

**Assignment for PMIM-702 Dissertation**

|  |  |
| --- | --- |
| **Module number:** | PMIM-702 |
| **Module name:** | Dissertation (Research Paper) |
| **Title of assignment:** | *Predicting the diagnosis of Alzheimer’s disease with supervised machine learning* |
| **Student ID number:** | 952781 |
| **Word count:** | 5009 |
| **Declaration:** | I understand the following conditions which apply throughout this course:   1. I confirm that I am the sole author of this work. 2. I understand that proof reading by a third party is discouraged, but if used, records should be available as per guidelines. 3. I understand the need for academic integrity and that all my submitted work will adhere to its principles. 4. I understand that the teaching team will take measures to deter, detect and report any academic misconduct. 5. I agree to my work being submitted to the TurnItIn academic database. 6. I understand the importance of assignment deadlines and the need to seek help in good time where personal circumstances interrupt my work. |
| **Please copy and paste this declaration onto the front of the submission.** | |

Predicting the diagnosis of Alzheimer’s disease with supervised machine learning

By Andre Ng

952781

Swansea University Medical School

Word Count: 5009 words

**Abstract**

**Objectives**

The ability to identify Alzheimer’s disease (AD) patients early and accurately is crucial for early prevention strategies. In this study, routinely collected electronic health records (EHR) were used to train models with supervised machine learning (ML) algorithms to predict a five-year risk of developing AD.

**Methods**

A case-control study was performed, using GP and hospital data from the Secure Anonymised Information Linkage (SAIL) databank. AD patients were categorically matched with controls by year of birth, gender, and deprivation score. Two datasets were produced, an unbalanced dataset with 1 case to 5 controls (n = 92674), and a balanced dataset with 1 case to 1 control (n=30858). The features were the 50 most commonly found Read codes in AD patients five years before diagnosis, where the number of times a patient had the Read code was counted, alongside their corresponding average event value. Random forest, naive Bayes, decision trees, generalised linear/logistic models and SVM were trialled.

**Results**

Unbalanced dataset produced results heavily skewed results towards sensitivity, except for naïve Bayes. In the balanced dataset, results were more even, and random forest (k=10) performed best with 0.964 AUC (0.873 accuracy, 0.860 sensitivity and 0.887 specificity), followed by decision trees (k=10) with 0.887 AUC, while other models performed significantly poorer. The top features from random forest were ALT/SGPT serum level, serum bilirubin level, GFR (abbreviated MDRD), and serum urea level.

**Conclusions**

Liver and kidney biomarkers were shown to be important predictors, suggesting a correlation between AD and liver/kidney functioning. This study also demonstrated that ML with routinely collected EHR is a viable option to predict a diagnosis of AD early and accurately.

**Introduction & previous work**

Alzheimer’s disease (AD) is the most common cause of dementia and it is a progressive neurological disorder that can ultimately prove fatal (Vickers et al., 2000). This age-associated condition is the largest unmet medical need in neurology (Citron, 2010), and it is because currently there are no available disease-modifying treatments, with only symptomatic medications that can be given to patients to maintain function and manage symptoms (British National Formulary, 2021; Joe & Ringman, 2019). In Wales, AD is prevalent in up to 5% of the population for 75 years or older (Abdulrahman, 2014). Despite a decline in dementia incidence, the number of AD patients will increase as life expectancy improve (Ahmadi-Abhari et al., 2017). Its symptoms include memory loss, irritability, depression, anxiety, and others, which could, in turn, lead to secondary effects such as impairment in daily activities, caregiver stress and increased cost of care (Lanctôt et al., 2017). With the 44 million cases of AD in 2015 predicted to double by 2050, it is an increasing economic and social burden for countries around the world (Mendiola-Precoma, Berumen, Padilla, & Garcia-Alcocer, 2016). In the US, it is estimated that 277 billion US dollars had been spent on long term healthcare for dementia patients in the year 2018 alone (Alzheimer Association, 2018). However, mathematical models suggest an early and accurate diagnosis of AD can help save up to 7.9 *trillion* dollars in medical care costs, as well as possible important personal benefits (Alzheimer Association, 2018). These benefits include early prevention strategies like changes to personal lifestyles and prevention therapies, or in the future, early AD treatments as drug development shifts to early prevention of cellular pathology (Graham, Bonito-Oliva, & Sakmar, 2017). The ability of early AD prediction is likely to have a profound positive impact both on a personal level, and on a national scale.

Currently, AD diagnosis may include neuropsychological assessments, neuroimaging and cerebrospinal fluid analyses, which are respectively time-consuming, expensive to set up or operate, and invasive (Laske et al., 2015). Therefore, some researchers had opted for a different approach by taking advantage of the growing collection of health-related data and recent advances in machine learning (ML). By using existing data from health organisations, supervised ML algorithms can be trained to create models that predict the diagnosis of AD. These models may be trained from a variety of different health data types, such as neuroimages (Ahmed et al., 2019), unstructured clinical notes (Miled et al., 2020), wearables (Saif et al., 2020), and electronic health records (EHR) (Ford et al., 2019). While each data type has its strengths and limitations, EHR is the most widely collected form of data, allowing for detailed insight into AD patients from an inexpensive source. One EHR database is the Secured Anonymised Information Linkage databank (SAIL), which stores anonymised and routinely collected EHR from the National Health Service in Wales, UK. Containing over 4.4 million unique participants (Schnier et al., 2020), it is the database used in this study. For this paper, supervised ML models will be trained by using routinely collected data, predicting between two classes: AD cases and controls, five years in advance. The objective is to train an accurate predictive model and extract key features.

**Methods**

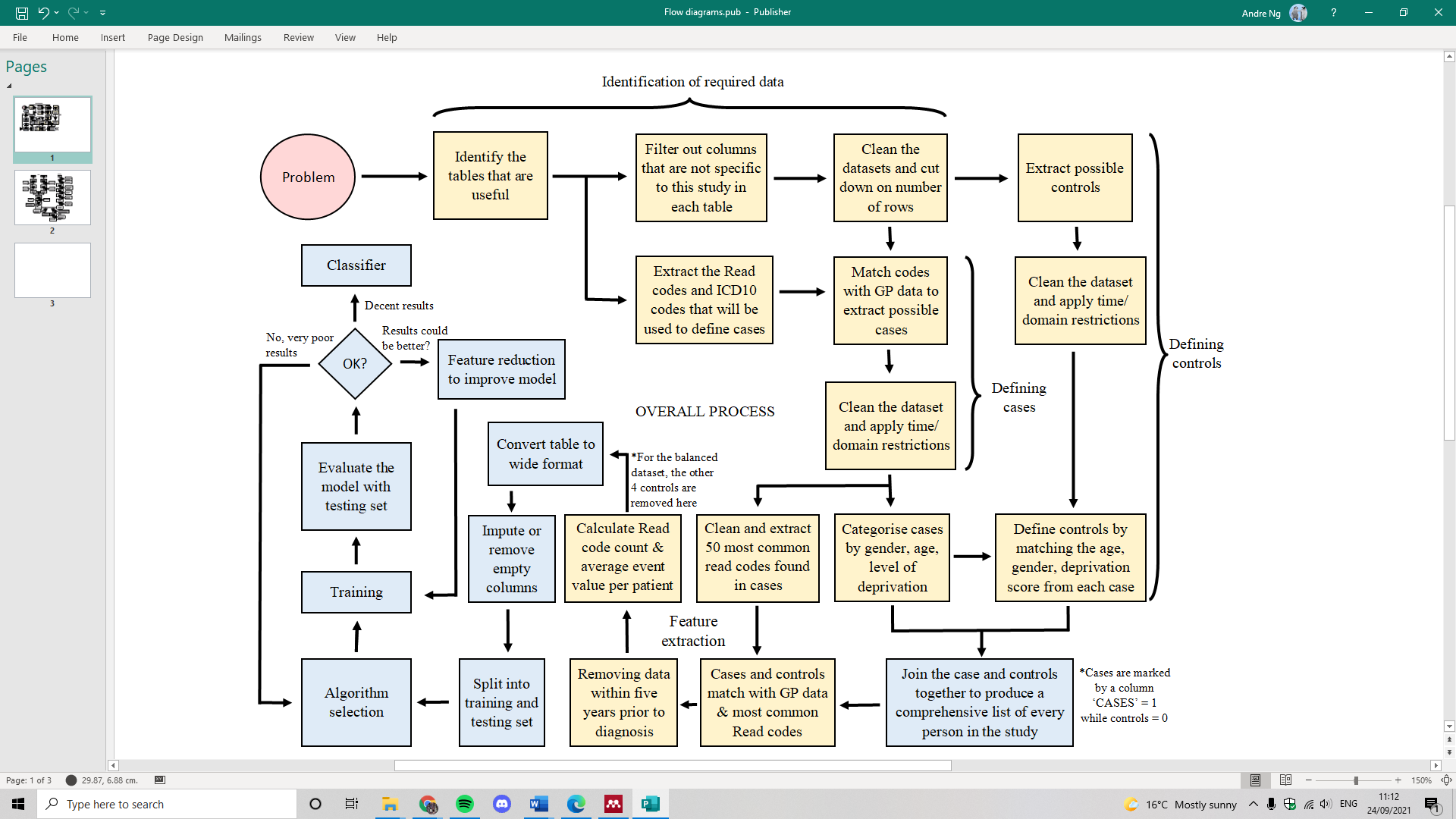


Figure 1: The overall process of training a model to predict Alzheimer’s diagnosis in five years by using supervised machine learning algorithms (Yellow coloured – written in SQL, Blue coloured – written in R)

In this study, data from the Secured Anonymised Information Linkage (SAIL) databank was used, which includes general practitioner (GP) datasets, hospital records, death data and demographics data from across Wales, UK. As there were many factors and only two outcomes, a case-control study was performed on patients who were diagnosed with AD between 2015 and 2019. This allowed models to be trained with classifier algorithms to predict a five-year risk of developing AD, all data available before the five years prior to diagnosis was used to simulate a prediction scenario (Figure 2). The AD patients (cases) were defined by Read codes (version 2) from the GP dataset and ICD10 codes from the hospital datasets, while the features tested were extracted from Read codes in the GP dataset. Controls were categorically matched with each case by their year of birth, gender, and deprivation score of where they lived. For both cases and controls, domain and time restrictions were applied (Figure 3) with all the datasets mentioned above. The computer software used in this study were Eclipse – which uses the programming language SQL, and R studios – which uses R (Figure 1).

**Definition of cases and controls**

Table 1: Case definition by Read codes and ICD10 codes

|  |  |  |
| --- | --- | --- |
| **Code** | **Code Type** | **Description** |
| Eu00. | Read V2 | Dementia in Alzheimer's disease |
| Eu000 | Read V2 | Dementia in Alzheimer's disease with early onset |
| Eu001 | Read V2 | Dementia in Alzheimer's disease with late onset |
| Eu00z | Read V2 | Dementia in Alzheimer's disease, unspecified |
| F110. | Read V2 | Alzheimer's disease |
| F1100 | Read V2 | Alzheimer's disease with early onset |
| F1101 | Read V2 | Alzheimer's disease with late onset |
| Fyu30 | Read V2 | Other Alzheimer's disease |
| 129B. | Read V2 | Alzheimer's disease |
| F000 | ICD10 | Dementia in Alzheimer's disease with early onset |
| F001 | ICD10 | Dementia in Alzheimer's disease with late onset |
| F009 | ICD10 | Dementia in Alzheimer's disease, unspecified |
| G300 | ICD10 | Alzheimer's disease with early onset |
| G301 | ICD10 | Alzheimer's disease with late onset |
| G308 | ICD10 | Other Alzheimer's disease |
| G309 | ICD10 | Alzheimer's disease, unspecified |

The cases were defined by both Read codes in GP data and ICD10 codes (or diagnosis codes) in hospital datasets to increase the sample size (exact codes are shown in table 1). The AD codes were extracted from two built-in datasets in SAIL that highlights all the Read codes and ICD10 codes in use within the databank (named SAILREFRV.READ\_CD and ICD10\_DIAG\_CD), where the word “Alzheimer’s” was searched for in the description boxes of the two datasets to extract the codes. The codes were then filtered and cross-checked with existing literature to produce table 1 (Stocks et al., 2017; Yang et al., 2020). Any patients found with these codes were considered as possible cases. The data were then cleaned and filtered to fit within time and domain restrictions, these include they must be diagnosed with AD within the year 2015 to 2019, must be alive when diagnosed with AD and must be at least 18 years old (Flow chart – figure 3). The controls were then categorically matched with cases by three categories: year of birth, gender, and deprivation score. The deprivation score is between 1 to 5 (1 most deprived, 5 least deprived) and it is derived from where an individual’s address score on the Welsh Index of Multiple Deprivation (WIMD) quintile (2014). As there may be multiple addresses across time, the address that was recorded on their date of diagnosis was used. In the unbalanced case-control dataset, five controls were matched with one case, while the balanced dataset has a one-to-one ratio. The controls also go through cleaning and filtering to fit within the time and domain restrictions. These include: Alive, does not already exist in cases, does not have AD or any other dementias, did not move address during the study period (see figure 3). A total of 15429 cases and 77145 controls were included in the study.

**Feature selection**

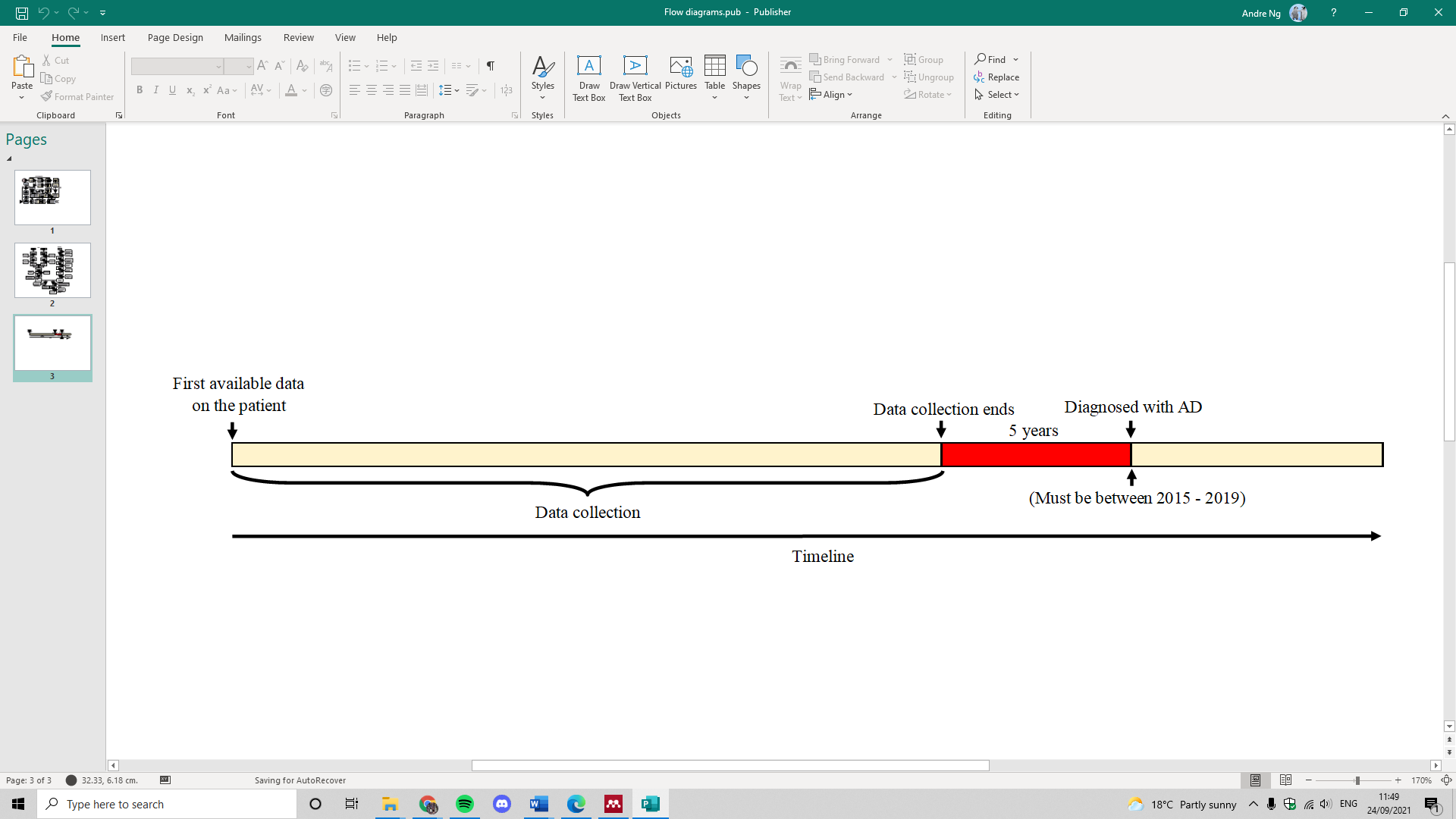


Figure 2: Data collected in terms of patient timeline

The cases were then matched with GP data to identify the top 50 most common Read codes in AD cases in the years leading up to the five years before diagnosis (Figure 2). As the study is data-driven, minimal input had been placed to select features. Despite this, the Read codes were still checked for duplicates, blanks, “NA”s, and for certainty – they must not already exist within the AD list of Read codes. In addition, they must not be dementia codes as well (Appendix 1). Dementia Read codes were determined by the same method used in the creation of the AD code lists – by searching for the word “dementia” in the description column of the SAILREFRV.READ\_CD dataset (Appendix 2). The list was then checked and cleaned. The most common features were matched with the dementia list to ensure there were no matching codes. Once the 50 Read codes were determined, they were matched with both cases and controls. The total number of occurrences per patient, per Read code was counted. In addition, their “average event value” were also calculated. The “event value” is related to its particular Read code, and its meaning could range from the value of diastolic blood pressure to the number of tablets taken. If a patient had multiple records of the same Read code with varying “event values”, the mean of the “event values” was determined as “average event value”.

After converting the table from long to wide with one-hot-encoding, two sets of columns (or features) were created: 1) the total number of occurrences per Read code, and 2) the Read code’s average event value. However, there may be empty rows of average event values. If an average event value column had more than 50% rows filled, then the empty cells were imputed with the median (separated by cases and controls). This is similar to the population mean imputation method used by Pekkala *et al*. (2017), though median was used instead as it is less susceptible to outliers. Otherwise, the column is removed completely. Therefore, only some Read codes would have an associating column of “average event value”.

**ML algorithms**

Once the data was prepared, it was split into a ‘test’ set and ‘train’ set, at a ratio of 25% to 75%, respectively. The model training was performed on an unbalanced dataset (1 case to 5 controls), where the models with the best performance were then optimised further. However, as the unbalanced dataset produced results heavily skewed towards specificity (Table 2), a balanced dataset (1 case to 1 control) was later used instead. By using the ‘caret’, ‘rpart’, ‘randomForest’, ‘kernlab’ packages in R studios, various classifiers were trialled for the best performance. This included the most frequently used algorithms in similar studies, such as random forest, support vector machines (SVM), naïve Bayes, generalised linear model, generalised logistic model, and decision trees. As k-fold cross-validation was time-consuming, only the best model was k-fold cross-validated to further improve accuracy and avoid overfitting. In addition, the best model’s variable importance list was also used for feature reduction in a bid to improve its performance. Most models produced were accompanied with their corresponding receiver operating characteristic (ROC) curves and Area Under Curve (AUC/AUROC) scores, by using the package ‘yardstick’ and ‘pROC’.

**Ethics**

This project is covered and approved by the Information Governance Review Panel (IGRP) under project ID 1074. Since this study uses anonymised and routinely collected data from SAIL, it was exempt from other specific approvals. SAIL databank is also independently certified by ISO27001, a best practice standard for information security management systems. All data extracted were approved by SAIL Data Guardians.

**Results**

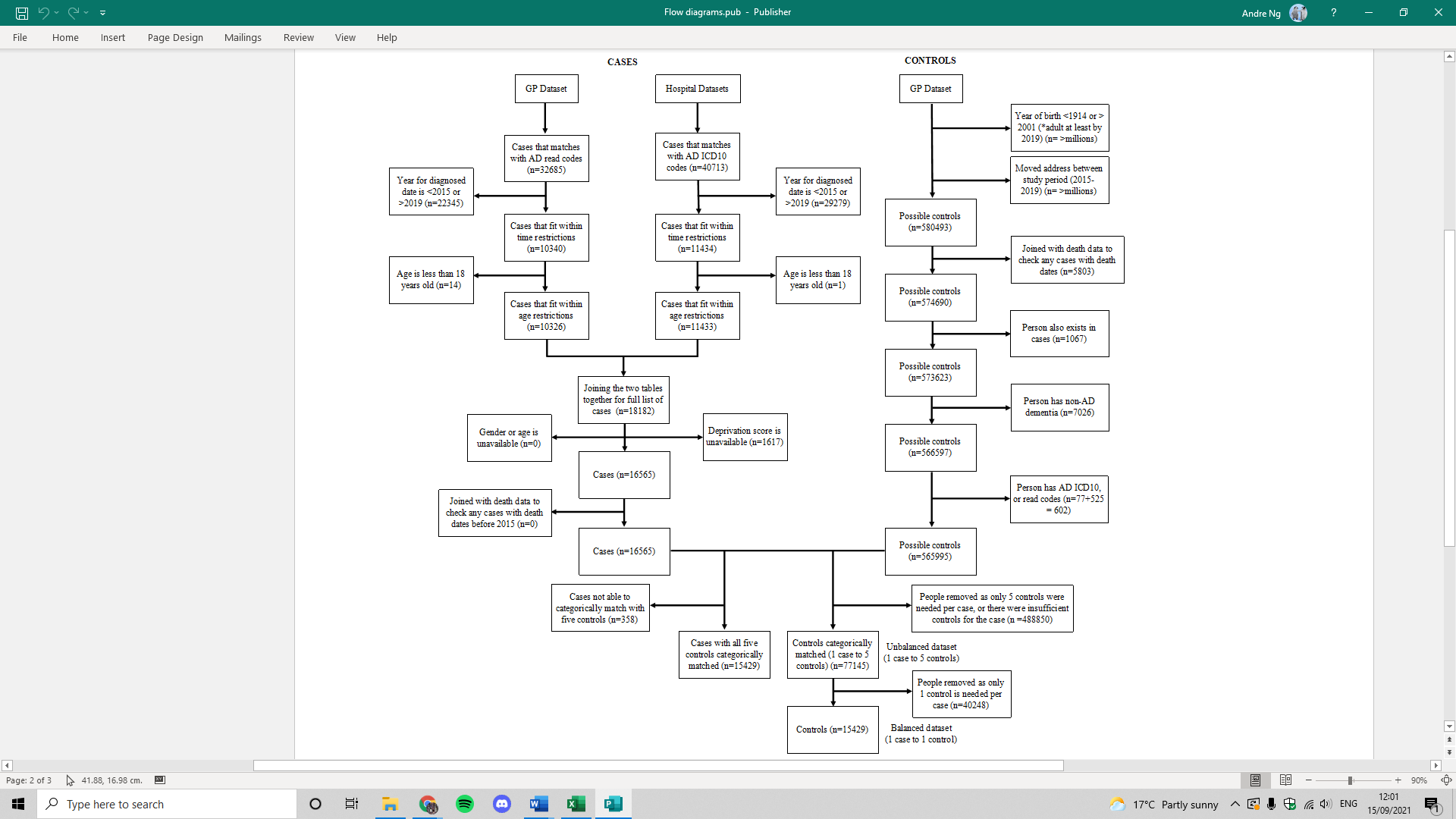


Figure 3: Flow diagram for case and control definitions (see appendix 1 for dementia Read codes used to exclude controls, appendix 2 for all 50 most common AD features)

Table 2: Results from models trained with unbalanced dataset (k = 10, means 10-folds cross validation performed to produce the model)

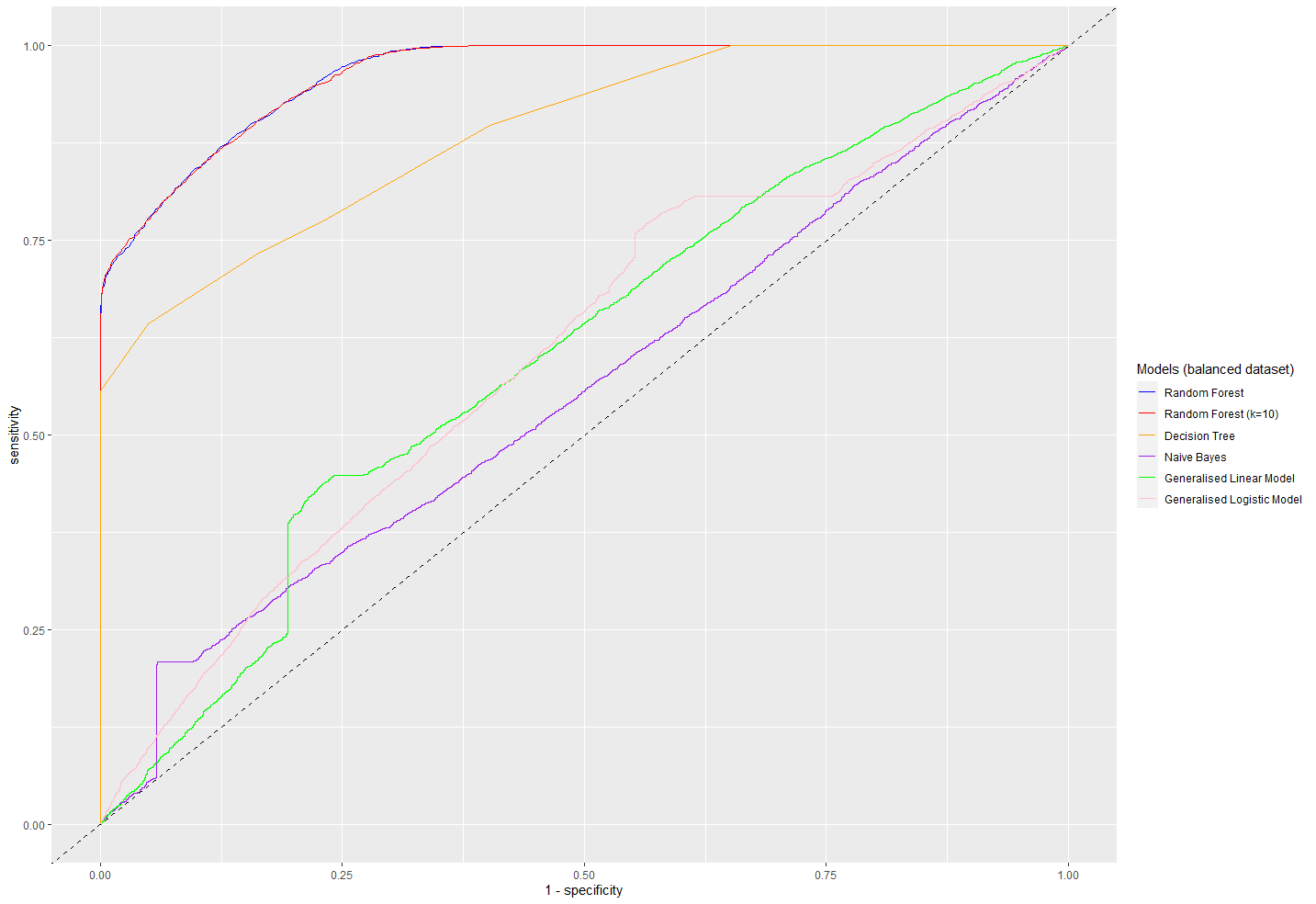
|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| Algorithm | Model Name | Accuracy | Sensitivity | Specificity | AUC |
| Random Forest | rfmodel1 | 0.9501 | 0.7009 | 1.0000 | 0.9620 |
| Naïve Bayes | nbFit | 0.4114 | 0.7112 | 0.3515 | 0.5570 |
| Decision Tree (k=10) | dtFit1 | 0.9314 | 0.7943 | 1.0000 | 0.8920 |

As seen in table 2, the unbalanced dataset with one case to five controls led to a heavily skewed sensitivity and specificity ratio. Most models were skewed towards specificity with complete 1.0000 specificity, aside from naïve Bayes, which was heavily skewed towards sensitivity in both unbalanced and balanced datasets (Table 3). Therefore, the balanced dataset (1 case to 1 control) was prioritised.

Table 3: Results from models trained with balanced dataset (k = 10, means 10-folds cross validated performed to produce the model) (AUC for SVM unavailable due to long processing time and time constraints)

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| Algorithm | Model Name | Accuracy | Sensitivity | Specificity | AUC |
| Random Forest | bal\_rfmodel1 | 0.8714 | 0.8548 | 0.888 | 0.963 |
| **Random Forest (k=10)** | **bal\_rfmodel3** | **0.8716** | **0.8577** | **0.8854** | **0.9630** |
| SVM | bal\_svmFit1 | 0.5638 | 0.7107 | 0.4168 | N/A |
| Naïve Bayes | bal\_nbFit | 0.5022 | 0.0625 | 0.9419 | 0.5560 |
| Generalised linear model (Binomial model) | bal\_glmFit1 | 0.5723 | 0.6356 | 0.5091 | 0.6030 |
| Generalised logistic model (Binomial model) | bal\_glgmFit1 | 0.5723 | 0.6356 | 0.4909 | 0.6029 |
| **Decision tree (k=10)** | **bal\_dtFit1** | **0.7967** | **0.6426** | **0.9508** | **0.887** |

The results collected from models trained with the balanced dataset (one to one control) showed much more balanced ratios between sensitivity and specificity. Random forest and decision tree models performed best with accuracy between 0.797 to 0.872, sensitivity between 0.643 to 0.858, specificity between 0.885 to 0.951. Despite naïve Bayes achieving a higher sensitivity than other models (0.941), its specificity was very low at 0.063. As random forest was the best performer, it was trained again with 10-fold cross-validation, which leads to marginally better performance in accuracy and sensitivity, while specificity decreased by 0.003. Additionally, the R package used for the decision tree model (‘rpart’) auto-implements 10-fold cross-validation. Other models all performed relatively poorly with accuracy less than 0.600, with naïve Bayes the worst performer at 0.502 accuracy, 0.063 sensitivity and 0.942 specificity.

Figure 4: ROC curves from all models (balanced dataset) except for SVM

As seen in figure 4, most models performed poorly, aside from the random forest and decision tree models. Between the two random forest models (one with 10-fold while the other does not), they had an almost identical ROC curve, with a slight variation at 0.25 1-specificity. Naïve Bayes and generalised linear model had a surge in sensitivity at low 1-specificity but evens out afterwards.

Table 4: Results from the two best models (random forest & decision tree) after feature selection

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| Algorithm | Model Name | Accuracy | Sensitivity | Specificity | AUC |
| Random Forest (k=10) | bal\_rfmodel3A | 0.8734 | 0.8595 | 0.8872 | 0.9640 |
| Decision Tree (k=10) | bal\_dtFit1A | 0.7967 | 0.6427 | 0.9508 | 0.8870 |

In an attempt to increase performance in both random forest and decision tree models from the balanced dataset, feature selection was performed on the best models (bal\_rfmodel3 and bal\_dtFit1), while performance improved marginally in random forest (Figure 4), it maintained the same with decision trees (Table 4).

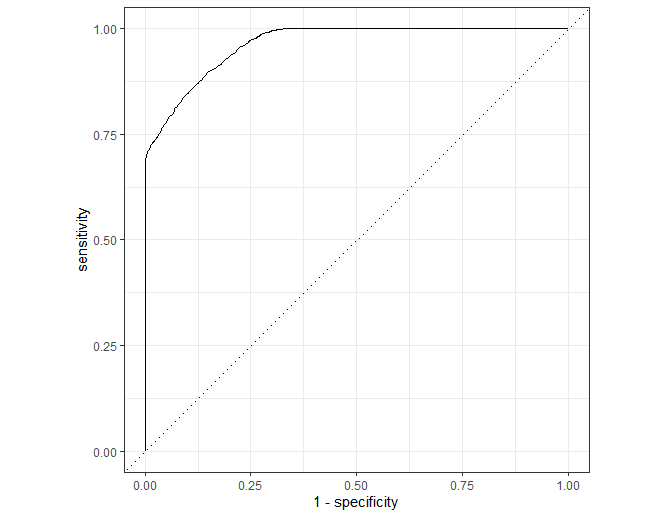
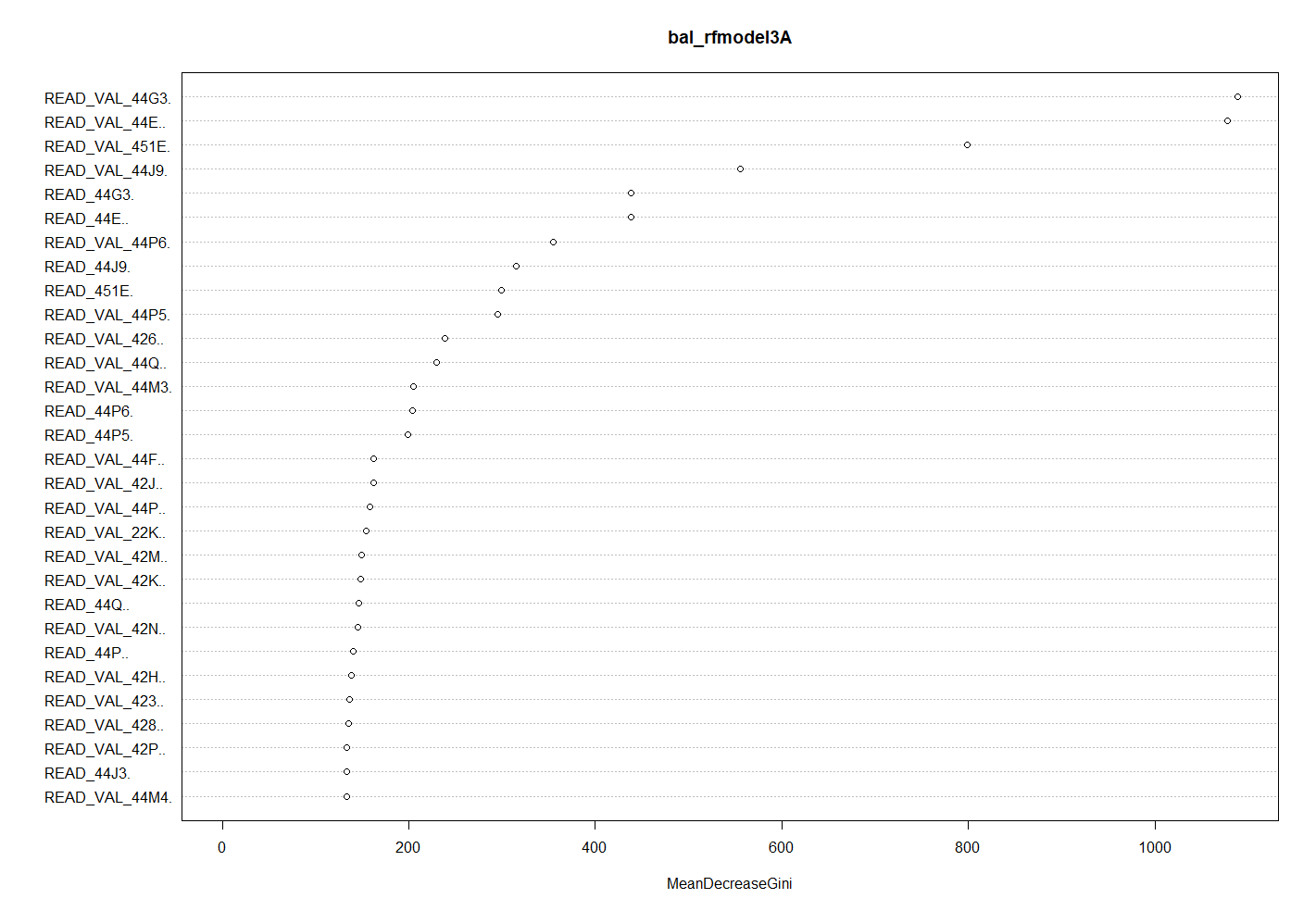


Figure 5: ROC curve of the *feature selected* random forest model (k=10) trained from the balanced dataset (bal\_rfmodel3A) (Produced with R studios and ‘yardstick’ package)

Figure 6: The top 30 most important features from the *feature selected* random forest model (k=10) trained from the balanced dataset (bal\_rfmodel3A) with the average event value (marked by \_VAL) and Read codes count (without the \_VAL) as the features on the y-axis, while the mean decrease Gini show variable significance (Produced with R studios and ‘yardstick’ package)

As seen in figure 6, the average event value features were significantly more important than the count features, with average event value features taking the top four places for most important variables.

Table 5: Feature descriptions, mean, median of event count (COUNT) and average event value (VAL) between cases and controls for bal\_rfmodel3A, where mean decrease Gini show level of variable importance (As some patients may have no values at all for certain features, the mean/median shown here are only for those who *does* *have* a value before imputation. Additionally, median values were further rounded up due to data protection reasons)

|  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- |
| Read Code | Type | Descriptions  (\* = repeated) | Mean | | Median | | Mean Decrease Gini |
| Control | Case | Control | Case |
| 44G3. | VAL | ALT/SGPT serum level\* | 25.228 | 24.199 | 22.3 | 21.2 | 1089.1 |
| 44E.. | VAL | Serum bilirubin level\* | 10.893 | 10.824 | 9.9 | 9.7 | 1078.2 |
| 451E. | VAL | GFR calculated abbreviated MDRD\* | 67.193 | 67.386 | 67.0 | 67.3 | 798.6 |
| 44J9. | VAL | Serum urea level\* | 6.109 | 5.976 | 5.9 | 5.8 | 556.0 |
| 44G3. | COUNT | ALT/SGPT serum level\* | 7 | 7 | 5 | 5 | 438.1 |
| 44E.. | COUNT | Serum bilirubin level\* | 7 | 7 | 6 | 6 | 437.9 |
| 44P6. | VAL | Serum LDL cholesterol level\* | 3.002 | 3.022 | 3 | 3 | 354.8 |
| 44J9. | COUNT | Serum urea level\* | 7 | 7 | 5 | 5 | 315.1 |
| 451E. | COUNT | GFR calculated abbreviated MDRD\* | 7 | 7 | 5 | 5 | 299.4 |
| 44P5. | VAL | Serum HDL cholesterol level\* | 1.499 | 1.518 | 1.4 | 1.4 | 295.2 |
| 426.. | VAL | Red blood cell (RBC) count | 4.637 | 4.820 | 4.5 | 4.5 | 238.5 |
| 44Q.. | VAL | Serum triglycerides\* | 1.567 | 1.548 | 1.4 | 1.4 | 229.6 |
| 44M3. | VAL | Serum total protein | 70.799 | 70.783 | 70.9 | 70.8 | 204.7 |
| 44P6. | COUNT | Serum LDL cholesterol level\* | 7 | 6 | 6 | 5 | 203.7 |
| 44P5. | COUNT | Serum HDL cholesterol level\* | 7 | 7 | 6 | 6 | 199.0 |
| 44F.. | VAL | Serum alkaline phosphatase | 90.646 | 87.274 | 79.2 | 77.9 | 162.6 |
| 42J.. | VAL | Neutrophil count | 5.105 | 5.211 | 4 | 4.1 | 162.1 |
| 44P.. | VAL | Serum cholesterol\* | 7.065 | 5.691 | 5.1 | 5.1 | 158.3 |
| 22K.. | VAL | Body Mass Index | 176.326 | 217.441 | 26.8 | 26.4 | 154.7 |
| 42M.. | VAL | Lymphocyte count | 2.663 | 2.556 | 2 | 2 | 149.3 |
| 42K.. | VAL | Eosinophil count | 0.485 | 0.467 | 0.2 | 0.2 | 148.1 |
| 44Q.. | COUNT | Serum triglycerides\* | 8 | 8 | 7 | 7 | 146.0 |
| 42N.. | VAL | Monocyte count | 0.666 | 0.671 | 0.5 | 0.5 | 145.4 |
| 44P.. | COUNT | Serum cholesterol\* | 9 | 10 | 8 | 8 | 140.3 |
| 42H.. | VAL | Total white cell count | 6.966 | 6.998 | 6.7 | 6.8 | 138.4 |
| 423.. | VAL | Haemoglobin estimation | 19.141 | 19.440 | 14.1 | 14 | 136.6 |
| 428.. | VAL | Mean corpusc. haemoglobin (MCH) | 30.364 | 30.399 | 30.4 | 30.4 | 135.0 |
| 42P.. | VAL | Platelet count | 261.284 | 261.440 | 254.7 | 254.5 | 133.9 |
| 44J3. | COUNT | Serum creatinine | 12 | 12 | 5 | 5 | 133.5 |
| 44M4. | VAL | Serum albumin | 41.945 | 41.816 | 42.1 | 42 | 133.0 |

In table 5, the top features from bal\_rfmodel3A was shown. Most of the top 15 features had the word “serum”, showing the significance of serum-related features. Whereas the lower 15 features included many blood-related Read codes such as monocyte counts, white blood cell counts, platelet counts etc. When comparing the mean values between cases and controls, the differences were small. However, the differences in median were often even smaller. In Read code “22K..”, there was a significant difference between mean and median calculations (mean: 176.3 and 217.4, median: 26.8 and 26.4), suggesting there were several outliers in the average event values of the body mass index.

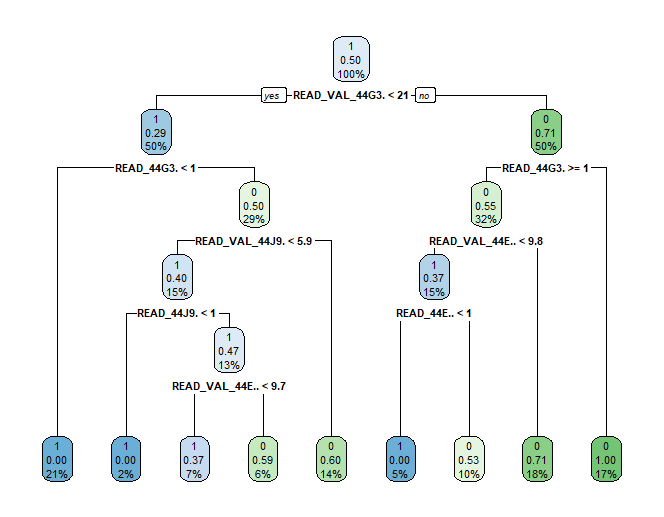


Figure 7: Decision tree of the model named “bal\_dtFit1” trained from the balanced dataset (k=10) (average event value = VAL, count = no VAL) (Produced with R studios and ‘yardstick’ package)

The decision tree model from the balanced dataset showed their most important features (Figure 7), with ‘44G3.’, ‘44J9.’ and ‘44E..’ are all present in the top features of the random forest model (bal\_rfmodel3A). However, when compared to the random forest model (Table 5), event counts seem to show more significance, as seen by ‘44G3.’ counts were above average event values of ‘44J9.’ and ‘44E..’.

**Discussion**

In this study, supervised machine learning algorithms were used to predict a diagnosis of AD five years ahead, where random forest was the most accurate model at 0.873 accuracy, 0.860 sensitivity, 0.887 specificity and 0.964 AUC. Aside from decision trees which had reasonable results (0.887 AUC), other models performed significantly poorer. For example, naïve Bayes had an AUC of 0.556 (Table 3) with a poor ROC curve (Figure 4), which may be due to naïve Bayes’ assumption that all features did not correlate with each other. As later shown in table 5, this is untrue as features were often serum-related or blood-related. For generalised linear and logistic models, they likely underperformed due to the large number of features found in this study. As there are many predictors for AD, the random forest algorithm benefitted from applying a weak learning strategy to extract features that are most predictive (Breiman, 2001), outperforming “greedy” algorithms like decision trees (Table 3). Additionally, the high performance can be attributed to the vast collection of data from SAIL. As reported by Yang *et al*. (2020), their total number of features decreased as their prediction window widens from one year to five years ahead, leading to a decrease in model performance. While Yang *et al*. (2020) had a total of 4669 unique features in their five-year prediction window (resulting in 0.680 AUC), the SAIL databank allowed this study to collect a total of ~36992 unique features from cases. Despite only the 50 most common Read codes from AD cases were used, the wide variety of features would allow distinctive features uncommonly found in other studies to be included.

**Top features**

As seen in table 5, the top two features for the random forest model (bal\_rfmodel3A) were ALT/SGPT serum level and serum bilirubin level, showing significantly higher importance than most in classifying AD patients and controls (Figure 6). ALT (alanine aminotransferase), SGPT (serum glutamic pyruvic transaminase) are liver-related enzymes. ALT, SGPT and bilirubin were shown to correlate with liver diseases and injury (Cohen & Kaplan, 1979; Kwak et al., 2012; Srivastava, Kumar, Agarwal, & Ranjan, 2007). And liver problems are correlated with AD, as demonstrated by Bassendine, Taylor-Robinson, Fertleman, Khan, & Neely (2020). They found that Amyloid-β deposits – which accumulates in the brain of an AD patient “long before cognitive decline is evident”, could be originated from the liver. Additionally, they also suggested the liver is involved in peripheral clearance of Amyloid-β deposits in blood circulation. In another paper by D. G. Kim *et al*. (2016), they found that mice induced with chronic liver diseases led to a decrease in a protein involved with Amyloid-β clearance (low-density lipoprotein receptor-related protein-1), which leads to pathological signs of AD. This suggests that ALT and SGPT could potentially be important predictors for early diagnosis of AD, as they forecast complications in the liver, which could lead to an increase in Amyloid-β in the brain. Despite liver diseases is not identified as an important risk factor for dementia in the Lancet commission, nor in the Lifestyle for Brain Health (Libra) index –a common method to calculate dementia risk (Bin-Hezam & Ward, 2019; Livingston et al., 2020), diabetes – a condition closely related to liver diseases, was cited by both as an important predictor (Bin-Hezam & Ward, 2019). Additionally, ALT levels were known to be a good indicator for overall health in the context of obesity and cardiovascular diseases (W. R. Kim, Flamm, Di Bisceglie, & Bodenheimer, 2008). For which, obesity was cited by both Lancet commission and Libra index, whereas chronic heart disease was cited by the Libra index as an important predictor as well (Bin-Hezam & Ward, 2019). This suggests that ALT is likely to have a close relationship with AD as well as *the conditions that leads to AD*. Potentially, it could be a precursor for known AD biomarkers like Amyloid-β, though more research is needed to understand its role in early AD diagnosis.

The top two features were then followed by “GFR calculated abbreviated MDRD” (Table 5), which is a measure of kidney function. GFR (Glomerular Filtration Rate) is the filtration rate for plasma fluid in kidneys and GFR in adults can be estimated from the abbreviated MDRD (Modification of Diet in Renal Disease) equation – a simplified version of the original MDRD equation (Levey et al., 1999; National Kidney Foundation, 2014). Despite evidence had shown limitations, GFR is often used as a classification schema to diagnose patients with chronic kidney disease (CKD) (Glassock & Winearls, 2010; Ma et al., 2006). This feature suggests that kidney function is correlated with early diagnosis of AD, which would be on par with current research. CKD is one of the key factors to AD and related dementias, listed in both the Libra index and Lancet commission (Bin-Hezam & Ward, 2019; Livingston et al., 2020). However, the direct pathological mechanisms behind this connection are elusive, and it is made more difficult by the multifactorial nature of AD and CKD (Stanciu et al., 2020; Zhang, He, Su, Zhang, & Meng, 2020). However, GFR might be a better predictor for early AD diagnosis than CKD itself, despite CKD is already recognised as an important factor. This is because there is only a marginal difference (0.3 mL/min/1.73m^2) in GFR (Table 5), because CKD stages are relatively broad and there are limitations to its classification system.

The feature “GFR abbreviated MDRD” is then followed by serum urea level (Table 5), another predictor that is closely related to kidney functioning (Bamanikar, Bamanikar, & Arora, 2016). Further supporting the hypothesis that kidney functioning is an important AD predictor, though no literature had shown a direct association between AD and serum urea level. However, apart from the random forest model, serum urea level is also one of the top features in the decision tree model (Figure 7 – 44J9.), alongside ALT/SGPT serum levels and serum bilirubin levels. These three features and “GFR calculated abbreviated MDRD” would conclude the top nine features found in the random forest model (Table 5). Overall, the average event values were more important than the count values, yet for the top features, feature counts were surprisingly more important than any other average event values, showing the significance of these four features.

**Other features**

For the rest of the features, most were information collected from blood tests. For example, red blood cell count, total white blood cell count, platelets count and mean corpuscular haemoglobin were collected from routine complete blood count tests (Fletcher & Sampson, 2020), these blood-related features likely indicate the overall health of an individual (Reddy, 2014). Whereas cholesterol features like serum LDL/HDL (High/Low-Density Lipoprotein) cholesterol are features specifically detected from lipoprotein panels, which may be applied to a patient if the GP suspect possible coronary heart diseases and other atherosclerotic problems (Fletcher & Sampson, 2020). These features would indicate an association between AD and cholesterol, or heart diseases, which is supported by recent research. It had been shown that LDL cholesterol influences the development of AD, but the causality between AD and cholesterol is unclear (Sáiz-Vazquez, Puente-Martinez, Ubillos-Landa, Pacheco-Bonrostro, & Santabárbara, 2020). Additionally, Liu *et al*. (2020) showed that serum triglyceride levels – another lipid detected from the lipoprotein panels, and serum LDL cholesterol levels were found to be higher in AD patients. While the increase in LDL cholesterol for AD cases was supported by this study (Controls: 3.002 mmol/dL, Cases 3.022 mmol/dL), serum triglycerides in AD cases decreased when compared to controls (Controls: 1.567 mmol/L, Cases: 1.548 mmol/L) (Table 5). The exact reasoning is difficult to determine, but the limited evidence on serum triglycerides and AD suggests further study is required. Finally, other features were again related to liver/kidney diseases, further supporting their relationships with AD. For example, serum albumin – are produced by the liver into bloodstream (Pupim, Martin, & Ikizler, 2013), and serum alkaline phosphatase – found in kidneys, livers, are associated with bilirubin levels and chronic liver diseases (Iluz-Freundlich et al., 2021).

**Comparison to other studies, limitations** **& future work**

In comparison to other studies, this study was more focused on internal validity. For example, this study predicted between “pure” AD cases and non-AD controls, whereas Ford *et al*. (2019) predicted dementia cases, and Boustani *et al*. (2020) predicted “Alzheimer’s disease and related dementias” rather than AD alone. Additionally, some studies predicted between non-AD controls, AD, and Mild Cognitive Impairment (MCI) – an intermediate stage between dementia and normal cognitive ageing (Aschwanden et al., 2020; Stephan, Kurth, Matthews, Brayne, & Dufouil, 2010). However, its definition varies between studies and may lead to inaccurate diagnosis of dementia (Nori et al., 2019). Henceforth, it was not included in this study. Furthermore, some studies predicted between multiple time windows, such as Yang *et al*. (2020) who predicted zero, one, three, five years in advance, and Boustani *et al*. (2020) – one, three, and five years, yet this study predicted five years only. This approach had trade off the external validity of the model for better model performance, which would limit its generalisability on public use. For limitations, the use of GP and hospital datasets as case definitions might cause bias, leading to more hospital-correlated features, and giving people who go hospitals more a higher chance of being diagnosed. As controls were defined by GP datasets only, the controls are also less likely to go to hospitals than cases. Furthermore, feature imputations could skew controls’ average event values, as empty cells could mean the person had “normal” values (eg. person is healthy thereby not needing a test). Additionally, there could be misdiagnosis and delayed diagnosis of AD from the data itself, which should be addressed in future work.

Nevertheless, the best performing model – random forest, achieved an AUC of 0.964 which would still be very useful in early diagnosis of AD, a type of dementia that accounts for 60-70% of all dementia cases (WHO, 2020). In addition, random forest was also the best model for Park *et al*. (2020), who also applied linear regression and SVM. Their data-driven approach also led to similar important predictors, such as haemoglobin levels, age, and urine protein levels, supporting the importance of blood-related features, and proteins like alkaline phosphatase (Table 5) which is detected in blood and urine (Amador, Zimmerman, & Wacker, 1963). Future work should investigate the association between AD and liver/kidney diseases, and ML studies could include the use of multiple classes, prediction windows, wider definition of AD, and more datasets, to include features that cannot be detected by GP or hospital datasets alone (eg. education level).

**Conclusion**

To conclude, this study shows that the use of routinely collected data and machine learning is an effective tool to predict AD, despite the limitations posed by an internally focused study. Additionally, it has the advantage of using information that is comparatively cheap and non-invasive to collect, as shown by the large number of important predictors collected by different blood tests (Table 5). Between random forest, decision trees, naïve Bayes, SVM, generalised linear and logistic models, random forest performed best with 0.873 accuracy, 0.860 sensitivity, 0.887 specificity and 0.964 AUC (bal\_rfmodel3A), followed by decision trees with 0.797 accuracy, 0.643 sensitivity, 0.951 specificity and 0.887 AUC, while other models performed significantly poorer. Furthermore, the random forest model suggested that liver and kidney functions were crucial to an accurate and early diagnosis of AD, with ALT/SGPT, bilirubin, GFR and serum urea as the most important predictors. Liver correlated proteins like ALT/SGPT and bilirubin could potentially be precursor biomarkers to known AD biomarkers like Amyloid-β, as they are correlated with conditions (eg. diabetes) that lead to AD. Whereas GFR and serum urea further supports that kidney functioning is related to AD. Other features mostly consist of blood-related counts, features that are related to heart diseases (cholesterols and triglycerides), and features that also correlate with liver and kidney diseases.

**References**

Abdulrahman, G. O. (2014). Alzheimer’s disease: Current trends in Wales. *Oman Medical Journal*, *29*(4), 280–284. https://doi.org/10.5001/omj.2014.73

Ahmadi-Abhari, S., Guzman-Castillo, M., Bandosz, P., Shipley, M. J., Muniz-Terrera, G., Singh-Manoux, A., … Brunner, E. J. (2017). Temporal trend in dementia incidence since 2002 and projections for prevalence in England and Wales to 2040: Modelling study. *BMJ (Online)*, *358*(July), 1–15. https://doi.org/10.1136/bmj.j2856

Ahmed, M. R., Zhang, Y., Feng, Z., Lo, B., Inan, O. T., & Liao, H. (2019). Neuroimaging and machine learning for dementia diagnosis: Recent advancements and future prospects. *IEEE Reviews in Biomedical Engineering*, *12*, 19–33. https://doi.org/10.1109/RBME.2018.2886237

Alzheimer Association. (2018). 2018 Alzheimer’s disease facts and figures. In *Alzheimer’s and Dementia* (Vol. 14). https://doi.org/10.1016/j.jalz.2018.02.001

Amador, E., Zimmerman, T. S., & Wacker, W. E. C. (1963). Urinary alkaline phosphatase activity: I. Elevated urinary LDH and alkaline phosphatase activities for the diagnosis of renal adenocarcinomas. In *JAMA* (Vol. 185). https://doi.org/10.1001/jama.1963.03060100049016

Aschwanden, D., Aichele, S., Ghisletta, P., Terracciano, A., Kliegel, M., Sutin, A., … Allemand, M. (2020). Predicting cognitive impairment and dementia: A machine learning approach. *Journal of Alzheimer’s Disease*, *75*(3), 717–728. https://doi.org/10.3233/JAD-190967

Bamanikar, S., Bamanikar, A., & Arora, A. (2016). Study of serum urea and creatinine in diabetic and non-diabetic patients in a tertiary teaching hospital. *The Journal of Medical Research*, *2*, 12–15. Retrieved from www.medicinearticle.com

Bassendine, M. F., Taylor-Robinson, S. D., Fertleman, M., Khan, M., & Neely, D. (2020). Is Alzheimer’s Disease a liver disease of the brain? *Journal of Alzheimer’s Disease*, *75*(1), 1–14. https://doi.org/10.3233/jad-190848

Bin-Hezam, R., & Ward, T. E. (2019). A machine learning approach towards detecting dementia based on its modifiable risk factors. *International Journal of Advanced Computer Science and Applications*, *10*, 148–158. https://doi.org/10.14569/ijacsa.2019.0100820

Boustani, M., Perkins, A. J., Khandker, R. K., Duong, S., Dexter, P. R., Lipton, R., … Monahan, P. (2020). Passive digital signature for early identification of Alzheimer’s disease and related dementia. *Journal of the American Geriatrics Society*, *68*(3), 511–518. https://doi.org/10.1111/jgs.16218

Breiman, L. (2001). Random Forests. *Machine Learning*, *45*, 5–32. https://doi.org/https://doi.org/10.1023/A:1010933404324

British National Formulary. (2021). Dementia. Retrieved from National Institute for Health and Care Excellence website: https://bnf.nice.org.uk/treatment-summary/dementia.html

Citron, M. (2010). Alzheimer’s disease: strategies for disease modification. *Nature Reviews Drug Discovery*, *9*(5), 387–398. https://doi.org/10.1038/nrd2896

Cohen, J. A., & Kaplan, M. M. (1979). The SGOT/SGPT ratio—An indicator of alcoholic liver disease. *Digestive Diseases and Sciences*, *24*(11), 835–838. https://doi.org/10.1007/BF01324898

Fletcher, J., & Sampson, S. (2020). All about blood tests. *MedicalNewsToday*. Retrieved from https://www.medicalnewstoday.com/articles/blood-tests

Ford, E., Rooney, P., Oliver, S., Hoile, R., Hurley, P., Banerjee, S., … Cassell, J. (2019). Identifying undetected dementia in UK primary care patients: A retrospective case-control study comparing machine-learning and standard epidemiological approaches. *BMC Medical Informatics and Decision Making*, *19*(248). https://doi.org/10.1186/s12911-019-0991-9

Glassock, R. J., & Winearls, C. (2010). Diagnosing chronic kidney disease. *Current Opinion in Nephrology and Hypertension*, *19*(2). Retrieved from https://journals.lww.com/co-nephrolhypertens/Fulltext/2010/03000/Diagnosing\_chronic\_kidney\_disease.2.aspx

Graham, W. V., Bonito-Oliva, A., & Sakmar, T. P. (2017). Update on Alzheimer’s Disease Therapy and Prevention Strategies. *Annual Review of Medicine*, *68*, 413–430. https://doi.org/10.1146/annurev-med-042915-103753

Iluz-Freundlich, D., Grubert Van Iderstine, M., Uhanova, J., Zhang, M., Knowles, C., & Minuk, G. Y. (2021). Low serum alkaline phosphatase levels in patients with chronic liver diseases: Possible contributions to disease pathogenesis. *Clinics and Research in Hepatology and Gastroenterology*, *45*(4), 101694. https://doi.org/10.1016/j.clinre.2021.101694

Joe, E., & Ringman, J. M. (2019). Cognitive symptoms of Alzheimer’s disease: Clinical management and prevention. *The BMJ*, *367*, 1–14. https://doi.org/10.1136/bmj.l6217

Kim, D. G., Krenz, A., Toussaint, L. E., Maurer, K. J., Robinson, S. A., Yan, A., … Bynoe, M. S. (2016). Non-alcoholic fatty liver disease induces signs of Alzheimer’s disease (AD) in wild-type mice and accelerates pathological signs of AD in an AD model. *Journal of Neuroinflammation*, *13*(1), 1–18. https://doi.org/10.1186/s12974-015-0467-5

Kim, W. R., Flamm, S. L., Di Bisceglie, A. M., & Bodenheimer, H. C. (2008). Serum activity of alanine aminotransferase (ALT) as an indicator of health and disease. *Hepatology*, *47*, 1363–1370. https://doi.org/10.1002/hep.22109

Kwak, M. S., Kim, D., Chung, G. E., Kang, S. J., Park, M. J., Kim, Y. J., … Lee, H. S. (2012). Serum bilirubin levels are inversely associated with nonalcoholic fatty liver disease. *Clinical and Molecular Hepatology*, *18*(4), 383–390. https://doi.org/10.3350/cmh.2012.18.4.383

Lanctôt, K. L., Amatniek, J., Ancoli-Israel, S., Arnold, S. E., Ballard, C., Cohen-Mansfield, J., … Boot, B. (2017). Neuropsychiatric signs and symptoms of Alzheimer’s disease: New treatment paradigms. *Alzheimer’s and Dementia: Translational Research and Clinical Interventions*, *3*(3), 440–449. https://doi.org/10.1016/j.trci.2017.07.001

Laske, C., Sohrabi, H. R., Frost, S. M., López-De-Ipiña, K., Garrard, P., Buscema, M., … O’bryant, S. E. (2015). Innovative diagnostic tools for early detection of Alzheimer’s disease. *Alzheimer’s and Dementia*, *11*(5), 561–578. https://doi.org/10.1016/j.jalz.2014.06.004

Levey, A. S., Bosch, J. P., Lewis, J. B., Greene, T., Rogers, N., & Roth, D. (1999). A more accurate method to estimate glomerular filtration rate from serum creatinine: a new prediction equation. *Annals of Internal Medicine*, *130*(6), 461–470. Retrieved from http://www.ncbi.nlm.nih.gov/pubmed/10075613

Liu, Y., Zhong, X., Shen, J., Jiao, L., Tong, J., Zhao, W., … Wei, M. (2020). Elevated serum TC and LDL-C levels in Alzheimer’s disease and mild cognitive impairment: A meta-analysis study. *Brain Research*, *1727*(July 2019). https://doi.org/10.1016/j.brainres.2019.146554

Livingston, G., Huntley, J., Sommerlad, A., Ames, D., Ballard, C., Banerjee, S., … Mukadam, N. (2020). Dementia prevention, intervention, and care: 2020 report of the Lancet Commission. *The Lancet*, *396*, 413–446. https://doi.org/10.1016/S0140-6736(20)30367-6

Ma, Y. C., Zuo, L., Chen, J. H., Luo, Q., Yu, X. Q., Li, Y., … Wang, H. Y. (2006). Modified glomerular filtration rate estimating equation for Chinese patients with chronic kidney disease. *Journal of the American Society of Nephrology*, *17*(10), 2937–2944. https://doi.org/10.1681/ASN.2006040368

Mendiola-Precoma, J., Berumen, L. C., Padilla, K., & Garcia-Alcocer, G. (2016). Therapies for prevention and treatment of Alzheimer’s disease. *BioMed Research International*, (2). https://doi.org/10.1155/2016/2589276

Miled, Z. Ben, Haas, K., Black, C. M., Khandker, R. K., Chandrasekaran, V., Lipton, R., & Boustani, M. A. (2020). Predicting dementia with routine care EMR data. *Artificial Intelligence in Medicine*, *102*(101771). https://doi.org/10.1016/j.artmed.2019.101771

National Kidney Foundation. (2014). Frequently asked questions about GFR estimates. In *kidney.org*.

Nori, V. S., Hane, C. A., Crown, W. H., Au, R., Burke, W. J., Sanghavi, D. M., & Bleicher, P. (2019). Machine learning models to predict onset of dementia: A label learning approach. *Alzheimer’s and Dementia: Translational Research and Clinical Interventions*, *5*, 918–925. https://doi.org/10.1016/j.trci.2019.10.006

Park, J. H., Cho, H. E., Kim, J. H., Wall, M. M., Stern, Y., Lim, H., … Cha, J. (2020). Machine learning prediction of incidence of Alzheimer’s disease using large-scale administrative health data. *Npj Digital Medicine*, *3*(1), 1–29. https://doi.org/10.1038/s41746-020-0256-0

Pekkala, T., Hall, A., Lötjönen, J., Mattila, J., Soininen, H., Ngandu, T., … Solomon, A. (2017). Development of a late-life dementia prediction index with supervised machine learning in the population-based CAIDE study. *Journal of Alzheimer’s Disease*, *55*, 1055–1067. https://doi.org/10.3233/JAD-160560

Pupim, L. B., Martin, C. J., & Ikizler, T. A. (2013). Chapter 10 - Assessment of Protein and Energy Nutritional Status. In J. D. Kopple, S. G. Massry, & K. B. T.-N. M. of R. D. Kalantar-Zadeh (Eds.), *Nutritional Management of Renal Disease (Third Edition)* (pp. 137–158). https://doi.org/https://doi.org/10.1016/B978-0-12-391934-2.00010-2

Reddy, V. H. (2014). Automatic red blood cell and white blood cell counting for telemedicine system. *International Journal of Research in Advent Technology*, *2*(1), 2321–9637.

Saif, N., Yan, P., Niotis, K., Scheyer, O., Rahman, A., Berkowitz, M., … Isaacson, R. S. (2020). Feasibility of using a wearable biosensor device in patients at risk for Alzheimer’s disease dementia. *The Journal of Prevention of Alzheimer’s Disease*, *7*(2), 104–111. Retrieved from http://www.embase.com/search/results?subaction=viewrecord&from=export&id=L2004546734%0Ahttp://dx.doi.org/10.14283/jpad.2019.39

Sáiz-Vazquez, O., Puente-Martinez, A., Ubillos-Landa, S., Pacheco-Bonrostro, J., & Santabárbara, J. (2020). Cholesterol and Alzheimer’s Disease risk: A meta-meta-analysis. *Brain Sciences*, *10*, 386.

Schnier, C., Wilkinson, T., Akbari, A., Orton, C., Sleegers, K., Gallacher, J., … Sudlow, C. (2020). The Secure Anonymised Information Linkage databank Dementia e-cohort (SAIL-DeC). *International Journal of Population Data Science*, *5*.

Srivastava, A. R., Kumar, S., Agarwal, G. G., & Ranjan, P. (2007). Blunt abdominal injury: Serum ALT-A marker of liver injury and a guide to assessment of its severity. *Injury*, *38*, 1069–1074. https://doi.org/10.1016/j.injury.2007.04.019

Stanciu, G. D., Ababei, D. C., Bild, V., Bild, W., Paduraru, L., Gutu, M. M., & Tamba, B. I. (2020). Renal contributions in the pathophysiology and neuropathological substrates shared by chronic kidney disease and alzheimer’s disease. *Brain Sciences*, *10*(8), 1–20. https://doi.org/10.3390/brainsci10080563

Stephan, B. C. M., Kurth, T., Matthews, F. E., Brayne, C., & Dufouil, C. (2010). Dementia risk prediction in the population: Are screening models accurate? *Nature Reviews Neurology*, *6*, 318–326. https://doi.org/10.1038/nrneurol.2010.54

Stocks, S. J., Kontopantelis, E., Webb, R. T., Avery, A. J., Burns, A., & Ashcroft, D. M. (2017). Antipsychotic prescribing to patients diagnosed with dementia without a diagnosis of psychosis in the context of national guidance and drug safety warnings: longitudinal study in UK general practice. *Drug Safety*, *40*, 679–692. https://doi.org/10.1007/s40264-017-0538-x

Vickers, J. C., Dickson, T. C., A. Adlard, P., Saunders, H. L., King, C. E., & McCormack, G. (2000). The cause of neuronal degeneration in Alzheimer’s disease. *Progress in Neurobiology*, *60*(2), 139–165. https://doi.org/10.1016/S0301-0082(99)00023-4

WHO. (2020). Dementia. Retrieved from World Health Organization website: https://www.who.int/news-room/fact-sheets/detail/dementia

Yang, X., Li, Q., Wu, Y., Bian, J., Lyu, T., Guo, Y., … Maraganore, D. (2020). Early prediction of Alzheimer’s disease and related dementias using electronic health records. *MedRxiv*. https://doi.org/10.1101/2020.06.13.20130401

Zhang, C. Y., He, F. F., Su, H., Zhang, C., & Meng, X. F. (2020). Association between chronic kidney disease and Alzheimer’s disease: an update. *Metabolic Brain Disease*, *35*(6), 883–894. https://doi.org/10.1007/s11011-020-00561-y

**Appendix 1 – Dementia read codes used to remove patients from the controls**

|  |  |
| --- | --- |
| Read codes | Descriptions |
| 3AE.. | Global deterioration scale: assessment of prim deg dementia |
| 9hD.. | Exception reporting: dementia quality indicators |
| 9hD1. | Excepted from dementia quality indicators: Informed dissent |
| E0011 | Presenile dementia with delirium |
| E0012 | Presenile dementia with paranoia |
| E0021 | Senile dementia with depression |
| E002z | Senile dementia with depressive or paranoid features NOS |
| E0041 | Arteriosclerotic dementia with delirium |
| E0043 | Arteriosclerotic dementia with depression |
| E012. | Other alcoholic dementia |
| E041. | Dementia in conditions EC |
| Eu00. | [X]Dementia in Alzheimer's disease |
| Eu00z | [X]Dementia in Alzheimer's disease, unspecified |
| 1281 | FH: Senile dementia |
| 8Hla. | Referral to dementia care advisor |
| 9Ou2. | Dementia monitoring second letter |
| E001. | Presenile dementia |
| E002. | Senile dementia with depressive or paranoid features |
| E0020 | Senile dementia with paranoia |
| E0042 | Arteriosclerotic dementia with paranoia |
| E004z | Arteriosclerotic dementia NOS |
| E02y1 | Drug-induced dementia |
| Eu010 | [X]Vascular dementia of acute onset |
| Eu013 | [X]Mixed cortical and subcortical vascular dementia |
| Eu022 | [X]Dementia in Huntington's disease |
| Eu02z | [X] Unspecified dementia |
| 3AD.. | Dementia test |
| 3AD1. | Ten item dementia test |
| 66h.. | Dementia monitoring |
| 6AB.. | Dementia annual review |
| 9Ou1. | Dementia monitoring first letter |
| 9Ou3. | Dementia monitoring third letter |
| E0013 | Presenile dementia with depression |
| E001z | Presenile dementia NOS |
| E003. | Senile dementia with delirium |
| Eu000 | [X]Dementia in Alzheimer's disease with early onset |
| Eu002 | [X]Dementia in Alzheimer's dis, atypical or mixed type |
| Eu012 | [X]Subcortical vascular dementia |
| Eu01y | [X]Other vascular dementia |
| Eu023 | [X]Dementia in Parkinson's disease |
| Eu040 | [X]Delirium not superimposed on dementia, so described |
| 1461 | H/O: dementia |
| 14Od. | At risk of dementia |
| 1JA2. | Suspected dementia |
| 918y. | Carer of person with dementia |
| 9Ou5. | Dementia monitoring telephone invite |
| 9hD0. | Excepted from dementia quality indicators: Patient unsuitabl |
| E000. | Uncomplicated senile dementia |
| E0010 | Uncomplicated presenile dementia |
| E004. | Arteriosclerotic dementia |
| Eu01. | [X]Vascular dementia |
| Eu011 | [X]Multi-infarct dementia |
| Eu01z | [X]Vascular dementia, unspecified |
| Eu02. | [X]Dementia in other diseases classified elsewhere |
| Eu024 | [X]Dementia in human immunodef virus [HIV] disease |
| Eu02y | [X]Dementia in other specified diseases classif elsewhere |
| Eu041 | [X]Delirium superimposed on dementia |
| 3AD2. | Thirty seven item dementia test |
| 9Ou.. | Dementia monitoring administration |
| 9Ou4. | Dementia monitoring verbal invite |
| E0040 | Uncomplicated arteriosclerotic dementia |
| Eu001 | [X]Dementia in Alzheimer's disease with late onset |
| Eu020 | [X]Dementia in Pick's disease |
| Eu021 | [X]Dementia in Creutzfeldt-Jakob disease |
| Eu025 | [X]Lewy body dementia |

**Appendix 2 – Top 50 most commonly found Read codes in AD patients five years prior to diagnosis**

|  |  |
| --- | --- |
| Read codes | Descriptions |
| 42K.. | Eosinophil count |
| 246.. | O/E - blood pressure reading |
| a6b1. | OMEPRAZOLE 20mg e/c capsules |
| 426.. | Red blood cell (RBC) count |
| bu2c. | ASPIRIN 75mg soluble tablets |
| 4.... | Laboratory procedures |
| bu23. | ASPIRIN 75mg dispersible tablets |
| 44M4. | Serum albumin |
| b211. | BENDROFLUMETHIAZIDE 2.5mg tablets |
| bd35. | ATENOLOL 50mg tablets |
| 42P.. | Platelet count |
| 424.. | Full blood count - FBC |
| b312. | FUROSEMIDE 40mg tablets |
| 44P5. | Serum HDL cholesterol level |
| 44P.. | Serum cholesterol |
| 44I4. | Serum potassium |
| 428.. | Mean corpusc. haemoglobin(MCH) |
| 44J9. | Serum urea level |
| 1371 | Never smoked tobacco |
| 44J3. | Serum creatinine |
| 42H.. | Total white cell count |
| 423.. | Haemoglobin estimation |
| 9N11. | Seen in GP's surgery |
| a6c3. | LANSOPRAZOLE 15mg capsules |
| 42J.. | Neutrophil count |
| bxd2. | SIMVASTATIN 20mg tablets |
| 451E. | GFR calculated abbreviated MDRD |
| 44G3. | ALT/SGPT serum level |
| 42QE. | International normalised ratio |
| blb1. | AMLODIPINE 5mg tablets |
| 44M3. | Serum total protein |
| f41y. | METFORMIN HYDROCHLORIDE 500mg tablets |
| f922. | LEVOTHYROXINE SODIUM 50micrograms tablets |
| 44I5. | Serum sodium |
| 44P6. | Serum LDL cholesterol level |
| 44F.. | Serum alkaline phosphatase |
| 42M.. | Lymphocyte count |
| bxd5. | SIMVASTATIN 40mg tablets |
| 22K.. | Body Mass Index |
| 22A.. | O/E - weight |
| 42A.. | Mean corpuscular volume (MCV) |
| dia2. | CO-CODAMOL 8mg/500mg tablets |
| 44Q.. | Serum triglycerides |
| 65E.. | Influenza vaccination |
| f923. | LEVOTHYROXINE SODIUM 100micrograms tablets |
| 42N.. | Monocyte count |
| 44E.. | Serum bilirubin level |
| 42L.. | Basophil count |
| di21. | PARACETAMOL 500mg tablets |
| 44M5. | Serum globulin |