An investigation of the predictors of thyroid cancer in patients with thyroid nodules

Ovie Edafe

Neil Shephard

Karen Sisley

Sabapathy P Balasubramanian

2024-04-26

Abstract

An abstract summarising the work undertaken and the overall conclusions can be placed here. Sub-headings are currently removed because they conflict with those in the body of the text and mess up the links in the Table of Contents.

Source: [Article Notebook](https://ns-rse.github.io/sheffield-thyroid/index-preview.html)

## 1 Introduction

Thyroid nodules are common. The challenge in the management of thyroid nodules is differentiating between benign and malignant nodule thyroid nodules.The use fine needle aspiration and cytology (FNAC) still leaves around 20% of patients that cannot be clearly classified as either benign or malignant. This scenario traditionally leads to diagnostic hemithyroidectomy for definitive histology. Other clinical variables such as patients’ demographics, clinical and biochemical factors have been shown to be associated with thyroid cancer in patients with thyroid nodules. This has been utilised in studies evaluating predictors of thyroid cancer with a view of creating a model to aid prediction. Standard practice on the management of thyroid nodules does not utilise these non ultrasound and non cytological factors. Combination of these variables considered to be significant with ultrasound and cytological characteristics may improve management of patients with thyroid nodules. Thyroid nodules are increasingly being incidentally detected with increased use of imaging in the evaluation of non thyroid related pathologies. Thus, leading to increase investigation of thyroid nodules and subsequent increased number of thyroid operations in non diagnostic cases. There are morbidities associated with thyroid surgery including scar, recurrent laryngeal nerve injury, hypothyroidism and hypoparathyroidism. We performed a systematic review to evaluate for predictors of thyroid cancer specifically in patients presenting with thyroid nodules. The systematic review a number of potential important variables that may be useful in the prediction of thyroid cancer in patients with thyroid nodules. The aim of this study was to evaluate the predictors of thyroid cancer with a view of improving prediction of thyroid cancer using computer age statistical inference techniques (Efron and Hastie (2016)).

## 2 Methods

This study was reported as per the Strengthening the Reporting of Observational Studies in Epidemiology (STROBE) guidelines

### 2.1 Study design

This was a retrospective cohort study.

### 2.2 Setting

The study was conducted at the Sheffield Teaching hospitals NHS Foundation Trusts. This is a tertiary referral centre for the management of thyroid cancer

### 2.3 Participants

We included all consecutive patients who presented with thyroid nodule(s) or that were found to have thyroid nodule(s) on ultrasound done for thyroid pathology or for other non thyroid related pathologies

### 2.4 Variables

Variable evaluated was based on findings from a systematic review evaluating predictors of thyroid cancer in patients with thyroid nodules. Data on the following variables were collected: patient demographics (age, gender, ethnicity), nodule presentation (incidental nodule, palpable nodule, rapid enlargement, compressive symptoms, vocal paresis), past medical history (hypertension, Graves’ disease, Hashimotos’ thyroiditis, family history of thyroid cancer, exposure to neck radiation), biochemistry (thyroid stimulating hormone, lymphocytes, monocytes), ultrasound characteristics ([British Thyroid Association ultrasound (BTA U)](https://radiopaedia.org/articles/bta-ultrasound-u-classification-of-thyroid-nodules)), nodule size, solitary nodule, nodule consistency, cervical lymphadenopathy), Royal College of Pathology (RCP) FNAC classification, type of thyroid surgery, and histological diagnosis.

### 2.5 Data source

Data was collected from patients’ case notes and electronic patients’ database using a standardised data collection proforma. This was initially piloted on 30 patients and revised to improve data entry. In addition a number of variables that were not standard collected during workout of patients were not further checked; these include body mass index (BMI), serum thyroglobulin, serum triiodothyronine (T3), thyroxine (T4), thyroglobulin antibody (TgAb), thyroid peroxidase antibody (TP0Ab), and urinary iodine.

### 2.6 Study size

We sought to have a large data set of at least 100 thyroid nodules with a cancer diagnosis using consecutive sampling technique. We aimed for a total of 1500 patients with thyroid nodules to achieve our target sample size. With the use of modern statistical techniques, we proposed such number will be appropriate to detect important variables if it exists.

### 2.7 Data analysis

Data was cleaned and analysed using the R Statistical Software R Core Team (2023) and the Tidyverse (Wickham et al. (2019)), Tidymodels (Kuhn and Wickham (2020)) collection of packages.

### 2.8 Imputation

The dataset is incomplete and there are missing observations across all variables to varying degrees. In order to maximise the sample available for analysis imputation was used to infer missing values. The Multivariat Imputation via Chained Equations ([MICE](https://amices.org/mice/) and implemented in the eponymous R package Buuren and Groothuis-Oudshoorn (2011)) was employed which assumes data is missing at random (a difficult assumption to formally test). The approach takes each variable with missing data and attempts to predict it using statistical modelling based on the observed values. In essence it is the same approach as the statistical methods being employed to try and predict Thyroid Cancer and there are a range of statistical techniques available which include

### 2.9 Modelling

We used a selection of statistic modelling techniques to evaluate association between variables and thyroid cancer in patients with thyroid nodules. The patient population was split into training and testing cohorts in a ratio of 0.75:0.25 and each model is fitted using the training cohort. This split ratio is generally used in traditional machine learning techniques. The training set of the data was used to estimate the relation between variables and thyroid cancer. The larger the training data, the better it is for the model to learn the trends. The test set was used to determine the accuracy of the model in predicting thyroid cancer; the bigger the test data the more confidence we have in the model prognostic values. We used simple randomisation technique for the split to prevent bias in the data split. We ensured that there was no duplicate in the data sets so any test data was not accidentally trained. Furthermore, cross validation was used to estimate the accuracy of the various machine learning models. The k-fold techniques splits the data in ?10 folds, and the data was trained on all but one of the the fold, and the one fold not trained is used to test the data. This was repeated multiple times using a different fold for test and the others for training until all the folds is utilised for training and testing. Following multiple training process with k-fold, we selected the model that has the best predictive value for thyroid cancer in the test cohort. We also used the leave one out (loo) cross-validation to train and test the data set.In this technique, all but one observation is use to train the data set and one observation is use to test the data; this is repeated until all the data test is used for testing and training. The model with the best predictive value was selected.

#### 2.9.1 LASSO / Elastic Net

LASSO (Least Absolute Shrinkage and Selection Operatror) and Elastic Net Zou and Hastie (2005) are regression methods that perform variable selection. The original LASSO method proposed by “Regression Shrinkage and Selection via the Lasso” (1996) allows the coefficients for independent/predictor variables to “shrink” down towards zero, effectively eliminating them from influencing the model, this is often referred to as L1 regularisation. The Elastic Net Zou and Hastie (2005) improves on the LASSO by balancing L1 regularisation with ridge-regression or L2 regularisation which helps avoid over-fitting.

Both methods avoid many of the shortcomings/pitfalls of stepwise variable selection Thompson (1995) Smith (2018) and have been shown to be more accurate in clinical decision making in small datasets with well code, externally selected variables Steyerberg et al. (2001)

#### 2.9.2 Random Forest

To add reference The random forest plot is an extension of the decision tree methodology to reduce variance. Decision trees are very sensitive to the training data set and can lead to high variance; thus potential issues with generalisation of the model. The random forest plot selects random observation of the dataset to create multiple decision trees. Random variables are selected for each tree in the training of the data set. The aggregated output of the generated decision trees is then used to create an estimate.

#### 2.9.3 Gradient Boosting

Gradient boosting is a machine learning algorithm that uses decision tree as a base model. The data is initially trained on this decision tree, but the initial prediction is weak, thus termed a weak based model. In gradient boosting the process is iterative; a sequence of decision trees is added to the initial tree. Each tree learns from the prior tree(s) to improve the model, increasing strength and minimising error.

#### 2.9.4 SVM

Support Vector Machines is an approach that allows observation with a binary classifications to be separated using a hyperplane. It finds a hyperplane that best stratify the two classes i.e benign versus malignant nodules. SVM finds the hyperplane with the maximum margin of separation between the two classes. The support vectors are the data point that are positioned close to the margin of the hyperplane and these used to select the most appropraite hyperplane. The support vectors are the only data points that have an influence on the maximum margin in SVM.

#### 2.9.5 Comparision

## 3 Results

Source: [Article Notebook](https://ns-rse.github.io/sheffield-thyroid/index-preview.html)

### 3.1 Data Description

A summary of the variables that are available in this data set can be found in [Table 15](#tbl-variables).

## Demographics

## Clinical Characteristics

## Biomarkers

## Ultrasound

## BTA U

## Thyroid Classification

## Cytology

[Table 1](#tbl-patient-demographics) shows the demographics of patients included in this study. A total of 1364 patients were included in this study with a median (IQR) age of 55 ( 41-69). [Table 2](#tbl-clinical-characteristics) shows the distribution of clinical variables evaluated between benign and malignant thyroid nodules.

#### 3.1.1 Missing Data

The completeness of the original data is shown in tables [Table 11](#tbl-imputation-summary-pmm), [Table 12](#tbl-imputation-summary-cart), [Table 13](#tbl-imputation-summary-rf), along with summaries from four rounds of imputation for each of three imputation methods. Where variables continuous (e.g. age or size\_nodule\_mm) basic summary statistics in the form of mean, standard deviation, median and inter-quartile range are given. For categorical variables that are logical TRUE/FALSE (e.g. palpable\_nodule) the number of TRUE observations and the percentage (of those with observed data for that variable) are shown along with the number that are *Unknown*. For categorical variables such as gender percentages in each category are reported. For all variables an indication of the number of missing observations is also given and it is worth noting that there are 214 instances where the final\_pathology is not known which reduces the sample size to 1150.

More detailed tabulations of missing data by variable are shown in [Table 9](#tbl-naniar-miss-var-summary) which shows the number and percentage of missing data for each variable and by case in [Table 10](#tbl-naniar-miss-case-table) which shows how much missing data each case has. A visualisation of this is shown in [Figure 1](#fig-visdat-vis-missing) .

**NB** - Currently there is a [bug in the stable release of Quarto](https://github.com/quarto-dev/quarto-cli/issues/10196) which prevents rendering of the missing data figures. It is fixed in development version [v1.6.1](https://github.com/quarto-dev/quarto-cli/releases/tag/v1.6.1) (currently available as pre-release, so if things don’t render upgrade).

## Variables

|  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| Table 9: Summary of missing data by variable.  Summary of missing data by variable.   | Variable | N | % | | --- | --- | --- | | albumin | 515 | 44.8 | | tsh\_value | 413 | 35.9 | | monocyte | 363 | 31.6 | | lymphocytes | 359 | 31.2 | | size\_nodule\_mm | 319 | 27.7 | | family\_history\_thyroid\_cancer | 281 | 24.4 | | hypertension | 126 | 11.0 | | compressive\_symptoms | 106 | 9.22 | | vocal\_cord\_paresis | 76 | 6.61 | | hashimotos\_thyroiditis | 73 | 6.35 | | graves\_disease | 67 | 5.83 | | palpable\_nodule | 58 | 5.04 | | bta\_u\_classification | 50 | 4.35 | | ethnicity | 48 | 4.17 | | rapid\_enlargement | 43 | 3.74 | | cervical\_lymphadenopathy | 9 | 0.783 | | solitary\_nodule | 8 | 0.696 | | incidental\_nodule | 5 | 0.435 | | age\_at\_scan | 0 | 0 | | gender | 0 | 0 | | exposure\_radiation | 0 | 0 | | thy\_classification | 0 | 0 | | final\_pathology | 0 | 0 | |

Source: [Article Notebook](https://ns-rse.github.io/sheffield-thyroid/index-preview.html)

## Observations

|  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| Table 10: Summary of missing data by case, how much missing data is there per person?  Summary of missing data by case, how much missing data is there per person?   | Missing Variables | N | % | | --- | --- | --- | | 0 | 236 | 20.522 | | 1 | 262 | 22.783 | | 2 | 171 | 14.870 | | 3 | 155 | 13.478 | | 4 | 121 | 10.522 | | 5 | 84 | 7.304 | | 6 | 43 | 3.739 | | 7 | 18 | 1.565 | | 8 | 22 | 1.913 | | 9 | 15 | 1.304 | | 10 | 12 | 1.043 | | 11 | 5 | 0.435 | | 12 | 3 | 0.261 | | 13 | 2 | 0.174 | | 14 | 1 | 0.087 | |

Source: [Article Notebook](https://ns-rse.github.io/sheffield-thyroid/index-preview.html)

|  |
| --- |
| Figure 1 |

Source: [Article Notebook](https://ns-rse.github.io/sheffield-thyroid/index-preview.html)

The [MICE](https://amices.org/mice/) package also provides tools for visualising missing data and these are shown in figures [Figure 2](#fig-mice-vis-missing-clinical), **?@fig-mice-vis-missing-biomarker** and [Figure 4](#fig-mice-vis-missing-ultrasound).

The columns of these plots, labelled along the top, show the variable, if a cell is blue it indicates data is present, if it is red it indicates there is missing data. The left-hand side shows the total number of observations for that rows particular combination of variables with number of missing variables indicated on the right. The first row shows that for these variables there are 604 observations with zero missing data across the listed variables, the second row indicates there are 166 observations with *just* family\_history\_thyroid\_cancer but there are some with this missing *and other variables*. The numbers on the bottom of the figure indicate the total number of missing observations for that variable (e.g. for family\_history\_thyroid\_cancer there is a total of 281 missing observations).

**TODO** - Workout why out-width: "80%" isn’t applied to these figures and/or how to make the All figure readable.

## Clinical

|  |
| --- |
| Figure 2 |

Source: [Article Notebook](https://ns-rse.github.io/sheffield-thyroid/index-preview.html)

## Biomarkers

|  |
| --- |
| Figure 3 |

Source: [Article Notebook](https://ns-rse.github.io/sheffield-thyroid/index-preview.html)

## Ultrasound

|  |
| --- |
| Figure 4 |

Source: [Article Notebook](https://ns-rse.github.io/sheffield-thyroid/index-preview.html)

## All

|  |
| --- |
| Figure 5 |

Source: [Article Notebook](https://ns-rse.github.io/sheffield-thyroid/index-preview.html)

#### 3.1.2 Imputation

The [MICE](https://amices.org/mice/) package van Buuren and Groothuis-Oudshoorn (2011) offers a number of different methods for imputing variables (see [documentation][mice\_details]) we have investigated Predictive Mean Matching (PMM), Classification and Regression Trees (CART) and Random Forests (RF). Four rounds of imputation using each method were made.

A comparison of distributions/proportions before and after imputation are presented below to allow assessment of the utility of each method.

Source: [Article Notebook](https://ns-rse.github.io/sheffield-thyroid/index-preview.html)

The distribution of observed (blue) and imputed values for continous variables are shown in the Tab set immediately below and across all variables the imputed distrubtions follow closely that of the observed indicating that the imputation methods have worked well for all three methods tested.

## Albumin

|  |
| --- |
| Figure 6 |

Source: [Article Notebook](https://ns-rse.github.io/sheffield-thyroid/index-preview.html)

## Monocyte

|  |
| --- |
| Figure 7 |

Source: [Article Notebook](https://ns-rse.github.io/sheffield-thyroid/index-preview.html)

## Lymphocytes

|  |
| --- |
| Figure 8 |

Source: [Article Notebook](https://ns-rse.github.io/sheffield-thyroid/index-preview.html)

## TSH Value

|  |
| --- |
| Figure 9 |

Source: [Article Notebook](https://ns-rse.github.io/sheffield-thyroid/index-preview.html)

## Nodule Size

|  |
| --- |
| Figure 10 |

Source: [Article Notebook](https://ns-rse.github.io/sheffield-thyroid/index-preview.html)

Similarly for the discrete variables the distribution of proportions in the original and each imputed dataset are shown below. The observed (in pink/peach) always have slightly lower proportions of the observed values because of the presence of missing (evidences by only one group being present in the NA column) but as with the continuous variables across all imputation methods the proportions are roughly as expected again indicating that imputation has worked well.

**TODO** - Extract the legends from individual plots and add them to the end of each row, see the [cowplot shared legends article](https://wilkelab.org/cowplot/articles/shared_legends.html) for pointers on how to do this. Should ideally also get the fill colours to align with those used by ggmice.

## Incidental Nodule

|  |
| --- |
| Figure 11 |

Source: [Article Notebook](https://ns-rse.github.io/sheffield-thyroid/index-preview.html)

## Palpable Nodule

|  |
| --- |
| Figure 12 |

Source: [Article Notebook](https://ns-rse.github.io/sheffield-thyroid/index-preview.html)

## Rapid Enlargement

|  |
| --- |
| Figure 13 |

Source: [Article Notebook](https://ns-rse.github.io/sheffield-thyroid/index-preview.html)

## Compressive Symptoms

|  |
| --- |
| Figure 14 |

Source: [Article Notebook](https://ns-rse.github.io/sheffield-thyroid/index-preview.html)

## Hypertension

|  |
| --- |
| Figure 15 |

Source: [Article Notebook](https://ns-rse.github.io/sheffield-thyroid/index-preview.html)

## Vocal Cord Paresis

|  |
| --- |
| Figure 16 |

Source: [Article Notebook](https://ns-rse.github.io/sheffield-thyroid/index-preview.html)

## Graves Disease

|  |
| --- |
| Figure 17 |

Source: [Article Notebook](https://ns-rse.github.io/sheffield-thyroid/index-preview.html)

## Hashimotos Thyroiditis

|  |
| --- |
| Figure 18 |

Source: [Article Notebook](https://ns-rse.github.io/sheffield-thyroid/index-preview.html)

## Family History

|  |
| --- |
| Figure 19 |

Source: [Article Notebook](https://ns-rse.github.io/sheffield-thyroid/index-preview.html)

## Exposure Radiation

|  |
| --- |
| Figure 20 |

Source: [Article Notebook](https://ns-rse.github.io/sheffield-thyroid/index-preview.html)

## Solitary Nodule

|  |
| --- |
| Figure 21 |

Source: [Article Notebook](https://ns-rse.github.io/sheffield-thyroid/index-preview.html)

## BTA U-Classification

|  |
| --- |
| Figure 22 |

Source: [Article Notebook](https://ns-rse.github.io/sheffield-thyroid/index-preview.html)

## Cervical Lymphadenopathy

|  |
| --- |
| Figure 23 |

Source: [Article Notebook](https://ns-rse.github.io/sheffield-thyroid/index-preview.html)

## PMM

## CART

## RF

### 3.2 Modelling

**TODO** - And in light of having removed **?@tbl-data-completness** in favour of the imputed datesets this too has been removed? (@ns-rse 2024-07-11). **TODO** - This table feels like duplication of [Table 8](#tbl-data-completeness), perhaps have just one? (@ns-rse 2024-07-11).

The predictor variables selected to predict final\_pathology are shown in [Table 14](#tbl-predictors)

Section that sets up the modelling

Source: [Article Notebook](https://ns-rse.github.io/sheffield-thyroid/index-preview.html)

Source: [Article Notebook](https://ns-rse.github.io/sheffield-thyroid/index-preview.html)

Source: [Article Notebook](https://ns-rse.github.io/sheffield-thyroid/index-preview.html)

Source: [Article Notebook](https://ns-rse.github.io/sheffield-thyroid/index-preview.html)

Source: [Article Notebook](https://ns-rse.github.io/sheffield-thyroid/index-preview.html)

The following section is output from a Tidymodel approach to logistic regression to try and work out why variables are not being included.

Source: [Article Notebook](https://ns-rse.github.io/sheffield-thyroid/index-preview.html)

A total of 1150 patients had complete data for the selected predictor variables (see [Table 14](#tbl-predictors)). Because of the volume of missing data which if a saturated model were used would include only ~350 people with complete data across all co-variates imputed datasets were analysed instead.

Source: [Article Notebook](https://ns-rse.github.io/sheffield-thyroid/index-preview.html)

Source: [Article Notebook](https://ns-rse.github.io/sheffield-thyroid/index-preview.html)

Source: [Article Notebook](https://ns-rse.github.io/sheffield-thyroid/index-preview.html)

Source: [Article Notebook](https://ns-rse.github.io/sheffield-thyroid/index-preview.html)

Source: [Article Notebook](https://ns-rse.github.io/sheffield-thyroid/index-preview.html)

Source: [Article Notebook](https://ns-rse.github.io/sheffield-thyroid/index-preview.html)

Source: [Article Notebook](https://ns-rse.github.io/sheffield-thyroid/index-preview.html)

Source: [Article Notebook](https://ns-rse.github.io/sheffield-thyroid/index-preview.html)

Source: [Article Notebook](https://ns-rse.github.io/sheffield-thyroid/index-preview.html)

Source: [Article Notebook](https://ns-rse.github.io/sheffield-thyroid/index-preview.html)

Source: [Article Notebook](https://ns-rse.github.io/sheffield-thyroid/index-preview.html)

Source: [Article Notebook](https://ns-rse.github.io/sheffield-thyroid/index-preview.html)

Source: [Article Notebook](https://ns-rse.github.io/sheffield-thyroid/index-preview.html)

#### 3.2.1 Logistic Regression

#### 3.2.2 LASSO

#| label: fig-imputed-pmm-lasso  
#| purl: true  
#| eval: true  
#| echo: false  
#| output: false  
load(file = paste(r\_dir, "imputed\_cart\_model\_lasso.rds", sep = "/"))  
load(file = paste(r\_dir, "imputed\_pmm\_model\_lasso.rds", sep = "/"))  
load(file = paste(r\_dir, "imputed\_forest\_model\_lasso.rds", sep = "/"))

## PMM

#| label: fig-imputed-pmm-lasso  
#| purl: true  
#| eval: true  
#| echo: false  
#| output: true  
#| out-width: "80%"  
cowplot::plot\_grid(imputed\_pmm\_model\_lasso[[1]]$importance\_plot,  
 imputed\_pmm\_model\_lasso[[1]]$train\_roc\_curve\_plot,  
 imputed\_pmm\_model\_lasso[[1]]$test\_roc\_curve\_plot,  
 imputed\_pmm\_model\_lasso[[2]]$importance\_plot,  
 imputed\_pmm\_model\_lasso[[2]]$train\_roc\_curve\_plot,  
 imputed\_pmm\_model\_lasso[[2]]$test\_roc\_curve\_plot,  
 imputed\_pmm\_model\_lasso[[3]]$importance\_plot,  
 imputed\_pmm\_model\_lasso[[3]]$train\_roc\_curve\_plot,  
 imputed\_pmm\_model\_lasso[[3]]$test\_roc\_curve\_plot,  
 imputed\_pmm\_model\_lasso[[4]]$importance\_plot,  
 imputed\_pmm\_model\_lasso[[4]]$train\_roc\_curve\_plot,  
 imputed\_pmm\_model\_lasso[[4]]$test\_roc\_curve\_plot,  
 imputed\_pmm\_model\_lasso[[5]]$importance\_plot,  
 imputed\_pmm\_model\_lasso[[5]]$train\_roc\_curve\_plot,  
 imputed\_pmm\_model\_lasso[[5]]$test\_roc\_curve\_plot,  
 # labels = c("PMM", "Cart", "Random Forest"),  
 nrow = 5,  
 ncol = 3)

#| label: tab-imputed-pmm-lasso-importance  
#| purl: true  
#| eval: true  
#| echo: false  
#| output: true  
#| tbl-caption: "Importance of features from LASSO model of five PMM imputed datasets."  
tidy\_pmm\_lasso$importance |>  
 knitr::kable(caption = "Importance of features from LASSO model of five PMM imputed datasets.")

#| label: tab-imputed-pmm-lasso-train-metrics  
#| purl: true  
#| eval: true  
#| echo: false  
#| output: true  
#| tbl-caption: "Classification metrics from LASSO model of five PMM imputed datasets (Training Data)."  
tidy\_pmm\_lasso$train\_metrics |>  
 knitr::kable(caption = "Classification metrics from LASSO model of five PMM imputed datasets (Training Data).",  
 digits = 4)

#| label: tab-imputed-pmm-lasso-test-metrics  
#| purl: true  
#| eval: true  
#| echo: false  
#| output: true  
#| tbl-caption: "Classification metrics from LASSO model of five PMM imputed datasets (Testing Data)."  
tidy\_pmm\_lasso$test\_metrics |>  
 knitr::kable(caption = "Classification metrics from LASSO model of five PMM imputed datasets (Testing Data).",  
 digits = 4)

## CART

#| label: fig-imputed-cart-lasso  
#| purl: true  
#| eval: true  
#| echo: false  
#| output: true  
#| out-width: "80%"  
cowplot::plot\_grid(imputed\_cart\_model\_lasso[[1]]$importance\_plot,  
 imputed\_cart\_model\_lasso[[1]]$train\_roc\_curve\_plot,  
 imputed\_cart\_model\_lasso[[1]]$test\_roc\_curve\_plot,  
 imputed\_cart\_model\_lasso[[2]]$importance\_plot,  
 imputed\_cart\_model\_lasso[[2]]$train\_roc\_curve\_plot,  
 imputed\_cart\_model\_lasso[[2]]$test\_roc\_curve\_plot,  
 imputed\_cart\_model\_lasso[[3]]$importance\_plot,  
 imputed\_cart\_model\_lasso[[3]]$train\_roc\_curve\_plot,  
 imputed\_cart\_model\_lasso[[3]]$test\_roc\_curve\_plot,  
 imputed\_cart\_model\_lasso[[4]]$importance\_plot,  
 imputed\_cart\_model\_lasso[[4]]$train\_roc\_curve\_plot,  
 imputed\_cart\_model\_lasso[[4]]$test\_roc\_curve\_plot,  
 imputed\_cart\_model\_lasso[[5]]$importance\_plot,  
 imputed\_cart\_model\_lasso[[5]]$train\_roc\_curve\_plot,  
 imputed\_cart\_model\_lasso[[5]]$test\_roc\_curve\_plot,  
 # labels = c("PMM", "Cart", "Random Forest"),  
 nrow = 5,  
 ncol = 3)

## Random Forest

#| label: fig-imputed-rf-lasso  
#| purl: true  
#| eval: true  
#| echo: false  
#| output: true  
#| out-width: "80%"  
cowplot::plot\_grid(imputed\_rf\_model\_lasso[[1]]$importance\_plot,  
 imputed\_rf\_model\_lasso[[1]]$train\_roc\_curve\_plot,  
 imputed\_rf\_model\_lasso[[1]]$test\_roc\_curve\_plot,  
 imputed\_rf\_model\_lasso[[2]]$importance\_plot,  
 imputed\_rf\_model\_lasso[[2]]$train\_roc\_curve\_plot,  
 imputed\_rf\_model\_lasso[[2]]$test\_roc\_curve\_plot,  
 imputed\_rf\_model\_lasso[[3]]$importance\_plot,  
 imputed\_rf\_model\_lasso[[3]]$train\_roc\_curve\_plot,  
 imputed\_rf\_model\_lasso[[3]]$test\_roc\_curve\_plot,  
 imputed\_rf\_model\_lasso[[4]]$importance\_plot,  
 imputed\_rf\_model\_lasso[[4]]$train\_roc\_curve\_plot,  
 imputed\_rf\_model\_lasso[[4]]$test\_roc\_curve\_plot,  
 imputed\_rf\_model\_lasso[[5]]$importance\_plot,  
 imputed\_rf\_model\_lasso[[5]]$train\_roc\_curve\_plot,  
 imputed\_rf\_model\_lasso[[5]]$test\_roc\_curve\_plot,  
 # labels = c("PMM", "Rf", "Random Forest"),  
 nrow = 5,  
 ncol = 3)

#### 3.2.3 Elastic Net

#### 3.2.4 Random Forest

#### 3.2.5 Gradient Boosting

#### 3.2.6 Explainability

Which factors are important to classification can be assessed not just by the “importance” but by methods know as [LIME](https://search.r-project.org/CRAN/refmans/lime/html/lime-package.html) (Local Interpretable Model-Agnostic Explanations) Ribeiro, Singh, and Guestrin (2016) and [Shapley values](https://shap.readthedocs.io/en/latest/example_notebooks/overviews/An%20introduction%20to%20explainable%20AI%20with%20Shapley%20values.html) Lundberg and Lee (2017)

#### 3.2.7 Comparision

Comparing the sensitivity of the different models goes here.

* Table of sensitivity/specificity/other metrics.
* ROC curves

## 4 Conclusion

The take-away message is….these things are hard!

## 5 Appendix

### 5.1 Data Dictionary

Buuren, Stef van, and Karin Groothuis-Oudshoorn. 2011. “mice: Multivariate Imputation by Chained Equations in R.” *J. Stat. Soft.* 45 (December): 1–67. <https://doi.org/10.18637/jss.v045.i03>.

Efron, Bradley, and Trevor Hastie. 2016. *Computer Age Statistical Inference: Algorithms, Evidence, and Data Science*. *Cambridge Core*. Cambridge, England, UK: Cambridge University Press. <https://doi.org/10.1017/CBO9781316576533>.

Kuhn, Max, and Hadley Wickham. 2020. *Tidymodels: A Collection of Packages for Modeling and Machine Learning Using Tidyverse Principles.* <https://www.tidymodels.org>.

Lundberg, Scott M, and Su-In Lee. 2017. “A Unified Approach to Interpreting Model Predictions.” Edited by I. Guyon, U. V. Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, and R. Garnett, 4765–74. <http://papers.nips.cc/paper/7062-a-unified-approach-to-interpreting-model-predictions.pdf>.

R Core Team. 2023. *R: A Language and Environment for Statistical Computing*. Vienna, Austria: R Foundation for Statistical Computing. <https://www.R-project.org/>.

“Regression Shrinkage and Selection via the Lasso.” 1996. *Journal of the Royal Statistical Society. Series B (Methodological)*. <https://www.jstor.org/stable/2346178>.

Ribeiro, Marco Tulio, Sameer Singh, and Carlos Guestrin. 2016. “"Why Should I Trust You?": Explaining the Predictions of Any Classifier.” *arXiv*, February. <https://doi.org/10.48550/arXiv.1602.04938>.

Smith, Gary. 2018. “Step away from stepwise.” *J. Big Data* 5 (1): 1–12. <https://doi.org/10.1186/s40537-018-0143-6>.

Steyerberg, Ewout W., Marinus J. C. Eijkemans, Frank E. Harrell, and Dik. 2001. “Prognostic Modeling with Logistic Regression Analysis.” *Medical Decision Making* 21 (1): 45–56. <https://doi.org/10.1177/0272989x0102100106>.

Thompson, Bruce. 1995. “Stepwise Regression and Stepwise Discriminant Analysis Need Not Apply Here: A Guidelines Editorial.” *Educational and Psychological Measurement* 55 (4): 525–34. <https://doi.org/10.1177/0013164495055004001>.

van Buuren, Stef, and Karin Groothuis-Oudshoorn. 2011. “mice: Multivariate Imputation by Chained Equations in r.” *Journal of Statistical Software* 45 (3): 1–67. <https://doi.org/10.18637/jss.v045.i03>.

Wickham, Hadley, Mara Averick, Jennifer Bryan, Winston Chang, Lucy D’Agostino McGowan, Romain François, Garrett Grolemund, et al. 2019. “Welcome to the tidyverse.” *Journal of Open Source Software* 4 (43): 1686. <https://doi.org/10.21105/joss.01686>.

Zou, Hui, and Trevor Hastie. 2005. “Regularization and Variable Selection Via the Elastic Net.” *J. R. Stat. Soc. Ser. B. Stat. Methodol.* 67 (2): 301–20. <https://doi.org/10.1111/j.1467-9868.2005.00503.x>.