis and Machine Intelligence* 22.8, pp. 888–905 (cited on page 150).

T. Shi and S. Horvath (2006). 'Unsupervised Learning With Random Forest Predictors'. In: *Journal of Computational and Graphical Statistics* 15.1, pp. 118–138 (cited on page 142).

R. Sokal and C. Michener (1958). 'A Statistical Method for Evaluating Systematic Relationships'. In: *The University of Kansas Science Bulletin* 38, pp. 1409–1438 (cited on page 165).

L. Sørensen, M. Nielsen and Alzheimer's Disease Neuroimaging Initiative (2018). 'Ensemble support vector machine classification of dementia using structural MRI and mini-mental state examination'. In: *Journal of Neuroscience Methods* 302, pp. 66–74 (cited on page 130).

D. Stekhoven and P. Bühlmann (2012). 'MissForest - non-parametric missing value imputation for mixed-type data'. In: *Bioinformatics* 28.1, pp. 112–118 (cited on pages 57, 58 and 85).

A. Stemmer, T. Galili, T. Kozlovski, Y. Zeevi, M. Marcus-Kalish, Y. Benjamini and A. Mitelpunkt (2019). 'Current and Potential Approaches for Defining Disease Signatures: a Systematic Review'. In: *Journal of Molecular Neuroscience* 67.4, pp. 550–558 (cited on pages 2 and 139).

A. Strehl and J. Ghosh (2002). 'Cluster Ensembles - A Knowledge Reuse Framework for Combining Multiple Partitions'. In: *Journal of Machine Learning Research* 3.Dec, pp. 583–617 (cited on page 158).

F. Tang and H. Ishwaran (2017). 'Random Forest Missing Data Algorithms'. In: *Statistical Analysis and Data Mining* 10.6, pp. 363–377 (cited on pages 57 and 58).

M. Tanner and W. Wong (1987). 'The Calculation of Posterior Distributions by Data Augmentation'. In: *Journal of the American Statistical Association* 82.398, pp. 528–540 (cited on page 55).

The Genetic Frontotemporal dementia Initiative (2020). *GENFI: Genetic FTD Initiative*. Online. Accessed: October 2020. URL: `https://www.genfi.org/` (cited on page 143).

G. Tsoumakas and I. Katakis (2007). 'Multi-Label Classification: An Overview'. In: *International Journal of Data Warehousing and Mining* 3.3, pp. 1–13 (cited on page 135).

B. Twala (2005). 'Effective techniques for handling incomplete data using decision trees'. PhD thesis. The Open University (cited on pages 54 and 57).

B. Twala (2009). 'An Empirical Comparison of Techniques for Handling Incomplete Data using Decision Trees'. In: *Applied Artificial Intelligence* 23.5, pp. 373–405 (cited on page 57).

B. Twala, M. Jones and D. Hand (2008). 'Good methods for coping with missing data in decision trees'. In: *Pattern Recognition Letters* 29.7, pp. 950–956 (cited on pages 5, 59, 60, 64 and 172).

S. van Buuren (2018). *Flexible Imputation of Missing Data*. Second Edition. Boca Raton, FL: Chapman & Hall/CRC (cited on pages 55, 56, 57 and 58).

M. van de Velden, A. D'Enza and A. Markos (2019). 'Distance-based clustering of mixed data'. In: *WIREs Computational Statistics* 11.3, e1456 (cited on page 141).

W. van der Flier, Y. Pijnenburg, N. Prins, A. Lemstra, F. Bouwman, C. Teunissen, B. van Berckel, C. Stam, F. Barkhof, P. Visser et al. (2014). 'Optimizing Patient Care and Research: The Amsterdam Dementia Cohort'. In: *Journal of Alzheimer's Disease* 41.1, pp. 313–327 (cited on page 132).

C. Viroli (2012). 'Using factor mixture analysis to model heterogeneity, cognitive structure, and determinants of dementia: an application to the Aging, Demographics, and Memory Study'. In: *Statistics in Medicine* 31.19, pp. 2110–2122 (cited on page 143).

U. von Luxburg (2007). 'A tutorial on spectral clustering'. In: *Statistics and Computing* 17.4, pp. 395–416 (cited on pages 144, 149 and 150).

A. Weakley, J. Williams, M. Schmitter-Edgecombe and D. Cook (2015). 'Neuropsychological test selection for cognitive impairment classification: A machine learning approach'. In: *Journal of Clinical and Experimental Neuropsychology* 37.9, pp. 899–916 (cited on page 133).

L. Wehenkel and M. Pavella (1991). 'Decision Trees and Transient Stability of Electric Power Systems'. In: *Automatica* 27.1, pp. 115–134 (cited on page 11).

S. Weintraub, D. Salmon, N. Mercaldo, S. Ferris, N. Graff-Radford, H. Chui, J. Cummings, C. DeCarli, N. Foster, D. Galasko et al. (2009). 'The Alzheimer's Disease Centers' Uniform Data Set (UDS): The Neuropsychologic Test Battery'. In: *Alzheimer Disease & Associated Disorders* 23.2, pp. 91–101 (cited on page 16).

J. Whitwell, S. Przybelski, S. Weigand, R. Ivnik, P. Vemuri, J. Gunter, M. Senjem, M. Shiung, B. Boeve, D. Knopman et al. (2009). 'Distinct anatomical subtypes of the behavioural variant of frontotemporal dementia: a cluster analysis study'. In: *Brain* 132.11, pp. 2932–2946 (cited on page 143).

World Health Organisation (2020). *Dementia*. Online. Accessed: December 2020. URL: `https://www.who.int/news-room/fact-sheets/detail/dementia` (cited on page 1).

Y.-T. Wu, A. Beiser, M. Breteler, L. Fratiglioni, C. Helmer, H. Hendrie, H. Honda, M. Ikram, K. Langa, A. Lobo et al. (2017). 'The changing prevalence and incidence of dementia over time - current evidence'. In: *Nature Reviews Neurology* 13.6, pp. 327–339 (cited on page 1).

R. Xu and D. Wunsch (2009). *Clustering*. Hoboken, New Jersey: John Wiley & Sons, Inc. (cited on pages 139, 142 and 147).

A. Young, R. Marinescu, N. Oxtoby, M. Bocchetta, K. Yong, N. Firth, D. Cash, D. Thomas, K. Dick, J. Cardoso et al. (2018). 'Uncovering the heterogeneity and temporal complexity of neurodegenerative diseases with Subtype and Stage Inference'. In: *Nature Communications* 9.4273 (cited on page 143).

O. Zanetti, C. Geroldi and G. Frisoni (2009). *Mild Cognitive Impairment (MCI) is not a clinical entity*. Online. Accessed: September 2020. URL: https://www.bmj.com/rapid-response/2011/11/02/mild-cognitive-impairment-mci-not-clinical-entity (cited on page 103).

X. Zhu, C. Loy and S. Gong (2014). 'Constructing Robust Affinity Graphs for Spectral Clustering'. In: *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 1450–1457 (cited on pages 67 and 143).

M. Zwitter, M. Soklic, I. Kononenko and B. Cestnik (1988). *Lymphography Data Set*. Online. Accessed: November 2020. URL: https://archive.ics.uci.edu/ml/datasets/Lymphography (cited on page 155).