Routine Blood Tests for Mortality Prediction

Ricki Nabeshima
September 2018

This project is submitted under University of London regulations as part of the examination requirements for the MSc degree in Applied Statistics and Computational Data Analytics.

Any quotation or excerpt from the published or unpublished work of other persons is explicitly indicated and in each such instance a full reference of the source of such work is given. I have read and understood the requirements of the Birkbeck College Examinations Instructions to Candidates, including the relevant University of London regulations on Examination Tests and in accordance with those requirements submit this work as my own.

Ricki Nabeshima

_____

Date:

_____

1 Abstract

Effective risk stratification in medicine is important for efficiently allocating resources and selecting appropriate clinical interventions for patients.

Blood tests are an important and commonly-used diagnostic test, but the current approach to analysing them is not well-suited to risk stratification.

This project aims to develop a more sophisticated approach to analysing routine blood test results, with the goal of scoring lives by their short-term mortality risk. Such a model would be an effective tool for risk stratification – enabling doctors to make better decisions about when and how to treat their patients.

Data from the US population health study NHANESIII was used for the initial variable search. Cox regression was used to identify initial variables of interest and to identify which blood biomarkers had the strongest association with mortality. Some of the key variables identified were Red Blood Cell Distribution Width, Lymphocyte Percentage, Serum Creatinine, Serum Albumin, Serum Selenium, Cotinine (as a proxy for smoking habits) and C-Reactive Protein.

After the initial variable search a larger dataset using NHANESIII and records from the more contemporary Continuous NHANES was created. Hot-deck imputation was used to complete records with missing values and this dataset was used to train a random forest model to predict 5-year mortality.

ROC curves were used to assess the relative performance of models. A model using blood variables only achieved an AUC of 0.861. Adding age and sex to the model achieved an AUC of 0.885.

This model is shown to perform much better than the “reference range” approach which would achieve an AUC of 0.605 and I show that about 15% of the correctly-identified deaths from the blood-only model had all readings within normal ranges.

2 Background and motivation

Effective risk stratification is an important part of medical decision making.

From a clinical perspective, effective risk stratification enables doctors to decide on appropriate interventions for their patients. An excellent example of this is the QRisk algorithm (Hippisley-Cox, et al., 2008), which computes a probability of suffering a cardiovascular event given a person’s age, sex, blood pressure, cholesterol etc. This is now widely used by GPs to assess cardiovascular risk and its outputs are used to guide medical treatment.

Blood tests are an important and ubiquitous diagnostic tool. In the UK in 2017 alone, NHS hospitals requested 246 million biochemistry investigations and 43 million haematology investigations (NHS England, 2017). Modern automated blood analysers can process thousands of blood samples per hour to a high level of accuracy and with minimum human intervention, calculating a wealth of haematological and biochemical quantities.

However, despite the size and richness of data available the interpretation of blood test results is done in a very simple way, often by comparing individual readings against one- or two-dimensional reference ranges to identify abnormal readings.

Although this can provide useful diagnostic insights to doctors it is not a very effective means of risk stratification. I demonstrate in this project that many healthy people will have one or more readings that are

outside of normal ranges, and many unhealthy people have abnormal readings that in isolation are within normal ranges but combined together are suggestive of increased risk.

This project aims to develop a more sophisticated approach to analysing blood test results, with the goal of scoring lives by their short-term mortality risk. Such a model would be an effective tool for risk stratification – enabling doctors to make better decisions about when and how to treat their patients.

An overview of the main components of blood is presented in Table 1.

Component	Description
Plasma	The fluid part of blood which carries the blood cells around the body in suspension. The plasma also contains dissolved proteins, hormones and salts.
Red Blood Cells (erythrocytes)	Red blood cells carry the gases used for respiration around the body. They contain haemoglobin which picks up oxygen in the lungs and carries it to the tissues of the body, where it is swapped for carbon dioxide.
White Blood Cells (leukocytes)	White blood cells are the cells of the immune system and help the body to fight off infection. The three main types are: • Neutrophils – fight bacteria • Lymphocytes – produce antibodies and provide immunity • Monocytes – perform general blood cleaning tasks
Platelets (thrombocytes)	Cells that react to bleeding by clumping together and forming a blood clot.

Table 1: Summary of the main components of human blood. Source: Boyle and Senior (2008)

3 Existing research

Existing research has tended to focus on one or a few biomarkers, usually in relation to the incidence and prognosis of a particular disease. These are referenced extensively throughout section 6. Due to the scale and detail of the data collection, the NHANES studies are often used for this purpose.

Only a few studies look at the full spectrum of blood biomarkers for mortality prediction. One recent example (Liu, et al., 2018) suggests that blood biomarker analysis can be used to derive a "phenotypic age" which they describe as a "novel signature of mortality and morbidity risk" which they demonstrate correlates better with mortality than chronological age.

4 Project Structure

This project is structured as follows:

Section 5 gives an overview of the NHANES datasets and the data cleaning and manipulation undertaken. This includes a description of the demographic, health and laboratory variables and a detailed discussion of missing data considerations.

Section 6 explores the relationship between blood biomarkers and all-cause mortality in order to identify the most promising predictor variables. Cox regression modelling using penalised maximum likelihood is used to identify predictive variables and these are explored visually using standardised mortality ratios. This section includes extensive referencing to existing research.

Section 7 is the main part of the analysis and involves building a predictive model of 5-year all-cause mortality using techniques from statistical learning. This is then compared against models using more conventional predictors (e.g. BMI, smoker status, medical history). Model performance is assessed using Receiver-Operator-Characteristic (ROC) curves.

Sections 8 and 9 discuss the conclusions of this study, some of its limitations and some potential further areas of investigation.

4.1 Key assumptions / Decisions

I make the following assumptions in this project:

Time homogeneity

I am assuming that the relationship between measurements of blood biomarkers and mortality is time homogeneous i.e. results derived from historic data will still hold true today. Human biology has not really changed over this period - however it is possible that changes in environmental exposure, drugs and treatment protocols have had an impact. For example, in section 6.2.5 we observe that blood lead concentrations have fallen significantly since the 1970s due to stronger regulation and increased awareness of the toxicity of lead. Extending to more contemporary data is considered as a potential improvement in section 8.

Choice of metric

The ROC (receiver operator characteristic) curve and the area under this curve (AUC) are used to assess and compare model performance. For risk stratification ROC curves are especially useful as they enable users to choose a balance of sensitivity and specificity that suits their purpose.

Missingness mechanism

A proportion of records have missing blood readings. These are discussed in detail in section 5.3 where I make the assumption that the data can be treated as Missing-at Random.

5 Data

The datasets to be used are public-use, high-dimensional laboratory results datasets from the US NHANES population health studies, linked to national death registry records. These datasets are freely available from the CDC's website. Note that mortality follow-up is only available to 2011 (a 2015 release is expected in 2018 but was not available in time to be used in the analysis).

For the preliminary analysis I used only NHANESIII (1988-1993) data as this was in an easy-to-analyse format. From this I shortlisted a number of blood measurements which I found to be predictive of mortality – then I used both NHANESIII and Continuous NHANES to build and validate the predictive model.

A summary of the population characteristics of the datasets can be found in the following section. Note that all lives below the age of 20 at the survey have been removed as blood samples were only collected for participants over 20.

5.1 Population characteristics

5.1.1 NHANESIII

The NHANESIII (National Health and Nutrition Examination III) study was a large-scale health survey of non-institutionalised American citizens carried out in two phases from 1988-1993 by the American Centre for Disease Control and Prevention (CDC).

Data collection consisted of a survey to collect demographic, lifestyle and dietary data, a physical examination to obtain detailed body and health measurements and laboratory testing of blood and urine samples.

The CDC has also provided mortality follow up for those surveyed by linking the dataset to National Death Registry records (National Center for Health Statistics. Office of Analysis and Epidemiology, 2018). These include the latest known vital status, the primary cause of death and other contributory causes of death (e.g. diabetes, hypertension).

A summary of the population characteristics at baseline is provided in Table 2. Note that participants under the age of 20 were not subjected to the laboratory tests so have been excluded from analysis completely.

Sex	Smoker status	n	Average age at entry	Average follow-up (months)	Average BMI	Average impairment score	Number of deaths	Number of deaths within 1 year of follow up	Number of deaths within 5 years of follow up
Males	EX	2689	58.9	171.3	27.3	0.52	1420	84	436
	NS	2824	44.5	206.7	26.7	0.22	761	42	223
	SM	2440	43.5	203.0	25.7	0.21	807	36	198
	Total	**7953**	**49.1**	**193.6**	**26.6**	**0.32**	**2988**	**162**	**857**
Females	EX	1566	54.5	193.3	28.0	0.45	601	21	130
	NS	5580	49.0	203.2	27.7	0.27	1640	63	412
	SM	1931	42.4	211.5	26.6	0.21	527	6	89
	Total	**9077**	**48.6**	**203.3**	**27.5**	**0.29**	**2768**	**90**	**631**
Grand Total		**17030**	**48.8**	**198.7**	**27.1**	**0.30**	**5756**	**252**	**1488**

Table 2: Characteristics of the NHANESIII analysis dataset

5.1.2 Continuous NHANES

Since 1999 the CDC has run a similar study every 2 years, called Continuous NHANES. This takes a representative sample of the US population and performs many of the same surveys, examinations and laboratory tests. Since mortality follow-up is only available to 2011 I have used data from all the Continuous NHANES surveys from 1999 to 2007.

Sex	Smoker status	n	Average age at entry	Average follow-up (months)	Average BMI	Average impairment score	Number of deaths	Number of deaths within 1 year of follow up	Number of deaths within 5 years of follow up
Males	EX	3797	60.1	82.4	28.7	0.59	807	89	506
	NS	4913	46.9	87.3	28.5	0.26	477	40	279
	SM	3105	44.1	86.2	26.9	0.26	391	45	233
	Total	**11815**	**50.4**	**85.4**	**28.1**	**0.37**	**1675**	**174**	**1018**
Females	EX	2615	55.2	87.4	29.4	0.42	377	32	210

	NS	7942	48.5	89.9	29.0	0.26	728	63	389
	SM	2317	43.7	88.2	28.2	0.31	209	19	120
	Total	**12874**	**49.0**	**89.1**	**28.9**	**0.30**	**1314**	**114**	**719**
Grand Total		**24689**	**49.7**	**87.3**	**28.6**	**0.33**	**2989**	**288**	**1737**

Table 3: Characteristics of the Continuous NHANES dataset

The properties of the combined dataset are in Table 4.

Sex	Smoker status	n	Average age at entry	Average follow-up (months)	Average BMI	Average impairment score	Number of deaths	Number of deaths within 1 year of follow up	Number of deaths within 5 years of follow up
Males	EX	6486	59.6	119.3	28.1	0.57	2227	173	942
	NS	7737	46.0	130.9	27.8	0.25	1238	82	502
	SM	5545	43.8	137.6	26.4	0.24	1198	81	431
	Total	**19768**	**49.9**	**129.0**	**27.5**	**0.35**	**4663**	**336**	**1875**
Females	EX	4181	54.9	127.1	28.9	0.43	978	53	340
	NS	13522	48.7	136.6	28.5	0.26	2368	126	801
	SM	4248	43.1	144.3	27.4	0.27	736	25	209
	Total	**21951**	**48.8**	**136.3**	**28.3**	**0.30**	**4082**	**204**	**1350**
Grand Total		**41719**	**49.3**	**132.8**	**28.0**	**0.32**	**8745**	**540**	**3225**

Table 4: Characteristics of the Combined NHANESIII and Continuous NHANES dataset. Note that 4 records have missing smoker statuses and are omitted from this table.

Tables 2 and 3 show some of the trends we expect – smoking prevalence has decreased over this period and average BMIs have increased. Average follow up to 2011 is obviously much longer for the older study. We also note that death within 5 years are roughly evenly split between the two datasets.

5.2 Data Manipulation

5.2.1 Impairment Score

The NHANES health questionnaires include various questions about current health status which I use to allocate an "impairment score". The primary goal of this is so that we can control for known health conditions in the analysis. It is also used to isolate only healthy lives (impairment score = 0) for further analysis.

Impairment score is computed by taking the count of the number of major medical conditions that a life has (capped to 4) where the conditions considered are:

- Heart attack
- Heart failure
- Stroke
- Diabetes
- Emphysema/COPD
- Cancer

Note that these are based on self-reported medical history – not formal medical records – so they may not be completely reliable. In the rare cases where participants declined to answer or they said they were unsure I have assumed that the answer would have been negative.

As expected the number of impairments increases with increasing age, which is important to consider in analysis (e.g. by including an age x impairment interaction term in the model).

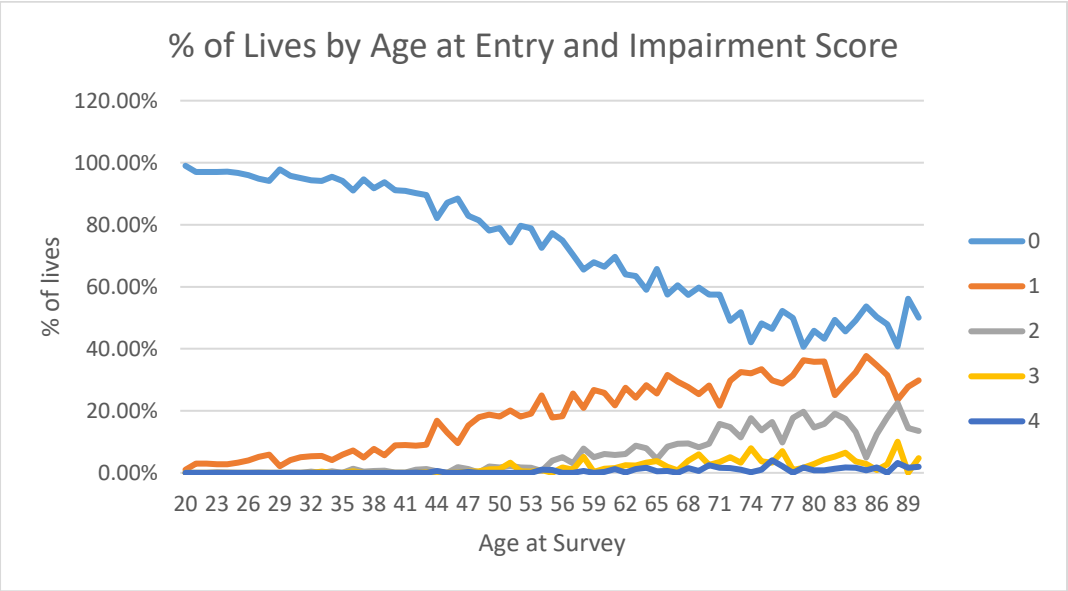

Figure 1: Percentage of lives by impairment count and age at entry. As expected younger lives are mostly unimpaired. After age 75 about 50% of lives have at least one impairment.

This impairment score acts as a strong predictor of mortality. For illustrative purposes Kaplan Meier survival curves for 60-69 year-old lives with differing levels of impairment are shown below in Figure 2.

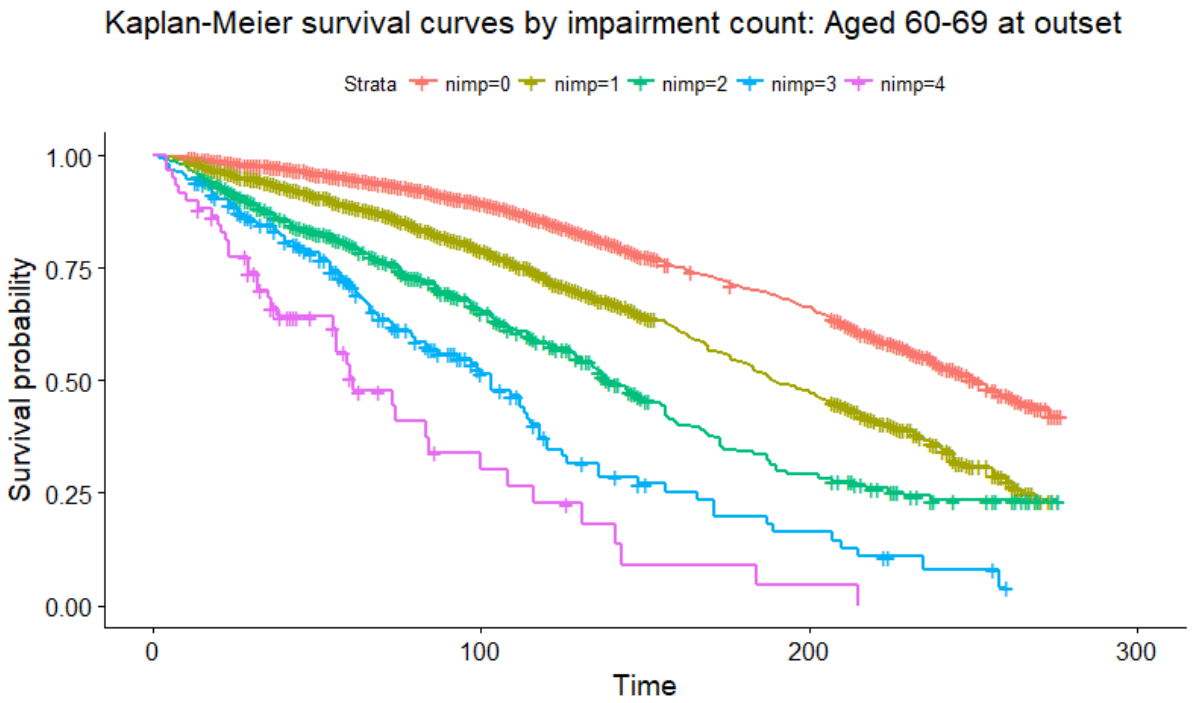

Figure 2: Kaplan -Meier survival plots for lives with differing levels of impairment. The time axis is months since the initial survey. Note that this graph has been produced using the combined dataset, so includes both NHANESIII and Continuous NHANES participants.

5.2.2 Smoker Status

After age and medical history, smoker status is the most important predictor of longevity. Smokers are over 20 times more prone to lung disease and at increased risk of cardiovascular and respiratory conditions (Pirie, et al., 2013).

The NHANES surveys include several questions on smoking history which I have processed into a single smoker status variable. A more comprehensive treatment would be to take account of smoking intensity and duration since quitting.

The definitions used are:

- Never Smoker: Someone whose survey responses indicate they have consumed fewer than 100 cigarettes in their lifetime
- Smoker: Someone whose survey responses indicate they currently smoke.
- Ex-Smoker: Someone whose survey responses indicate that they have smoked more than 100 cigarettes in their lifetime but do not currently smoke.

Note that the chemical cotinine is found in high concentrations in the blood following exposure to tobacco smoke and therefore may be a more reliable indicator of smoking habits than self-reported smoker status. The use of cotinine as a proxy for smoker status is explored more in section 6.2.7.

Kaplan Meier survival curves by smoker status are shown in Figure 3. As expected current smokers have the highest mortality, never-smokers the lowest and ex-smokers are somewhere in-between.

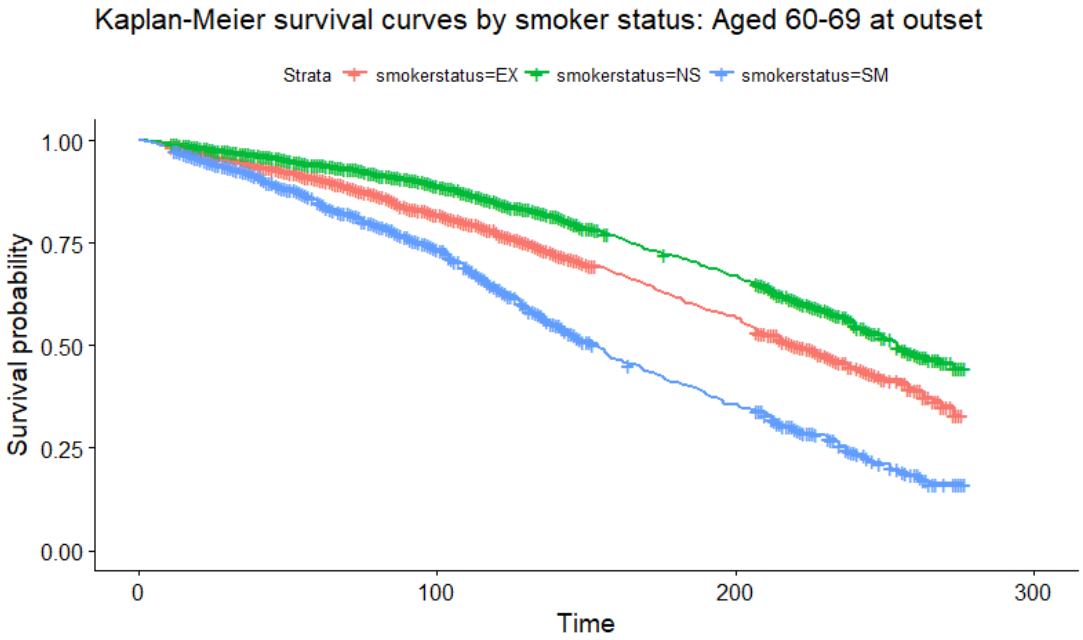

Figure 3: Kaplan-Meier survival plot for different smoker statuses. Time is in months since the initial survey.

5.2.3 BMI

Height and weight measurements were taken from all NHANES participants who attended the physical examination and BMI is included in the models.

5.2.4 Laboratory variables

The NHANESIII laboratory file contains data on 192 biomarkers collected through laboratory investigations of participants' blood and urine samples. Some of these are duplicates (e.g. the same values but converted to SI units) and others are ancillary variables (e.g. testing method).

Results from the urine tests, diabetes-specific and antibody tests are considered out-of-scope for this project as these are not part of a standard batch of blood tests.

The remaining 95 variables can be divided broadly into Haematology (39) and Biochemistry (56).

- Haematology refers to measurements of blood cells themselves, for example red blood cell count, platelet distribution width etc.
- Biochemistry relates to the presence of chemicals in the blood. For example the concentration of total proteins or the concentration of a particular enzyme.

A detailed summary of all the haematology and biochemistry variables, including missingness, can be found in the Appendix.

5.3 Missing Data

5.3.1 Considerations

We should consider the potential impact of missing data.

Most statistical procedures rely on complete observations i.e. for each observation all of the predictor variables are present. However simply taking the fully complete observations can lead to bias and a reduction in statistical power.

In the NHANES dataset, age, sex and survival time are complete. However, there are a material number of missing observations in the laboratory and questionnaire variables. Missing observations arise because:

- Some participants only completed the home questionnaire and did not attend the laboratory tests.
- Some laboratory results are discarded due to errors (e.g. lost samples)
- Respondents answered "Don't know" or refused to answer certain questions (only applicable to the survey items, e.g. smoking questions, medical history).

We note that the number of missing observations tend to increase with increasing age – it is possible that older and potentially sicker people are not able to attend the laboratory tests and hence have a larger proportion of missing values. Figure 4 shows and example of this for C-Reactive Protein.

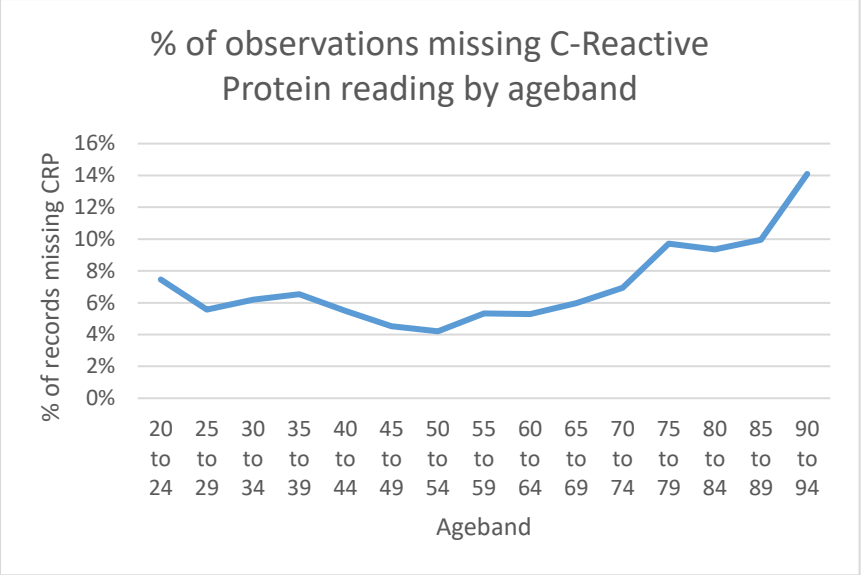

Figure 4: Proportion of lives missing C-Reactive Protein readings by age. This pattern of increasing missingness at older ages is seen in most of the laboratory variables.

There are 3 classes of missing data mechanisms, described briefly below:

- Missing completely at random (MCAR)
    - The probability that a value is missing is unrelated to both the observed and unobserved data for that observation. For example, if a random set of the blood tests is simply lost then this might be a MCAR process.
- Missing at random (MAR)
    - The probability that a value is missing is, conditional on the observed data, independent of the unobserved data. For example, smokers have significantly higher levels of the chemical cotinine in their blood. If among smokers the probability of observing cotinine is independent of the missing cotinine values themselves then this could be considered MAR.
- Missing not at random (MNAR)
    - The missingness of a variable cannot be fully explained by variables in the dataset and is dependent on the missing value itself. For example, if people with high BMIs were less likely to respond to the BMI questions, and this dependence could not be overcome using other observed variables, then this would be a MNAR mechanism.

Since we have observed that a greater proportion of readings are missing at older ages we conclude that missing observations are not missing completely at random.

Although it is by definition impossible to rule out MNAR missingness, I assume in this project that the missing data mechanism is MAR. This seems reasonable given the likely reasons for the missing data discussed above. Lost lab samples and failed lab tests are likely to be MCAR, whilst failure to attend the lab tests is likely explainable by a combination of older age and frailty.

5.3.2 Further Analysis

Figure 5 below shows the pattern of missing data in the final (combined) analysis dataset.

sex	age	d5y	nimp	smokerstatus	bmi	pbp	mpv	rdw	wbc	mcv	rbc	lmp	crp	alb	cre	bun	chl	sua	glu	pro	ldh	cot	tri	Number of observations with this missingness pattern
1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	33579
1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	0	3666
1	1	1	1	1	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1806
1	1	1	1	1	0	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	531
1	1	1	1	1	1	1	1	1	1	1	1	1	0	0	0	0	0	0	0	0	0	0	0	466
1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	0	1	434
1	1	1	1	1	1	1	1	1	1	1	1	1	1	0	0	0	0	0	0	0	0	1	0	187
1	1	1	1	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	162
1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	0	0	132
1	1	1	1	1	1	1	0	0	0	0	0	0	1	1	1	1	1	1	1	1	1	1	0	123
1	1	1	1	1	1	1	1	1	1	1	1	0	1	1	1	1	1	1	1	1	1	1	1	109
1	1	1	1	1	1	1	0	0	0	0	0	0	1	1	1	1	1	1	1	1	1	1	1	78
1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	0	1	1	75
1	1	1	1	1	1	1	1	1	1	1	1	1	1	0	0	0	0	0	0	0	0	0	0	58
1	1	1	1	1	1	1	1	1	1	1	1	1	0	0	0	0	0	0	0	0	0	1	0	57
All other missing patterns																								260
0	0	0	3	4	756	2065	2215	2215	2216	2216	2217	2336	2592	2833	2835	2835	2836	2837	2837	2857	2919	3168	6798	
Number of observations which are missing this reading																								

Figure 5: Missingness patterns for the combined dataset. Note that only the 15 most prevalent missingness patterns are shown.

We see that we have complete data in 33,579 (80.4%) of records. The second-most common missingness pattern is to be missing readings for triglycerides only (3,666 records).

We see that 1,968 records (1,806+162) have no data at all in the blood variables of interest, so these are not really useful for the analysis.

466 appear to be missing all of the standard biochemistry variables (CRP onwards) and a smaller number appear to be missing all of the standard haematology variables. I suggest that these are likely to be blood-sample handling errors – these participants must have attended the laboratory tests but one of their samples was likely lost or analysed incorrectly.

5.3.3 Choice of missing data method

A number of methods were considered. Table 5 provides an overview of the main missing data methods.

Method	Description
Deletion	Simply deleting all observations with missing values is the default behaviour of most statistical procedures, however this is not ideal as it doesn't make the best use of the available data. For example, an observation which is only missing, say, C-reactive protein measurements, still has useful data in the other variables. Simply disregarding these observations will typically lead to bias unless the missingness mechanism is Missing Completely at Random.
Single Mean / Median imputation	This involves replacing missing values with the mean / median of the non-missing values. In an MCAR situation this will produce unbiased estimates, however there is an artificial decrease in variance. Note also that relationships between variables will not be reflected in the imputed values. Median imputation is generally a good "naïve" approach and often a good starting point for dealing with missing values.
Single model-based imputation	This involves building a model of the missing variables, typically a regression model, although other methods are also popular (e.g. random forest imputation). This preserves relationships between the variables and under the MAR assumption should produce unbiased results. Additional noise can be added to the imputed values to preserve the variance.
Multiple imputation	Multiple imputation is considered a good general-purpose method for handling missing data. This involves creating a large number of imputed datasets, analysing each one separately and then combining the predictions produced from each dataset. If the MAR assumption is valid then this produces unbiased results that reflect the relationships between variables and also preserves the variance of the values.
Hot-deck imputation	"Hot deck imputation is a method for handling missing data in which each missing value is replaced with an observed response from a "similar" unit." (Andridge & Little, 2010) This is commonly-used method of handling MAR data and has the benefit of accurately reflecting the empirical distributions of the observed values, without any strong assumptions about their underlying distribution. Values are sampled at random (with replacement) from lives with similar properties.

Table 5: Summary of missing data methods

In this project I use two missing data methods:

For the **exploratory analysis** of section 6 (using only NHANESIII data to shortlist variables of interest) I used a pragmatic deletion approach.

- I ignored blood variables that have more than 10% missing values, and then restricted to the subset of the sample that has complete data for all of the remaining variables.
- This reduces the number of candidate variables to 57. A step-through of the reduction in variables and observations due to these restrictions is shown below in Table 6.
- This leads to discarding 22.5% of records from the exploratory dataset.

	Number of observations	Number of variables
Full NHA3 lab file	18162	192
Restrict to age >=20	17030	192
Restrict to blood vars	17030	95
Restrict to <10% missing	17030	57
Restrict to complete cases only	14078	57

Table 6: Summary of record and variable exclusions from the NHANESIII dataset

In the **main analysis** of section 7 I use a hot-deck imputation approach, partitioned by age band, smoker status and 5-year vital status.

- Hot deck imputation was performed using the VIM package in R (Kowarik & Templ, 2016).
- Table 7 shows the number of observations in each partition of the hot deck procedure. The highest and lowest age bands were combined in order to have a reasonable number of observations in each cell.

	EX		NS		SM	
Age band	**Survived**	**Died**	**Survived**	**Died**	**Survived**	**Died**
20-49	3460	39	12188	88	6395	127
50-59	2915	105	9496	128	5399	207
60-69	3569	287	7879	249	4251	356
70-79	3667	615	6554	501	2677	434
80+	2356	956	5560	1156	1504	414

Table 7: Number of lives in each partition of the hot deck imputation procedure

Note that the inclusion of the response (death in 5 years) in the hot deck partitions is important. This ensures that the imputed variables have the appropriate relationship with mortality. For example, in section 6.2.2 we will see that a low Red Blood Cell count is associated with higher mortality. We can check that this relationship is exhibited by the imputed variables, illustrated in Figure 6. Note that generally records with imputed data have higher mortality – since sicker lives are less more likely to not attend the lab tests.

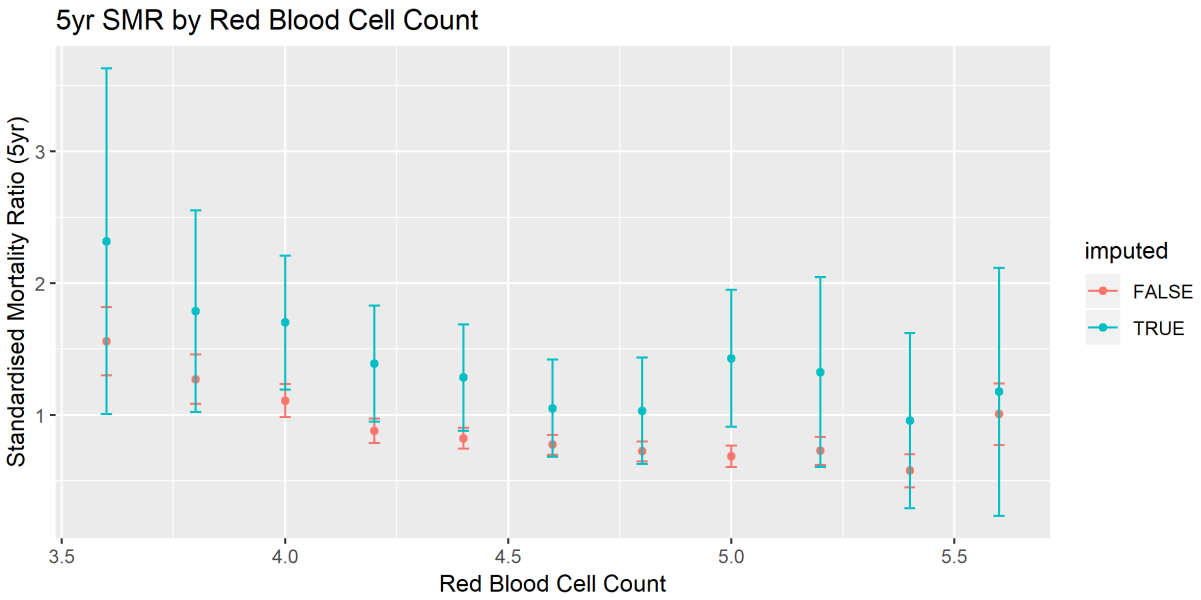

Figure 6: Illustration of 5-year age- and sex-standardised mortality ratios for imputed and unimputed values of Red Blood Cell count.

6 Exploratory Analysis

In the first part of this section I use a model-based approach to identify the variables of interest.

In the second part I look at the shortlisted variables in greater detail and use a graphical approach to investigate them further.

6.1 Variable selection

6.1.1 Cox Regression Analysis

Survival analysis methods are appropriate here as we interested in survival depending on values of measurements taken at the start of the observation period – an example of time-to-event data.

The Cox Proportional Hazards model is a standard semi-parametric approach used in survival analysis to model the relative "hazard" of different combinations of the predictor variables. Where hazard is defined as the instantaneous probability of death.

If T is a non-negative random value representing the time until death, the instantaneous hazard at time t is defined as:

$$h(t) = \lim_{dt \to 0} \left(\frac{\Pr(t \leq T \leq t + dt)}{dt} \right)$$

The form of the Cox Proportional Hazards model is then:

$$h(t) = h_0(t) \exp(\beta_1 x_1 + \beta_2 x_2 + \cdots \beta_p x_p)$$

Where $h_0(t)$ is the "baseline" hazard – a time-dependent hazard that is assumed to be the same for all lives. The β_i are the coefficients to be fitted and the x_i are the covariates of interest.

Hazard ratios are easily obtained by taking the exponent of the coefficients. For example, if x_1 is sex (coded as 0 for male, 1 for female) then the hazard ratio of female sex is simply $\exp(\beta_1)$.

The Cox Proportionate Hazards model has some disadvantages:

- The assumption of "proportional" hazards may not be appropriate.
- In its standard form it cannot allow for time-varying covariates. However, extension to time-varying covariates is possible with a straightforward extension of the model.
- It will estimate coefficients for all effects entered into the model – there is no way for the model to force small coefficients to zero to create a more parsimonious model.

Since we have many possible predictor variables (57 blood variables in the NHANESIII analysis), it is desirable to use some form of automatic effect selection to reduce the number of effects and produce a more parsimonious model. For example:

- Manual effect selection through trial and error
- Forwards / Backwards selection
- Stepwise selection
- Penalised maximum likelihood methods

In this project I use a penalised maximum likelihood method for model selection. This adds a penalty function onto the likelihood function, which is a function of the vector of coefficients $\vec{\beta}$, so that instead of simply maximising the log likelihood:

$$M(\vec{\beta}) = \log(L(\vec{\beta}|x))$$

We seek to maximise:

$$M(\vec{\beta}) = \log\left(L(\vec{\beta}|x)\right) - \lambda P(\vec{\beta})$$

Where λ is a hyperparameter that controls the strength of the penalty and P is typically either the LASSO penalty function:

$$P(\vec{\beta}) = \sum_{j=1}^{p} |\beta_i|$$

Or the Ridge Regression penalty function:

$$P(\vec{\beta}) = \sum_{j=1}^{p} \beta_i^2$$

The hyperparameter lambda is selected by cross validation i.e. testing the model performance on a subset of the data that has not been used to fit it.

I used the `glmnet` package (Friedman, et al., 2010) in R and the `coxph` function from the survival package (Therneau, 2015) to perform the penalised maximum likelihood model fit. The LASSO penalty function was selected.

I ran this model with the full set of (standardised) blood variables. As discussed in section 5.3part I only considered variables with fewer than 10% of readings missing and restricted the analysis to complete-cases only. This resulted in a dataset with 14,078 observations of 57 distinct blood variables.

I also included in the model terms for age, age^2, sex, impairment count (1-4), smoker status, BMI, BMI^2 and interactions between sex and BMI, sex and impairment count and impairment count and smoker status in order to properly control for these variables.

Further exploration of the data suggested that many of the variables had a U-shaped association with mortality so I also included quadratic terms for each of the blood variables.

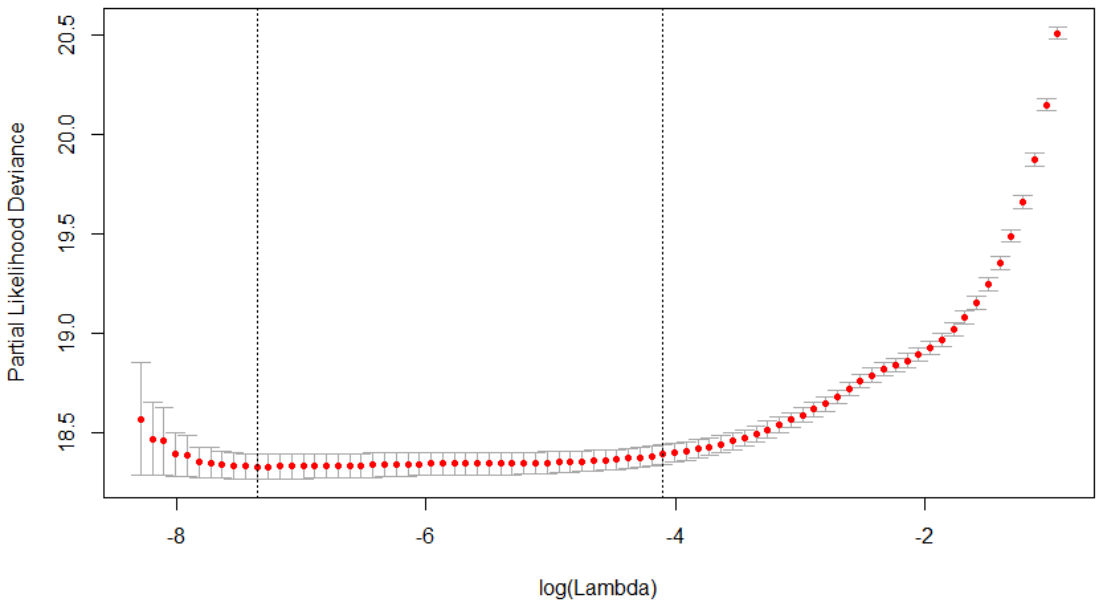

Figure 7: Cross validation partial deviance for varying levels of the penalty tuning parameter lambda.

Figure 7 shows the variation in cross-validation error seen from varying the tuning parameter lambda. Although the optimal value of lambda was log (-7.7) this does not appear to perform significantly better than a model with lambda = log (-4.0), which leads to a considerably more parsimonious model.

Therefore I decided to go with a lambda of log (-4) for model selection. This selects a model with non-zero coefficients for 21 blood variables (and the control variables age, sex, smoker status and impairment count).

Finally, I produced a new dataset by taking complete cases from only the variables selected by the regularisation procedure – yielding a slightly larger complete dataset (14,569 complete cases compared to 14,078) and fitted the selected model form.

The blood variables selected by this analysis and their level of significance can be seen in Table 8. Coefficient estimates for the full model can be found in the Appendix.

Variable	Description	p-value
wbc	White blood cell count	0.01
pbp	Lead (ug/dL)	<0.01
sel	Serum selenium (ng/mL)	<0.01
ldh	Serum lactate dehydrogenase: SI (U/L)	<0.01
sua	Serum uric acid (mg/dL)	0.01
rdw	Red cell distribution width (%)	<0.01
act	Serum alpha carotene (ug/dL)	<0.01
crp	Serum C-reactive protein (mg/dL)	0.65
aph	Serum alkaline phosphatase: SI (U/L)	<0.01
bun	Serum blood urea nitrogen (mg/dL)	0.14
pdw	Platelet distribution width (%)	<0.01
bcx	Serum beta cryptoxanthin (ug/dL)	0.28
scl	Serum chloride: SI (mmol/L)	0.01
alb	Serum albumin (g/dL)	<0.01
pro	Serum total protein (g/dL)	0.14
lyc	Serum lycopene (ug/dL)	<0.01

glu	Serum glucose (mg/dL)	<0.01
grp	Granulocyte percent (Coulter)	<0.01
cot	Serum cotinine (ng/mL)	0.01
cre	Serum creatinine (mg/dL)	<0.01
mcv	Mean cell volume: SI (fL)	0.74

Table 8: Summary of blood variables selected by the Cox analysis, with significance levels

This model achieved an R^2 of 0.50 and a concordance of 0.87. Concordance is the proportion of pairs of observations where one is correctly predicted to live longer than another. We naturally expect a high concordance for any model including age, the key driver of mortality risk.

Indeed, a baseline model with only age, sex, smoker status and impairment score achieves an R^2 of 0.47 and a concordance of 0.86, so the inclusion of blood variables appears to only be a modest improvement to the model.

6.1.2 Alternative models

In order to validate the results of the Cox model I fitted a number of other models:

Age-stratified model

An investigation of the proportional hazards assumption showed that several variables violated these – most importantly age. To resolve this, I fitted a separate model stratified by 10-year age bands, which allows the baseline hazard to vary for each age band. Note that the age term was still kept in the stratified models to capture within-band variations by age.

The stratified model resulted in almost exactly the same effect selection as the unstratified model, although coefficients were quite different. This suggests we would have reached a similar shortlist of variables using a stratified approach, but that the fitted coefficients may have been affected by non-proportionality.

Healthy-only model

The relationship between blood biomarkers and mortality is likely to vary by the health status of the individual. For example, elevated levels of C-reactive protein may be indicative of everyday stress rather than poor health. Therefore, it is useful to test the model on unimpaired lives only. I found that in general the fitted hazard ratios were similar (and all in the same direction) and variable significance was slightly lower as expected. Output from this model can be found in the Appendix.

No-blood model

This was a baseline model using only the non-blood effects from the final model. This includes age, sex, smoker status and number of impairments.

As expected performance is very good since the strongest predictors of mortality are age, impairment level and smoker status.

6.2 Exploratory data analysis

Following the initial model analysis, I carried out some deeper analysis of the variables that had been identified as predictive of mortality.

The graphs in this section depict standardised mortality ratios, calculated as:

$$SMR = \frac{observed\ number\ of\ deaths\ within\ 5\ years\ of\ entry}{expected\ number\ of\ deaths\ within\ 5\ years\ of\ entry}$$

Where the expected number of deaths is based on the overall observed mortality in each combination of age and sex, with appropriate smoothing. Note that "entry" here is defined as the date of the initial survey, not the date of the laboratory tests which were typically a few weeks later.

95% confidence intervals are also provided.

Figure 8: Raw and smoothed mortality curves split by age at entry and sex. Locally-weighted regression was used to obtained smoothed mortality curves and these are used as "Expected" mortality rates.

Note that some examples of the calculated mortality rates can be found in the Appendix.

6.2.1 WBC: White blood cell count

White blood cells are the cells of the immune system and are important in preventing and fighting off disease.

An elevated white blood cell count is associated with inflammation (Keshavarz, et al., 2013) and WBC has been shown to be an independent predictor of cardiovascular and all-cause mortality. (Ha Jee, et al., 2005).

This is quite clear in the NHANES data where we see a clear increase in mortality with increasing while blood cell count.

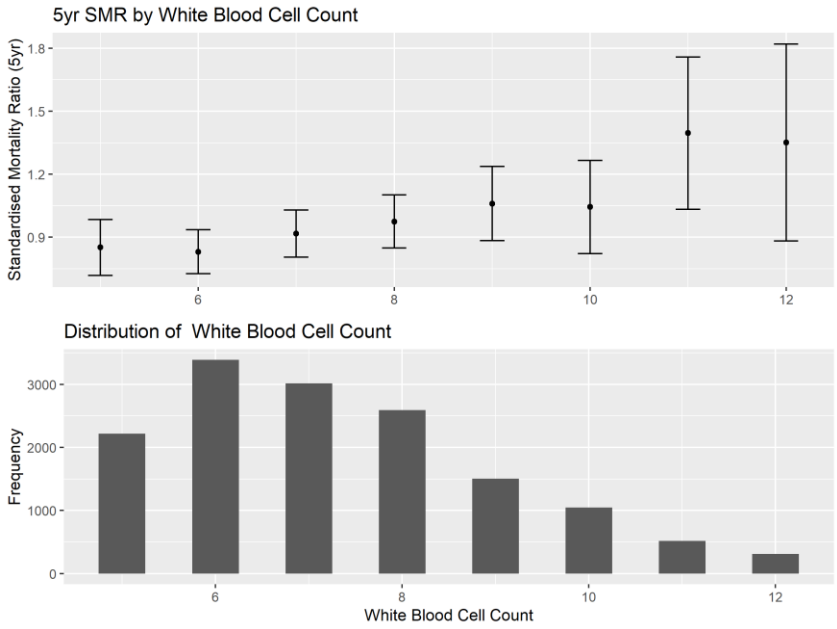

Figure 9: White Blood Cell Count

6.2.2 RBC: Red blood cell count

A low red blood cell count is associated with impaired kidney function and has been shown to be closely related to an impaired Glomerular Filtration Rate (GFR) – the main measure of renal function (Kim, et al., 2012). Kidney damage is a common side effect of metabolic diseases, notably diabetes.

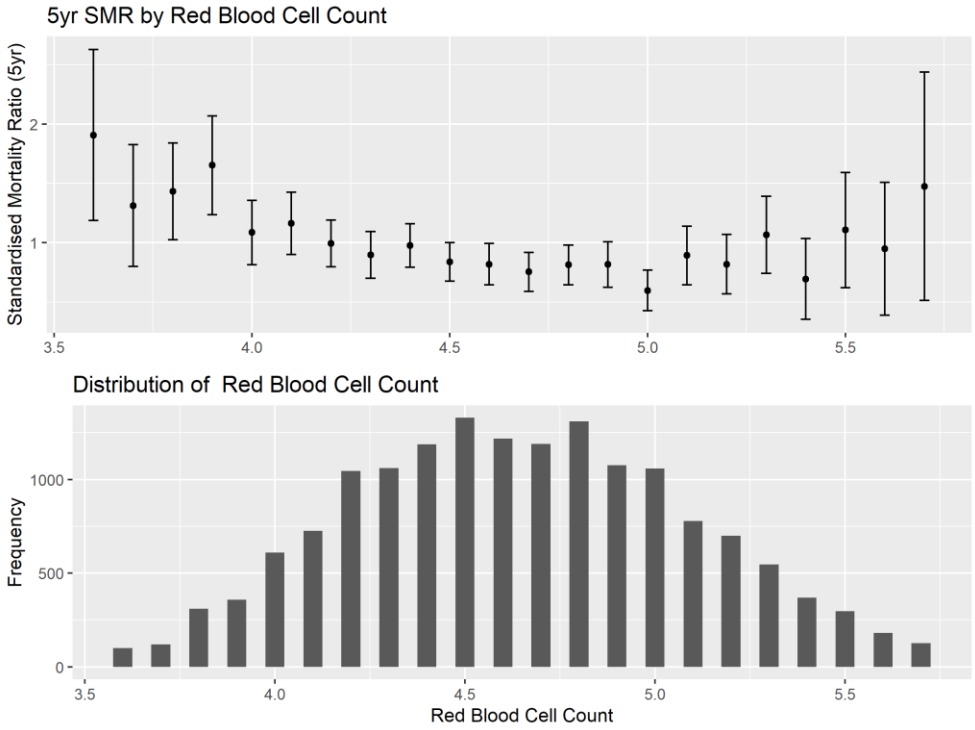

Figure 10: Red blood cell count

6.2.3 RDW: Red Blood Cell Distribution Width.

This is a measure of the variability in the size of red blood cells which is commonly used to diagnose anaemia – a lack of functioning red blood cells. We observe a clear relationship here between elevated RDW and enhanced mortality. Red blood cell distribution width has been shown to be a predictor of myocardial infarction (heart attack) in the general population (Skjelbakken, et al., 2014) and an independent predictor of cardiovascular and all-cause mortality (Perlstein, et al., 2009).

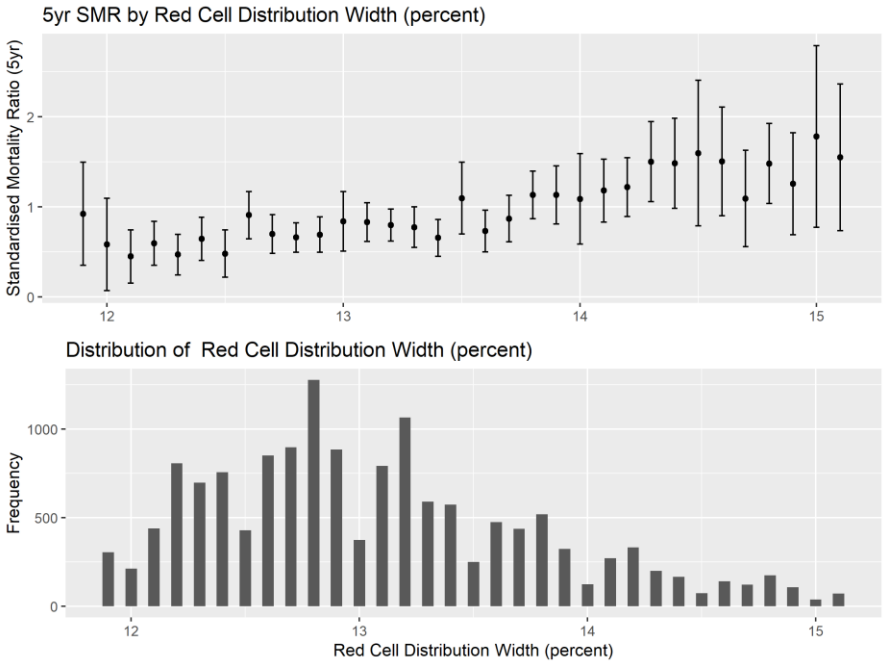

Figure 11: Red blood cell distribution width

6.2.4 PDW: Platelet Distribution Width

Platelet distribution width is a measure of the variability in the size of platelet cells in the blood. Increased platelet distribution width has been associated with higher mortality in hospitalised patients (Zhang, et al., 2015) and has been shown to be predictive of 1-year all-cause mortality in the elderly (Gonzalo-Calvo, et al., 2013).

Exploratory analysis in NHANESIII does not suggest this is a particularly strong predictor. Note that this variable is not available in the continuous NHANES dataset.

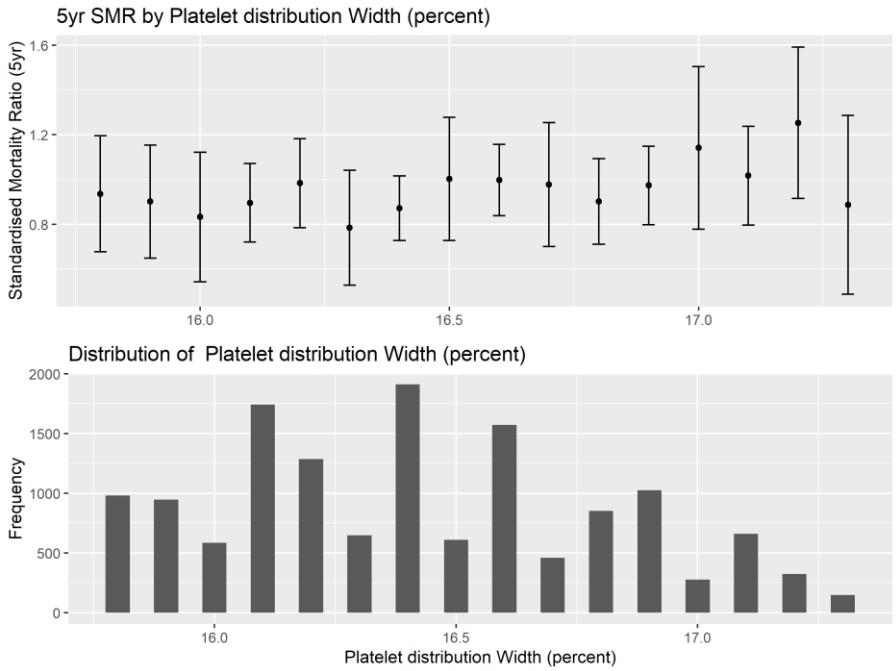

Figure 12: Platelet distribution width

6.2.5 PBP: Blood Lead

Lead is highly toxic and is associated with neurological impairment, kidney dysfunction and is likely carcinogenic.

It has been shown that blood lead concentrations greater than 5ug/DL are associated with significantly enhanced mortality (Shober, et al., 2006) – which we can see quite clearly here.

Due to greater awareness and regulation, blood lead concentrations have decreased significantly since the 1970s (Munter, et al., 2005) so its value as a mortality predictor is likely to be much lower in today's adults.

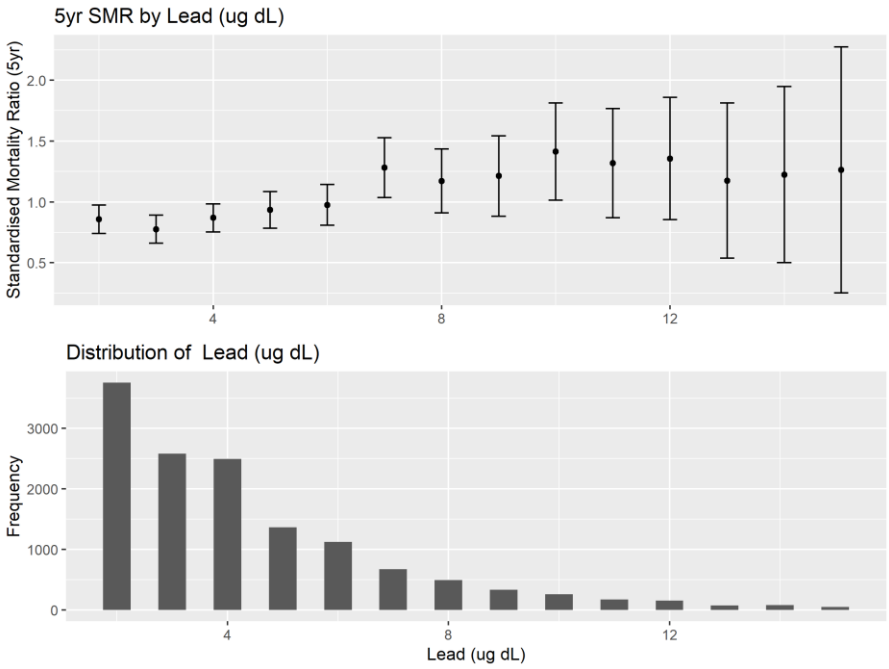

Figure 13: Blood lead

Table 9 shows the mean blood lead readings for each of the studies used in this project – and we can see a clear decrease over time.

Study	Mean blood lead (ug/dL)
NHANESIII 1988-1993	3.93
Cont. NHANES 1999	2.49
Cont. NHANES 2001	2.18
Cont. NHANES 2003	2.12
Cont. NHANES 2005	1.66
Cont. NHANES 2007	1.73
Cont. NHANES 2009	1.55

Table 9: Average levels of blood lead by study

6.2.6 CRP: C-reactive protein

CRP is a sensitive biomarker of inflammation in the body and has been shown to be predictive of all-cause mortality (Marsik, et al., 2008). This is thought to be indicative of underlying inflammatory diseases and not a causal link (Zacho, et al., 2010).

In the NHANESIII analysis we see a clear relationship between elevated CRP and increased mortality.

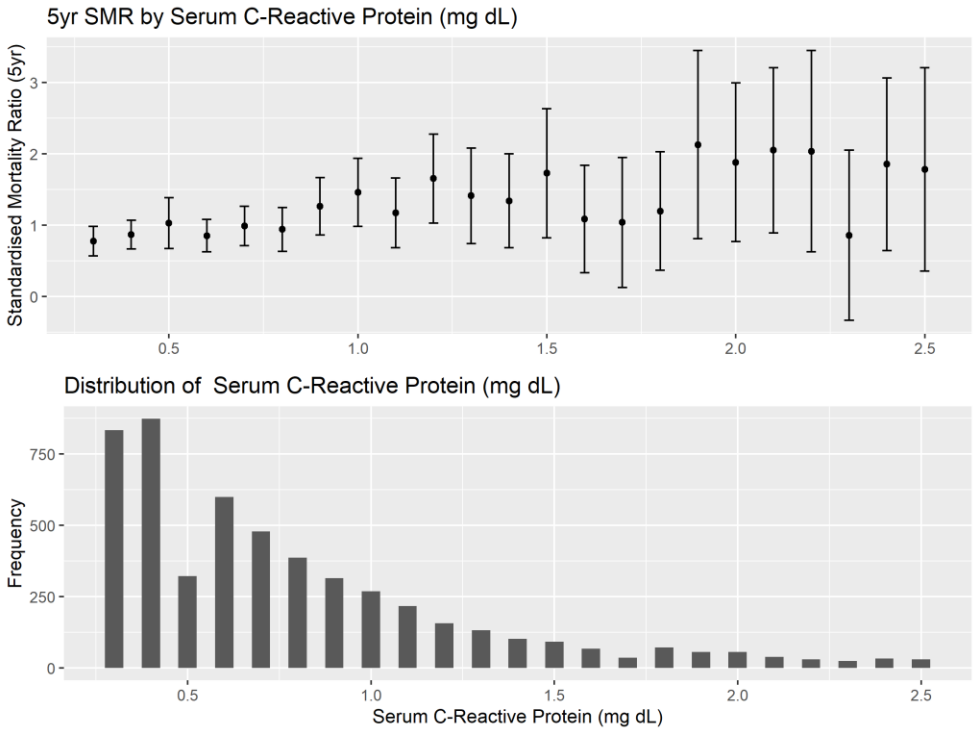

Figure 14: C-reactive protein

6.2.7 COT: Serum Cotinine

Cotinine is the chief metabolite of nicotine, and hence is present in significantly raised concentrations in the blood of smokers. Cotinine levels are negligible in non-smokers and former smokers (unless they are taking nicotine replacement therapy) (Wall, et al., 1988)

	Serum Cotinine (ng/mL)		
Smoker Status	**Lower Quartile**	**Median**	**Upper Quartile**
NS	0.03	0.06	0.28
EX	0.04	0.09	0.46
SM	102	203	308

Table 10: Summary values of Serum Cotinine for different smoker statuses

Compared to self-reported smoker status Cotinine has the advantage that it is not reliant on the participant telling the truth (as people often lie about their smoking) and is also a good indicator of smoking intensity – higher cotinine levels are found in heavier smokers.

The disadvantage is that cotinine leaves the bloodstream quickly (the half-life of cotinine is around 16 hours (Jarvis, et al., 1988)) so it cannot be used to verify ex-smokers.

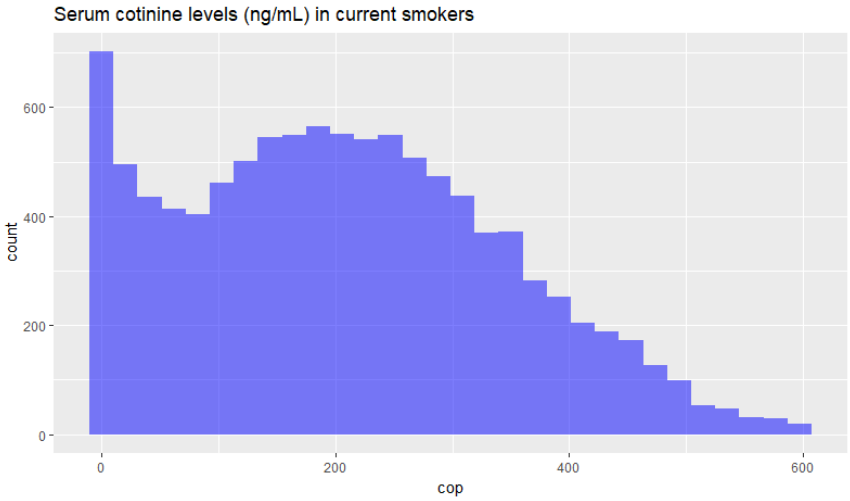

Figure 15: Distribution of cotinine levels in the blood of current smokers

Interestingly the distribution of cotinine in current smokers seems to have two peaks – possibly the difference between occasional/social and regular smokers or perhaps the difference between people who have had their first cigarette of the day and those who haven't.

Table 11 suggests a dose-response relationship between cotinine and mortality – smokers with higher levels of cotinine appear to have higher mortality. This seems reasonable, as increased levels of cotinine are associated with heavier smoking habits (Wall, et al., 1988), and there is a well-known dose-response relationship between smoking intensity and mortality (Rogers, et al., 2005).

Smoker status	Cotinine level	Observed deaths (5yr)	Expected deaths (5yr)	SMR (95% CI)
EX	<1 ng/mL	319	374	0.85 (0.76,0.95)
EX	<100 ng/mL	63	55	1.14 (0.86,1.43)
EX	>100ng/mL	62	47	1.32 (0.99,1.64)
NS	<1 ng/mL	370	487	0.76 (0.68,0.84)
NS	<100 ng/mL	53	52	1.01 (0.74,1.29)
NS	>100ng/mL	35	39	0.91 (0.61,1.21)
SM	<1 ng/mL	1	3	-
SM	<100 ng/mL	43	38	1.13 (0.79,1.47)
SM	>100ng/mL	198	146	1.36 (1.17,1.55)

Table 11: SMRs by smoker status and cotinine level.

6.2.8 CRE: Serum Creatinine

Serum creatinine is a waste product produced by the muscles and is commonly used as a proxy for glomerular filtration rate (GFR) – which is the standard measure of kidney function. Elevated levels of creatinine are associated with lower GFR and more severe kidney impairment (Damman, et al., 2012).

Additionally, low values of creatinine suggest a low muscle mass and possible frailty, which has been shown to be predictive of mortality in sick patients (Thongprayoon, et al., 2016).

These two effects are quite clearly reflected in the NHANESIII analysis where we see quite a clear U-shaped relationship with mortality.

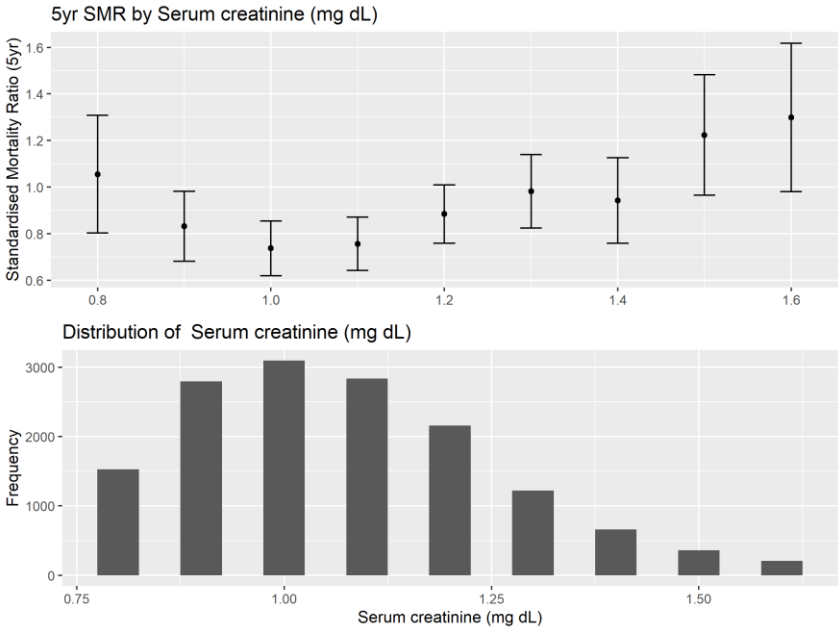

Figure 16: Serum creatinine

6.2.9 SEL: Serum Selenium

Selenium is a trace element that is thought to aid the body in defence of oxidative stress, and hence has been postulated to be protective against cancer (Brenneisen, et al., 2005). In the NHANES data we observe a beneficial effect of serum selenium levels above 130 ng/mL. Note that Selenium is not available in the Continuous NHANES dataset.

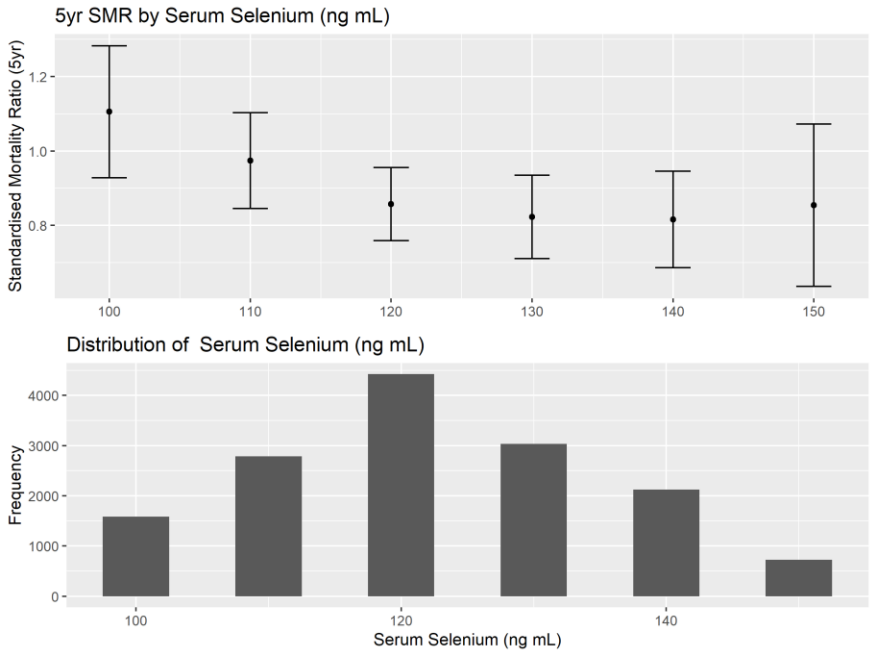

Figure 17: Serum Selenium

6.2.10 ALB: Serum Albumin

Serum albumin has been described as a "a highly sensitive indicator of preclinical disease and disease severity" (Goldwasser & Feldman, 1997). We see in NHANES quite clearly that values below 4 g/DL appear to be associated with higher mortality.

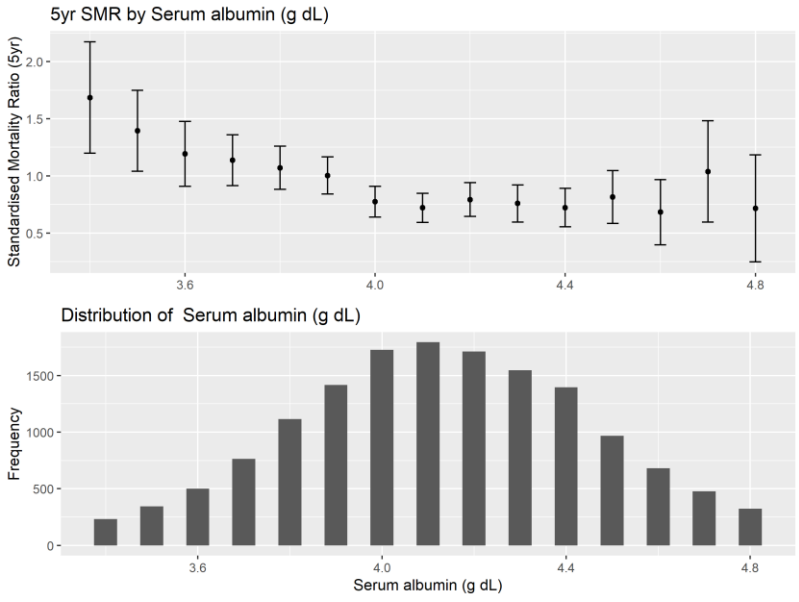

Figure 18: Serum albumin

6.2.11 PRO: Serum Total Protein

"The difference between total serum protein and albumin, i.e. the gamma gap, is a frequently used clinical screening measure for both latent infection and malignancy." (Juraschek, et al., 2015)

Although the Gamma gap wasn't explicitly investigated in the Cox model in section 6.1 we do observe greater mortality at higher protein levels and lower albumin levels.

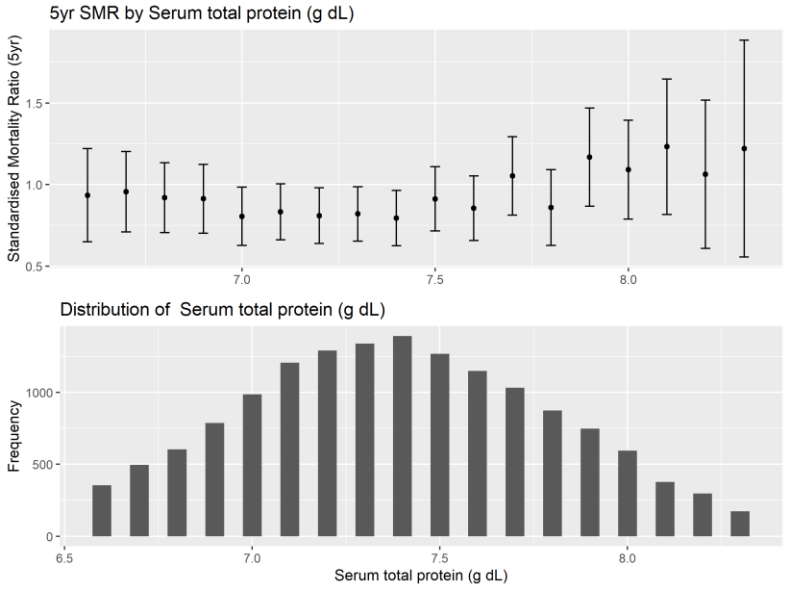

Figure 19: Serum total protein

6.2.12 LDH: Serum Lactate Dehydrogenase

Lactate Dehydrogenase is an enzyme used in the metabolic process which is typically only released into the bloodstream when cells are damaged or destroyed, so it has been used in the past as a measure of tissue damage (e.g. from heart attack).

LDH has also been demonstrated to be an effective measure of the progression of certain blood cancers (Teke, et al., 2014). In NHANESIII we can see evidence of increasing mortality with increasing levels of LDH.

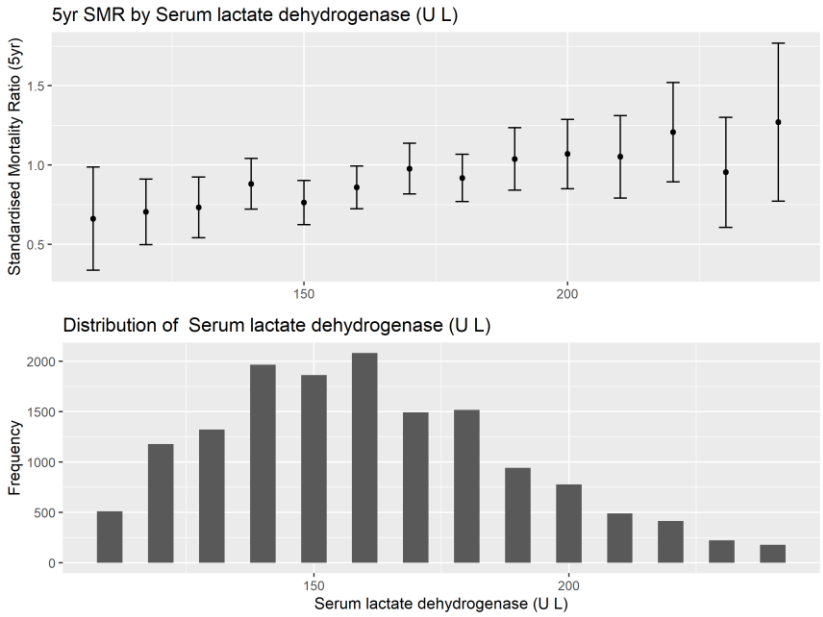

Figure 20: Serum lactate dehydrogenase

6.2.13 SCL: Serum Chloride

Low levels of chloride in the blood are known to be associated with increased mortality (McCallum, et al., 2013) although the exact mechanism is not well understood.

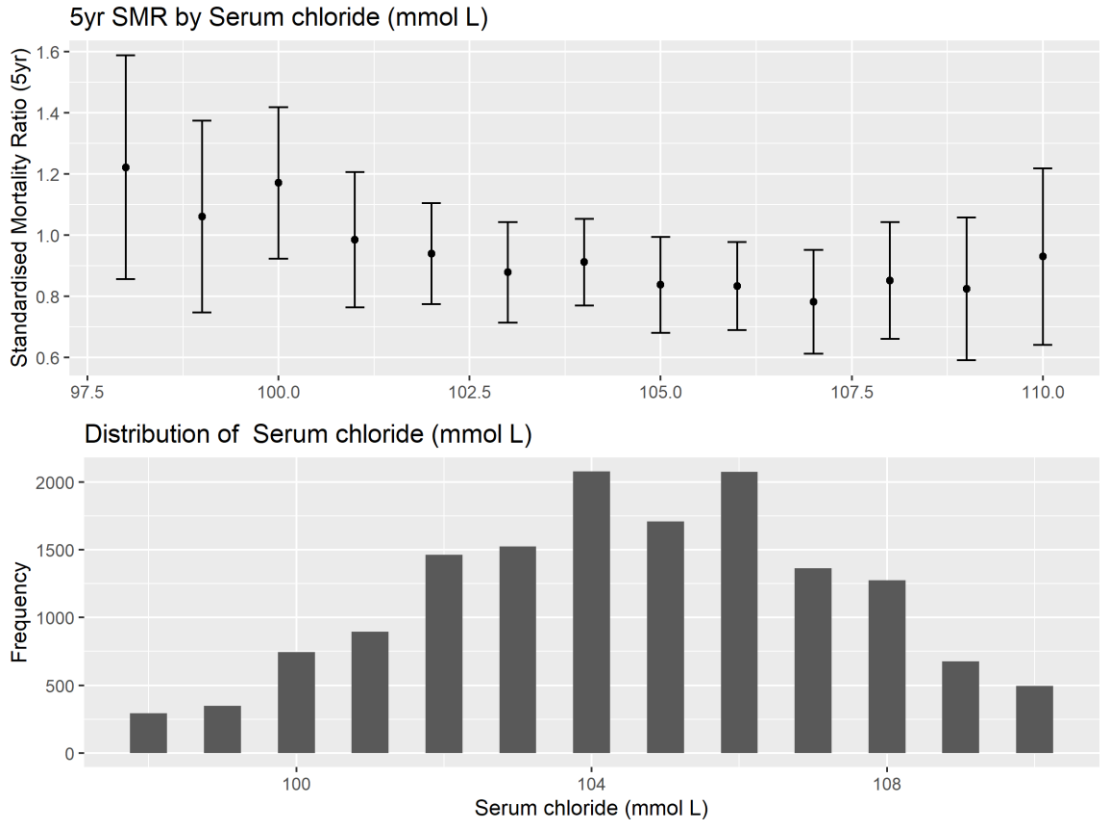

Figure 21: Serum chloride

6.2.14 BUN: Blood Urea Nitrogen

Blood Urea Nitrogen is another measure of renal (kidney) health. Urea nitrogen is a waste produce of the digestion of protein. Efficiently working kidneys will tend to keep the concentration of urea nitrogen to below 20mg/dL. Elevated Blood Urea Nitrogen has been associated with higher mortality in patients admitted to intensive care (Arihan, et al., 2018).

In NHANESIII there appears to be quite a clear increase in mortality over 20 mg/dL.

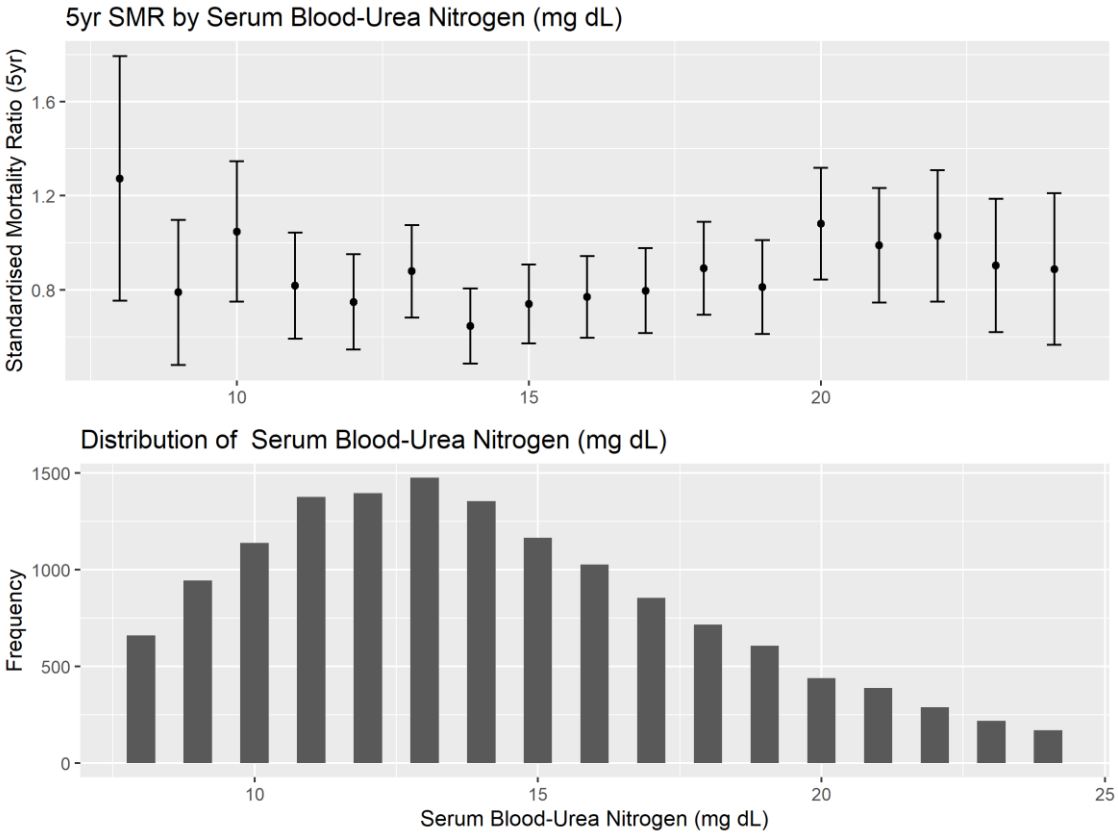

Figure 22: Blood urea Nitrogen

6.2.15 SUA: Serum Uric Acid

High levels of uric acid are associated with certain diets, poor kidney function and certain types of diuretic drugs. A study looking at an older NHANES dataset found evidence of an association between elevated Uric acid and cardiovascular mortality (Fang & Alderman, 2000).

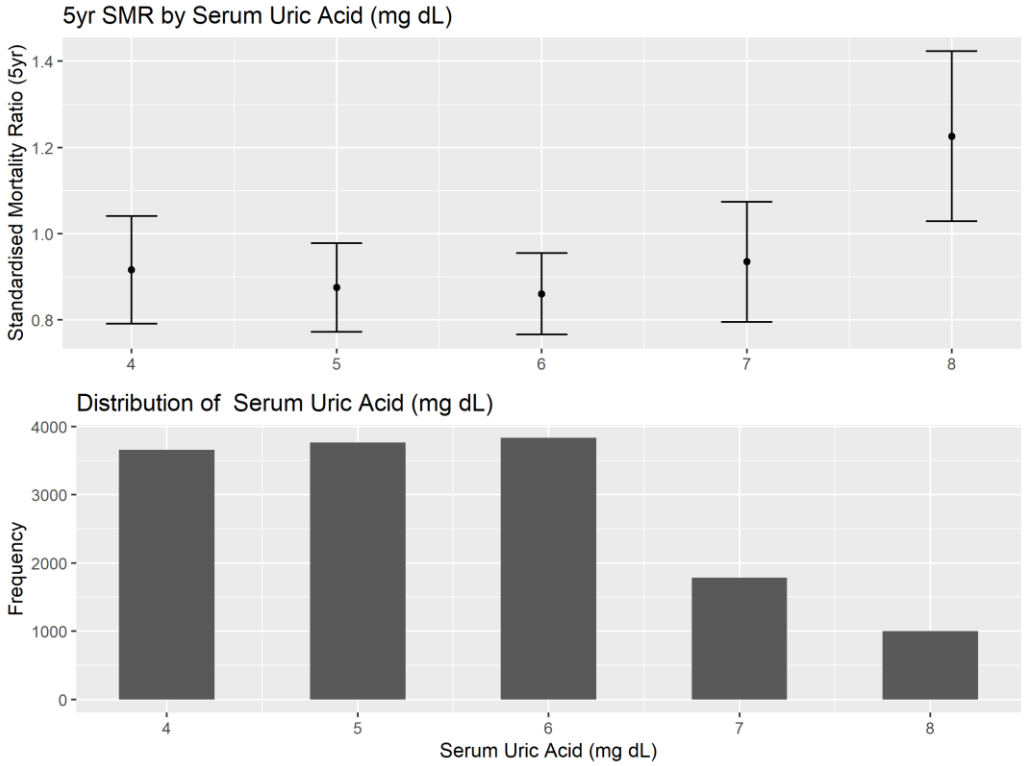

Figure 23: Serum uric acid

7 Predictive Modelling

The goal of this section is to produce a predictive model of 5-year mortality using the blood biomarkers identified in the previous section. Note that in this section I am using the full, combined and imputed dataset. This includes the full NHANESIII dataset and Continuous NHANES data, a summary of which was provided in section 5.1.

7.1 Variable changes

In the exploratory analysis I identified 21 blood biomarkers that were predictive of mortality. Unfortunately, not all of these variables are available in the Continuous NHANES datasets, and a number of variables have been dropped or swapped for a substitute in the combined dataset.

The following variables were found to be predictive but are not include in the combined dataset:

- Serum Selenium
- Alpha Carotene
- Alkaline Phosphatase
- Beta Cryptoxanthin
- Serum Chloride
- Serum Lycopene

Platelet distribution width has been replaced with another platelet measure – mean platelet volume. This was borderline for inclusion in the exploratory analysis and has found to be predictive of survival of hospitalised patients (Zhang, et al., 2015).

Granulocyte percentage is replaced with another white blood cell measure – Lymphocyte percentage. Granulocyte is a blanket term which includes most of the non-lymphocyte white blood cells – therefore we expect it to be strongly inversely correlated with Lymphocyte percentage. Indeed, in the NHANESIII dataset we observe a correlation coefficient of -0.96 between the two, so this is seen as a sensible replacement.

Finally, three additional variables were added due to known associations with morbidity and mortality. These are Red Blood Cell Count (which was borderline for inclusion in the exploratory analysis) and two measures of blood lipids – total Cholesterol and serum Triglycerides. The latter two have a known association with the incidence of stroke and heart disease so we expect these to be important for mortality prediction.

Variable	Description	Comment
wbc	White blood cell count: SI	From Exploratory Analysis
pbp	Lead (ug/dL)	From Exploratory Analysis
ldh	Serum lactate dehydrogenase: SI (U/L)	From Exploratory Analysis
sua	Serum uric acid (mg/dL)	From Exploratory Analysis
rdw	Red cell distribution width (%)	From Exploratory Analysis
crp	Serum C-reactive protein (mg/dL)	From Exploratory Analysis
bun	Serum blood urea nitrogen (mg/dL)	From Exploratory Analysis
alb	Serum albumin (g/dL)	From Exploratory Analysis
pro	Serum total protein (g/dL)	From Exploratory Analysis
glu	Serum glucose (mg/dL)	From Exploratory Analysis

cot	Serum cotinine (ng/mL)	From Exploratory Analysis
cre	Serum creatinine: SI (umol/L)	From Exploratory Analysis
mcv	Mean cell volume: SI (fL)	From Exploratory Analysis
chl	Serum cholesterol: SI (mmol/L)	Additional variable - blood lipids
tri	Serum triglycerides: SI (mmol/L)	Additional variable - blood lipids
rbc	Red blood cell count: SI	Additional variable - haematology
lmp	Lymphocyte percent (Coulter)	Replaces Granulocyte percent
mpv	Mean platelet volume: SI (fL)	Replaces Platelet distribution width

Table 12: Summary of the blood variables in the final combined dataset

7.2 Model evaluation

All models were fitted and evaluated using 5-fold cross-validation, using the area under the ROC curve as a performance metric.

The Receiver-Operator-Characteristic (ROC) curve is a commonly-used metric for binary classification problems, and is particularly suited to imbalanced classes like we have here.

The curve plots the True Positive Rate against the False Positive rate where:

- True positive rate (TPR) = Number of correctly predicted deaths / Total deaths
- False positive rate (FPR) = Number of incorrectly predicted deaths / Total survivors

Since the threshold for prediction of death can vary (the output of classification models is usually a class probability) there are many different possible combinations of TPR and FPR for a given model. These can be plotted as a curve, the area under which (Area under Curve or AUC) represents the discriminatory power of the model. I used the `pROC` package in R (Robin, et al., 2011) to compute this.

Clearly we seek a model which can obtain a high true positive rate with a low false positive rate. A perfect model would therefore have TPR=1 and FPR=0 and an AUC of 1.

7.3 Model summary

A high-level summary of the models fitted and their performance is shown below in Table 13. More detail is provided in the following sections.

Model	Model effects	Data used	Algorithm	Hyperparameters	AUC
m1	Age, sex, blood	Full (n=41,723,d=3,225)	Logistic regression	N/A	0.885
m2	Age, sex, blood	Full (n=41,723,d=3,225)	Neural network	hidden layers=1, nodes=30, decay=0.01	0.889
m3	Age, sex, blood	Full (n=41,723,d=3,225)	Random forest	ntrees=100	0.903
m3.1	Blood only	Full (n=41,723,d=3,225)	Random forest	ntrees=100	0.861
m3.2	Full model	Full (n=41,723,d=3,225)	Random forest	ntrees=100	0.905
m3.3	Blood only	Complete cases (n=34,111,d=2,309)	Random forest	ntrees=100	0.840
m4	Reference ranges	Full (n=41,723,d=3,225)	Count of extremes	N/A	0.605

Table 13: Summary of models fitted

7.3.1 M1-M3: Blood, age and sex

This initial modelling run was to compare the performance of a few different algorithms using only age, sex and blood variables.

7.3.2 M1: Logistic regression

The Logistic regression model was run with a model form of:

$$\ln\left(\frac{p}{1-p}\right) = \beta_0 + \beta_1 sex + \beta_2 age + \beta_3 age^2 + \beta_4 rdw + \beta_5 rdw^2 + \beta_6 wbc + \beta_7 crp + \beta_8 crp^2 + \beta_9 cre + \beta_{10} cre^2 + \beta_{11} alb + \beta_{12} alb^2$$

Where p is the probability of death within 5 years.

This model form was selected manually, using the results of the Cox regression analysis in section 6.1.1 as a starting point and manually adding / removing effects to improve the fit.

7.3.3 M2: Neural network

I fitted a neural network with a single hidden layer, with a model specification of:

$$d5y \sim age + age^2 + sex + bloodvars$$

Where d5y is a binary variable representing death within 5 years of follow-up (1=died, 0 =survived) and bloodvars includes all of the 18 different blood variables described in Table 12.

There are 18 bloodvars, therefore this is a neural network with 21 input nodes and 2 output nodes (corresponding to death or survival). The model was fitted using the `nnet` package in R (Venables & Ripley, 2002).

The single-layer neural network algorithm has two hyperparameters – the number of hidden nodes and the decay parameter – a variable that penalises overfitting.

Using the `caret` package (Kuhn, 2018) for model tuning, I found suitable values of 30 hidden nodes with a decay parameter of 0.01. The caret package works by carrying out a grid search of possible parameter values and identifying the combination that gives the best cross validation accuracy by repeatedly fitting and testing candidate models.

7.3.4 M3: Random Forest

The Random Forest algorithm is an ensemble learning method which works by fitting a large number of weakly-predictive decision trees and making predictions based on the mode of the predicted classes (Breiman, 2001). In order to prevent overfitting each tree is fitted with a random subset of features.

In R, Breiman's Random Forest algorithm is implemented in the package `randomForest` (Liaw & Wiener, 2002).

I fitted a random forest using the same model specification of as the neural network.

$$d5y \sim age + \ age^2 + sex + bloodvars$$

There are two hyperparameters in the random forest algorithm. The first is the number of trees and the second is the number effects to consider for each of the fitted trees.

I used the default assumption for the features to consider which is sqrt(p) where p is the total number of features – 21 in this case. I performed a manual search for a suitable number of trees to fit, using the AUC on the test partition to assess performance. I settled on 100 as increasing the number of trees beyond this did not appear to materially improve performance.

ROC curves for the performance of the first test partition are shown in Figure 24.

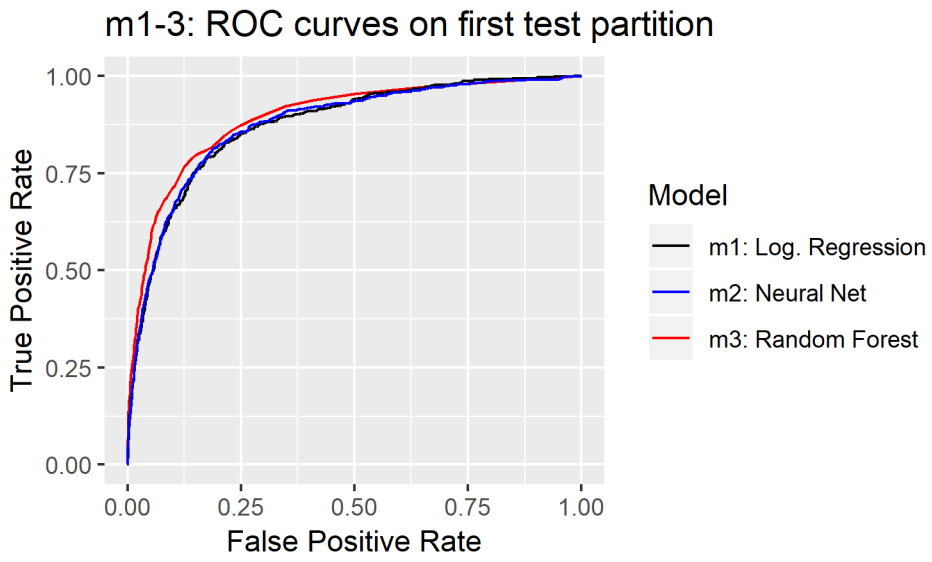

Figure 24: ROC curves for models m1-m3

AUCs for all 3 models across all the cross-validation partitions are provided below.

	Area under ROC curve (AUC)		
Test partition	**m1. Log Regression**	**m2. Neural Net**	**m3. Random Forest**
1	0.874	0.876	0.895
2	0.890	0.898	0.911
3	0.893	0.899	0.906
4	0.890	0.891	0.908
5	0.879	0.880	0.896
Average	0.885	0.889	0.903

Table 14: AUC statistics for models 1-3

The random forest algorithm clearly performs best and was chosen for deeper investigation in the rest of the models.

Although the random forest is algorithm is not as readily interpretable as a logistic regression or a single decision tree, it is possible to measure variable importance by calculating the average increase in node impurity for each variable, measured using either entropy or the Gini coefficient.

The variable importance plot is shown in Figure 25. This shows some similar findings to those of the exploratory analysis. Age is clearly the most important predictor, followed by RDW (Red Blood Cell Distribution width), BUN (Blood Urea Nitrogen), LMP (Lymphocyte percentage).

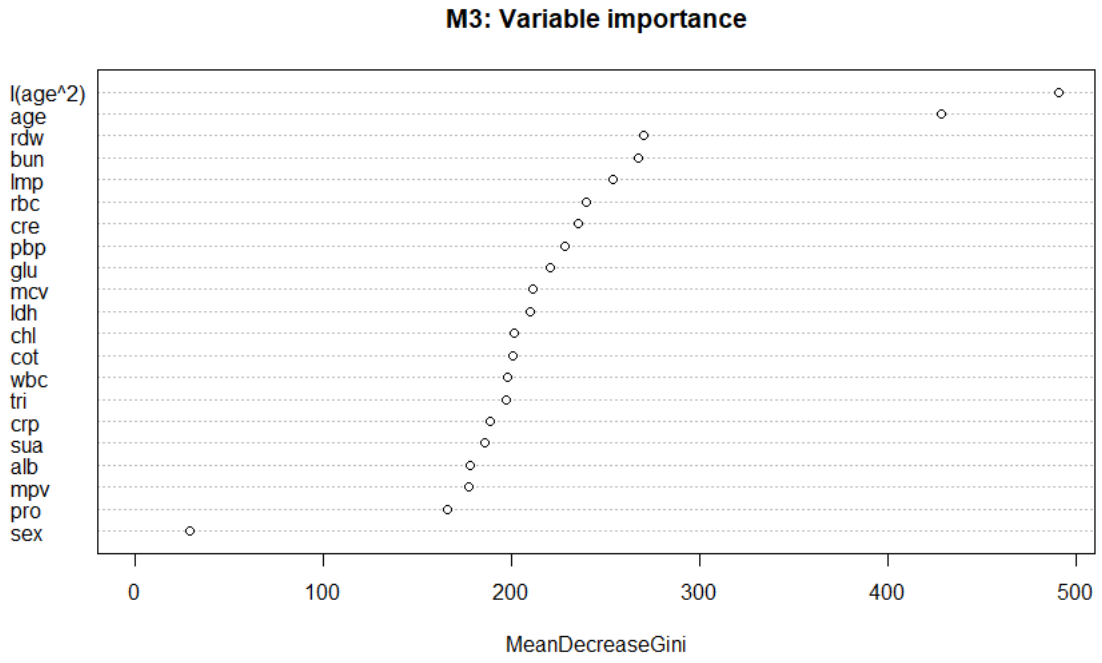

Figure 25: Variable importance plot for model M3 – a random forest using only age, sex and blood variables.

7.4 M3.1-M3.2: Alternative models

These model looks at the performance that can be achieved using:

- Only the blood variables
- All variables, including smoker status, impairment count and BMI.

Following the results of the previous section only random forests were investigated for these alternative models.

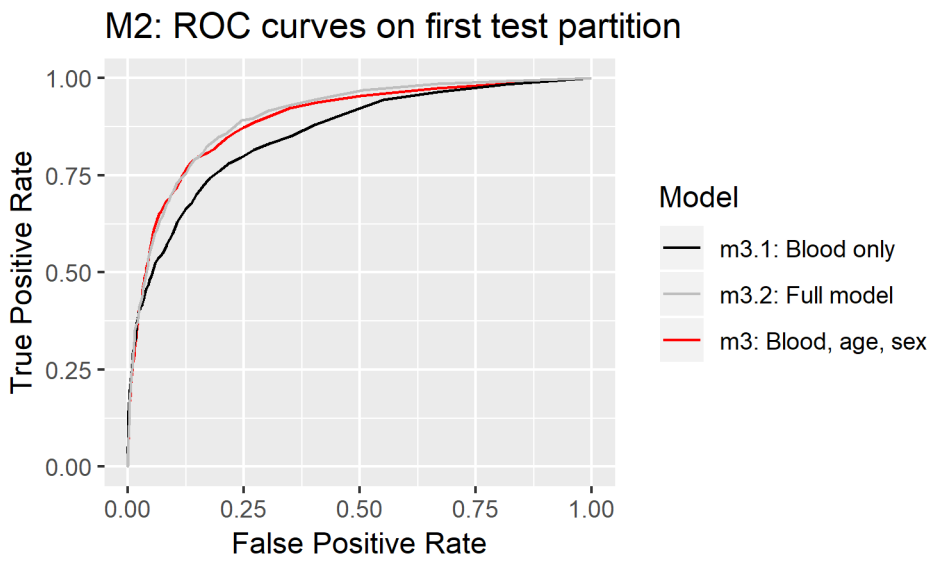

Figure 26: ROC curves for models m3, m3.1 and m3.2

Table 15 gives a summary of the relative performance of these models.

	Area under ROC curve (AUC)		
Test partition	**m3**	**m3.1 (blood only)**	**m3.2 (full model)**
1	0.895	0.858	0.902
2	0.911	0.870	0.912
3	0.906	0.860	0.906
4	0.908	0.865	0.910
5	0.896	0.852	0.896
Average	0.903	0.861	0.905

Table 15: AUC statistics for models m3, m3.1 and m3.2

Given the importance of age in model m3 it is unsurprising that we see a large decrease in performance when only blood variables are used. We see approximately twice as many false positives for the same true positive rate using this model.

Interestingly the addition of impairment count and smoker status don't seem to materially improve the model.

This is expected for smoker status – we saw in section 6.2.7 that smoker status was strongly associated with Serum Cotinine – so this is already accounted for in the model. Impairment count doesn't have a huge effect either, which is slightly surprising. It is likely however that there is extensive overlap between being sick and having impaired blood biomarkers.

The variable importance plot of Figure 27: Variable importance for m3.2 - full model reflects this too – smoker status is close to the bottom of the list and impairment count is around the middle.

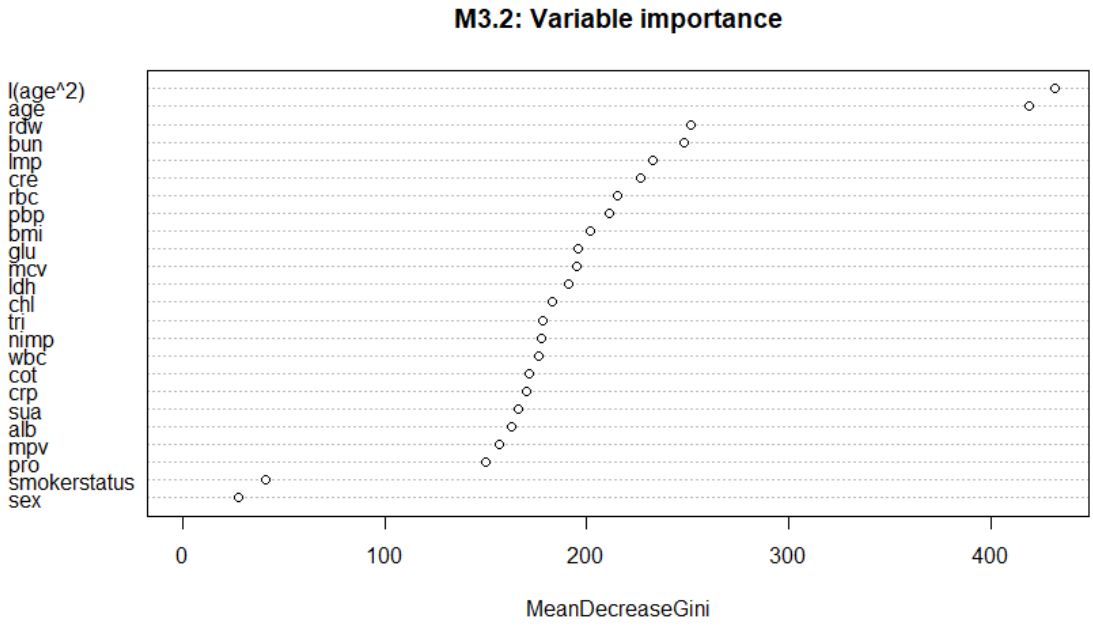

Figure 27: Variable importance for m3.2 - full model

7.5 M3.3: Complete-cases model

There is a danger that the hot-deck imputation procedure described in section 5.3 causes some bias in the dataset that invalidates the model. The hot-deck method should guard against this (since missing observations can only be filled in with real observations from similar lives) but there is a danger of bias in domains where only a small percentage of the lives have real values.

To check this I re-ran the blood only model (m3.1) using complete cases only. This gave an AUC of 0.846 – similar to the 0.864 obtained by m3.1.

Inspection of the variable importance plot also indicated that this model was similar to that fitted in m3.1. Therefore, I was satisfied that the imputation methodology hasn't substantially biased the dataset.

7.6 M4: Reference range comparison

Returning the original goal of the project, this section now aims to discuss the value of reference ranges.

Note that there is no universally-accepted set of normal reference ranges for blood tests. For this project I am using the approach used in the NHANES lab manual, which uses the 2.5th and 97.5th percentage quantiles as the lower and upper limits respectively.

One simple comparison we can do is to compare the strength of the association between having readings that exceed these reference ranges and 5-year mortality. For example, predicting that everyone who has N readings outside of the 2.5th and 97.5th percentile will die in the next five years.

Again we can compute an ROC curve for this using varying values for N and compare it against the other models.

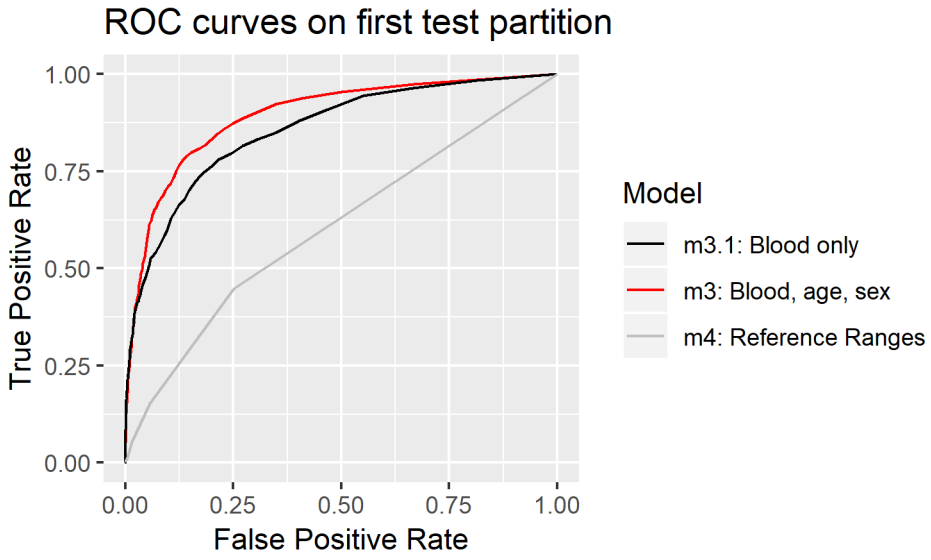

Figure 28: ROC comparison m3.1, m3 and m4

This gives an AUC of 0.605 – only slightly better than a 50/50 guess.

We conclude that simply being outside the reference ranges is a very poor predictor of mortality. This is as expected. Since reference ranges are not directly aligned to mortality we expect these in general to be poor predictors of mortality.

Clearly both the age, sex and blood random forest model and the blood-only random forest model are superior to simply using reference ranges for mortality prediction.

It is of particular clinical interest to examine cases where the blood-only model correctly predicts a death but none of the readings are outside the reference ranges. These are the lives whose blood-results are individually within normal limits but my model is able to identify that they are nonetheless at imminent risk of death.

Using the blood-only model m3.1, I find that typically about 15% of correctly predicted deaths actually have all readings within the reference ranges. Table 12 gives an overview of these for a number of different settings of the model.

Prediction Threshold	True positives	(within ref range)	False positives	TPR	FPR	% of TPs within ref range
0.1	1888	308	2964	58.5%	7.7%	16.3%
0.3	1401	191	1303	43.4%	3.4%	13.6%
0.5	1048	147	535	32.5%	1.4%	14.0%

Table 16: True / False positives and comparison against reference ranges

Inspection of a few of these indicates that these are typically where one or several biomarkers are close to, but not quite exceeding the reference ranges. In reality we would hope that these borderline cases would be noticed by doctors.

8 Limitations and Further Research

This section briefly considers some of the limitations of this project and some potential extensions.

8.1 Extension to disease prediction

The focus of this project has been on all-cause mortality, with the idea that this model could be used to identify high-risk lives and provide them with the medical interventions they need.

However, its usefulness as a diagnostic tool is obviously lacking. Although it may be useful for doctors to identify their highest risk patients, it doesn't actually tell them *what* is wrong with them or *how* to treat them.

Changing the response variable to a specific disease or cause of death could make it into a more useful tool for doctors.

8.2 Applicability to modern adults

As noted in section 4.1 we expect some changes in the relationship between blood biomarkers and mortality due to changes in lifestyle and drug technology. For example, we found that lead concentrations are much lower in today's adults than in the NHANESIII cohort due to tighter regulation around the use of lead in manufacturing. Although this was touched on the impact of these changes was not considered in detail.

One simple improvement would be to rerun the analysis with a more modern cohort. The CDC are in the process of releasing mortality follow-up data for the more modern NHANES cohorts so this could be used to improve the analysis. Alternatively, variables with known calendar-year changes should be considered for removal from the model.

8.3 Variables omitted from final analysis

Due to missingness in the continuous NHANES dataset, the combined dataset I used in the predictive modelling did not include some of the predictors I identified in the exploratory analysis of section 6.1.

While I felt this to be a sensible compromise in the interests of maximising the available data it is reasonable to assume that the model performance would have been improved by the addition of these variables.

Some of the variable omitted in the final analysis were:

- Selenium, which has been postulated to be protective against oxidative stress and cancer (Brenneisen, et al., 2005)
- Platelet distribution width, which has been demonstrated to be predictive of older age mortality (Gonzalo-Calvo, et al., 2013)

9 Conclusions

In this project I have investigated the relationship between routinely-collected blood biomarkers and all-cause mortality using US National Health Survey data.

18 different blood biomarkers were used in the final model. Some of the most important blood variables were found to be:

- Red Blood Cell Distribution Width
- Blood Urea Nitrogen
- Lymphocyte Percentage
- Serum Creatinine

Using the Random Forest algorithm, I have built a model that can be used to predict 5-year mortality from blood test results.

I propose that such a model could be applied to routine blood tests and the results could be used for more sophisticated risk stratification by doctors. This approach was demonstrated to be much more effective at identifying high mortality risk patients than the standard “reference range” approach. It was also found that typically around 15% of the deaths correctly predicted by my model had readings within normal blood reference ranges – these lives may be missed by doctors who are reliant on this approach.

An extended form of the model using age and sex was shown to give much better performance. Finally a model including BMI, smoker status and number of serious medical impairments was shown to provide even better predictive performance.

A clear extension of the model would be to disease prediction or a specific cause of death (e.g. stroke, cancer). Although the model is useful for identifying high risk patients it doesn’t necessarily help doctors to decide on an appropriate medical intervention or diagnosis since it doesn’t explain why they are at higher risk.

10 References

Andridge, R. R. & Little, R. J. A., 2010. A Review of Hot Deck Imputation for Survey Non-response. *Int Stat Rev,* 78(1), pp. 40-64.

Arihan, O., Wernly, B. & Lichtenauer, M., 2018. Blood Urea Nitrogen (BUN) is independently associated with mortality in critically ill patients admitted to ICU.. *PLoS One,* 13(1).

Boyle, M. & Senior, K., 2008. In: *Biology: Third Edition.* s.l.:Collins, pp. 226-231, 313.

Breiman, L., 2001. Random Forests. *Machine Learning,* 45(1), pp. 5-32.

Brenneisen, P., Steinbrenner, H. & Sies, H., 2005. Selenium, oxidative stress, and health aspects. *Molecular Aspects of Medicine,* 26(4-5), pp. 256-267.

Damman, K., Voors, A. A. & Navis, G., 2012. Current and novel renal biomarkers in heart failure. *Heart Failure Reviews,* 17(2), pp. 241-250.

Fang, J. & Alderman, M. H., 2000. Serum Uric Acid and Cardiovascular Mortality. *JAMA,* 283(18), pp. 2404-2410.

Friedman, J., Hastie, T. & Tibshiriani, R., 2010. Regularization Paths for Generalized Linear Models vis Coordinate Descent. *Journal of Statistical Software,* 33(1), pp. 1-22.

Goldwasser, P. & Feldman, J., 1997. Association of serum albumin and mortality risk. *Journal of Clinical Epidemiology,* 50(6), pp. 693-703.

Gonzalo-Calvo, D., Luxan-Delgado, B., Rodriguez-Gonzalez, S. & Garcia, M., 2013. Platelet distribution width is associated with 1-year all-cause mortality in the elderly population. *Journal of Clinical Gerontology and Geriatrics,* 4(1), pp. 12-16.

Ha Jee, S. et al., 2005. White Blood Cell Count and Risk for All-Cause, Cardiovascular, and Cancer Mortality in a Cohort of Koreans. *American Journal of Epidemiology.*

Hippisley-Cox, J. et al., 2008. Predicting cardiovascular risk in England and Wales: prospective derivation and validation of QRISK2. *BMJ,* 336(1475).

Jarvis, M. J., Russell, M. A., Benowitz, N. L. & Feyerabend, C., 1988. Elimination of cotinine from body fluids: implications for noninvasive measurement of tobacco smoke exposure. *American Journal of Public Health.*

Juraschek, S. P., Moliterno, A. R., Checkley, W. & Miller III, E. R., 2015. The Gamma Gap and All-Cause Mortality. *PLoS One,* 10(12).

Keshavarz, S.-A.et al., 2013. White Blood Cell Count in Women: Relation to Inflammatory Biomarkers, Haematological Profiles, Visceral Adiposity, and Other Cardiovascular Risk Factors. *Journal of Health, Population and Nutrition,* 31(1), pp. 58-64.

Kim, Y. C. et al., 2012. The Low Number of Red Blood Cells is an Important Risk Factor for All-Cause Mortality in the General Population. *The Tohoku Journal of Experimental Medicine,* 227(2), pp. 149-159.

Kowarik, A. & Templ, M., 2016. Imputation with the R Package VIM. *Journal of Statistical Software,* Volume 74, pp. 1-16.

Kuhn, M., 2018. *caret: Classification and Regression Training.* [Online]
Available at: https://CRAN.R-project.org/package=caret

Liaw, A. & Wiener, M., 2002. Classification and Regression by randomForest. *R News,* 2(3), pp. 18-22.

Liu, Z. et al., 2018. Phenotypic Age: a novel signature of mortality and morbidity risk [preprint]. *bioRxiv.*

Marsik, C. et al., 2008. C-Reactive Protein and All-Cause Mortality in a Large Hospital-Based Cohort. *Clinical Chemistry.*

McCallum, L., Jeemon, P., Hastie, C. E. & Patel, K. R., 2013. Serum Chloride Is an Independent Predictor of Mortality in Hypertensive Patients. *Hypertension,* 62(5), pp. 836-843.

Munter, P., Menke, A. & DeSalvo, K. B., 2005. Continued Decline in Blood Lead Levels Among Adults in the United States. *JAMA Internal Medicine,* 165(18), pp. 2155-2161.

National Center for Health Statistics. Office of Analysis and Epidemiology, 2018. *The Linkage of National Center for Health Statistics Survey Data to the National Death Index — 2015 Linked Mortality File (LMF): Methodology Overview and Analytic Considerations,,* s.l.: s.n.

NHS England, 2017. *National Schedule of Reference Costs 1,* s.l.: s.n.

Perlstein, T. S., Weuve, J., Pfeffer, A. M. & Beckman, J. A., 2009. Red blood cell distribution width and mortality risk in a community-based prospective cohort: NHANES III. *Arch Internal Medicine.*

Pirie, K., Peto, R., Reeves, G. K. & Beral, V., 2013. The 21st century hazards of smoking and benefits of stopping: a prospective study of one million women in the UK. *Lancet.*

Robin, X. et al., 2011. pROC: an open-source package for R and S+ to analyze and compare ROC curves. *BMC Bioinformatics,* Volume 12, p. 77.

Rogers, R. G., Hummer, R. A., Krueger, M. P. & Pampel, F. C., 2005. Mortality attributable to cigarette smoking in the United States. *Popul Dev Rev,* 31(2), pp. 259-292.

Shankar, A., Mitchell, P., Rochtchina, E. & Wang, J. J., 2007. The association between circulating white blood cell count, triglyceride level and cardiovascular and all-cause mortality: Population-based cohort study. *Atherosclerosis,* 192(1), pp. 177-183.

Shober, E. S. et al., 2006. Blood Lead Levels and Death from All Causes, Cardiovascular Disease, and Cancer: Results from the NHANES III Mortality Study. *Environmental Health Perspectives.*

Skjelbakken, T. et al., 2014. Red Cell Distribution Width Is Associated With Incident Myocardial Infarction in a General Population: The Tromso Study. *Journal of the American Heart Association,* 3(4).

Teke, H. U., Basak, M., Teke, D. & Kanbay, M., 2014. Serum Level of Lactate Dehydrogenase is a Useful Clinical Marker to Monitor Progressive Multiple Myeloma Diseases: A Case Report. *Turkish Journal of Hematology,* 31(1), pp. 84-87.

Therneau, T., 2015. *A Package for Survival Analysis in S version 2.38.* [Online]
Available at: https://CRAN.R-project.org/package=survival

Thongprayoon, C., Cheungpasitporn, W. & Kashani, K., 2016. Serum creatinine level, a surrogate of muscle mass, predicts mortality in critically ill patients. *Journal of Thoracic Disease,* 8(5), pp. 305-311.

Venables, W. N. & Ripley, B. D., 2002. Modern Applied Statistics with S. In: New York: Springer.

Wall, M. A., Johnson, J., Peyton, J. & Benowitz, N. L., 1988. Cotinine in the Serum, Saliva, and Urine of Nonsmokers, Passive Smokers and Active Smokers. *AM J Public Health,* 78(6), pp. 699-701.

Zacho, J., Tybjaerg-Hansen, A. & Nordestgaard, B. G., 2010. C-reactive protein and all-cause mortality—the Copenhagen City Heart Study. *European Heart Journal,* 31(13), pp. 1624-1632.

Zhang, S. et al., 2015. Use of Platelet Indices for Determining Illness Severity and Predicting Prognosis in Critically Ill Patients. *Chinese Medical Journal,* 128(15), pp. 2012-2018.

11 Appendix

11.1 Variable Summary

This is a summary of the Haematology and Biochemistry variables in the NHANESIII laboratory file.

Note that this is after restricting the dataset to ages 20+.

NHANESIII variable name	Report variable name	Missing	Miss %	Description
wcp	wbc	903	5.3%	White blood cell count
lmppcnt	lmp	905	5.3%	Lymphocyte percent (Coulter)
moppcnt	moppcnt	1214	7.1%	Mononuclear percent (Coulter)
grppcnt	grp	1214	7.1%	Granulocyte percent (Coulter)
rcp	rcp	905	5.3%	Red blood cell count
hgp	hgp	903	5.3%	Hemoglobin (g/dL)
htp	htp	905	5.3%	Hematocrit (%)
mvpsi	mcv	904	5.3%	Mean cell volume: SI (fL)
mcpsi	mcpsi	905	5.3%	Mean cell hemoglobin: SI (pg)
mhp	mhp	905	5.3%	Mean cell hemoglobin concentration
rwp	rdw	903	5.3%	Red cell distribution width (%)
plp	plp	906	5.3%	Platelet count
dwp	pdw	988	5.8%	Platelet distribution width (%)
grpdif	grpdif	12358	72.6%	Segment neutrophil(percent of 100 cells)
lmpdif	lmpdif	12358	72.6%	Lymphocytes (percent of 100 cells)
mopdif	mopdif	12358	72.6%	Monocytes (percent of 100 cells)
eop	eop	12358	72.6%	Eosinophils (percent of 100 cells)
bop	bop	12358	72.6%	Basophils (percent of 100 cells)
blp	blp	12358	72.6%	Blasts (percent of 100 cells)
prp	prp	12358	72.6%	Promyelocytes (percent of 100 cells)
mep	mep	12358	72.6%	Metamyelocytes (percent of 100 cells)
mlp	mlp	12358	72.6%	Myelocytes (percent of 100 cells)
bap	bap	12358	72.6%	Bands (percent of 100 cells)
lap	lap	12358	72.6%	Atyp lymphocytes (percent of 100 cells)
anp	anp	12358	72.6%	Anisocytosis (variation of cell size)
bsp	bsp	12358	72.6%	Basophilic stippling
hzp	hzp	12358	72.6%	Hypochromia (stain intensity of cell)
pkp	pkp	12358	72.6%	Poikilocytosis (cell shape variation)
pop	pop	12358	72.6%	Polychromatophilia(bluish color of cell)
mrp	mrp	12358	72.6%	Macrocytosis (large cell prevalence)
mip	mip	12358	72.6%	Microcytosis (small cell prevalence)
sip	sip	12358	72.6%	Sickle cells
shp	shp	12358	72.6%	Spherocytosis
ttp	ttp	12358	72.6%	Target cells
txp	txp	12358	72.6%	Toxic granulation
vup	vup	12358	72.6%	Vacuolated cells
pbp	pbp	749	4.4%	Lead (ug/dL)
epp	epp	791	4.6%	Erythrocyte protoporphyrin (ug/dL)
fep	fep	881	5.2%	Serum iron (ug/dL)
tip	tip	906	5.3%	Serum TIBC (ug/dL)

pxp	pxp	925	5.4%	Serum transferrin saturation (%)
frp	frp	916	5.4%	Serum ferritin (ng/mL)
fop	fop	909	5.3%	Serum folate (ng/mL)
rbp	rbp	1241	7.3%	RBC folate (ng/mL)
vbp	vbp	8803	51.7%	Serum vitamin B12 (pg/mL)
vcp	vcp	1845	10.8%	Serum vitamin C (mg/dL)
icpsi	icpsi	2870	16.9%	Serum normalized calcium: SI (mmol/L)
capsi	capsi	1431	8.4%	Serum total calcium: SI (mmol/L)
sep	sel	1301	7.6%	Serum selenium (ng/mL)
vap	vap	1065	6.3%	Serum vitamin A (ug/dL)
vep	vep	1065	6.3%	Serum vitamin E (ug/dL)
acp	act	1065	6.3%	Serum alpha carotene (ug/dL)
bcp	bcp	1065	6.3%	Serum beta carotene (ug/dL)
bxp	bcx	1066	6.3%	Serum beta cryptoxanthin (ug/dL)
lup	lup	1065	6.3%	Serum lutein/zeaxanthin (ug/dL)
lyp	lyc	1065	6.3%	Serum lycopene (ug/dL)
rep	rep	1065	6.3%	Serum sum retinyl esters (ug/dL)
cop	cot	1497	8.8%	Serum cotinine (ng/mL)
tcp	tcp	968	5.7%	Serum cholesterol (mg/dL)
tgp	tgp	1004	5.9%	Serum triglycerides (mg/dL)
lcp	lcp	10234	60.1%	Serum LDL cholesterol (mg/dL)
hdp	hdp	1079	6.3%	Serum HDL cholesterol (mg/dL)
aap	aap	9249	54.3%	Serum apolipoprotein AI (mg/dL)
abp	abp	9232	54.2%	Serum apolipoprotein B (mg/dL)
lpp	lpp	8813	51.7%	Serum lipoprotein(a) (mg/dL)
fhpsi	fhpsi	13908	81.7%	Serum FSH: SI (IU/L)
lhpsi	lhpsi	13910	81.7%	Serum luteinizing hormone: SI (IU/L)
fbp	fbp	7680	45.1%	Plasma fibrinogen (mg/dL)
crp	crp	1095	6.4%	Serum C-reactive protein (mg/dL)
napsi	napsi	1192	7.0%	Serum sodium: SI (mmol/L)
skpsi	skpsi	1192	7.0%	Serum potassium: SI (mmol/L)
clpsi	scl	1192	7.0%	Serum chloride: SI (mmol/L)
c3psi	c3psi	1194	7.0%	Serum bicarbonate: SI (mmol/L)
scp	scp	1193	7.0%	Serum total calcium (mg/dL)
psp	psp	1192	7.0%	Serum phosphorus (mg/dL)
uap	sua	1192	7.0%	Serum uric acid (mg/dL)
sgp	glu	1196	7.0%	Serum glucose (mg/dL)
bup	bun	1192	7.0%	Serum blood urea nitrogen (mg/dL)
tbp	tbp	1192	7.0%	Serum total bilirubin (mg/dL)
cep	cre	1193	7.0%	Serum creatinine (mg/dL)
sfp	sfp	5150	30.2%	Serum iron (ug/dL)
chp	chp	1194	7.0%	Serum cholesterol (mg/dL)
trp	trp	5150	30.2%	Serum triglycerides (mg/dL)
aspsi	aspsi	1192	7.0%	Aspartate aminotransferase: SI(U/L)
atpsi	atpsi	1192	7.0%	Alanine aminotransferase: SI (U/L)
ggpsi	ggpsi	4716	27.7%	Gamma glutamyl transferase: SI(U/L)
ldpsi	ldh	1193	7.0%	Serum lactate dehydrogenase: SI (U/L)
appsi	aph	1194	7.0%	Serum alkaline phosphatase: SI (U/L)

tpp	pro	1192	7.0%	Serum total protein (g/dL)
amp	alb	1192	7.0%	Serum albumin (g/dL)
gbp	gbp	5150	30.2%	Serum globulin (g/dL)
ospsi	ospsi	5150	30.2%	Serum osmolality: SI (mmol/Kg)

11.2 Cox model coefficients

A summary of the coefficient estimates from the selected Cox model of section 6.1.1.

Effect	Coefficient	exp(coef)	se(coef)	Wald Statistic (z)	Pr(>z)
age	0.05	1.05	0.01	7.02	0.00
I(age^2)	0.00	1.00	0.00	4.75	0.00
sex2	-0.19	0.83	0.04	-5.32	0.00
bmi	0.00	1.00	0.00	-1.09	0.28
nimp1	0.40	1.49	0.04	11.05	0.00
nimp2	0.67	1.95	0.05	12.85	0.00
nimp3	0.81	2.25	0.08	9.57	0.00
nimp4	0.94	2.57	0.14	6.66	0.00
smokerstatusNS	-0.16	0.85	0.04	-4.43	0.00
smokerstatusSM	0.35	1.42	0.05	7.45	0.00
wbc	0.02	1.02	0.01	2.57	0.01
rdw	0.12	1.13	0.01	8.94	0.00
pdw	0.07	1.07	0.02	2.88	0.00
pbp	0.02	1.02	0.00	4.70	0.00
act	-0.02	0.99	0.00	-3.75	0.00
bcx	0.00	1.00	0.00	-1.07	0.28
lyc	0.00	1.00	0.00	-2.88	0.00
cot	0.00	1.00	0.00	2.43	0.01
sel	0.00	1.00	0.00	-3.59	0.00
crp	-0.01	0.99	0.02	-0.45	0.65
scl	-0.01	0.99	0.00	-2.67	0.01
glu	0.00	1.00	0.00	7.44	0.00
cre	0.11	1.11	0.04	3.01	0.00
ldh	0.00	1.00	0.00	8.53	0.00
aph	0.00	1.00	0.00	4.13	0.00
alb	-0.41	0.66	0.05	-8.05	0.00
grp^2	0.00	1.00	0.00	4.18	0.00
mcv^2	0.00	1.00	0.00	0.92	0.36
sua^2	0.01	1.01	0.00	3.62	0.00
bun^2	0.00	1.00	0.00	3.90	0.00
pro^2	0.06	1.06	0.03	1.93	0.05
grp	-0.04	0.96	0.01	-3.51	0.00
mcv	-0.01	0.99	0.04	-0.33	0.74
sua	-0.12	0.88	0.05	-2.54	0.01
bun	-0.01	0.99	0.01	-1.46	0.14
pro	-0.67	0.51	0.45	-1.49	0.14

Concordance	0.873
R-square	0.502

11.3 Cox model coefficients – Healthy only

Effect	Coefficient	exp(coef)	se(coef)	Wald Statistic (z)	Pr(>z)
age	0.04	1.04	0.01	4.45	0.00
I(age^2)	0.00	1.00	0.00	5.70	0.00
sex2	-0.19	0.82	0.05	-3.86	0.00
bmi	0.00	1.00	0.00	0.64	0.52
smokerstatusNS	-0.14	0.87	0.05	-2.62	0.01
smokerstatusSM	0.42	1.52	0.06	6.54	0.00
wbc	0.01	1.01	0.01	1.04	0.30
rdw	0.14	1.15	0.02	7.49	0.00
pdw	0.07	1.07	0.03	2.31	0.02
pbp	0.02	1.02	0.01	3.52	0.00
act	-0.01	0.99	0.01	-2.09	0.04
bcx	-0.01	0.99	0.00	-2.46	0.01
lyc	-0.01	0.99	0.00	-2.61	0.01
cot	0.00	1.00	0.00	3.25	0.00
sel	0.00	1.00	0.00	-2.90	0.00
crp	-0.01	0.99	0.02	-0.47	0.64
scl	-0.01	0.99	0.01	-1.53	0.13
glu	0.00	1.00	0.00	3.36	0.00
cre	0.12	1.13	0.07	1.65	0.10
ldh	0.00	1.00	0.00	6.84	0.00
aph	0.00	1.00	0.00	3.01	0.00
alb	-0.34	0.71	0.07	-4.77	0.00
grp^2	0.00	1.00	0.00	2.80	0.01
mcv^2	0.00	1.00	0.00	0.35	0.72
sua^2	0.00	1.01	0.01	0.87	0.39
bun^2	0.00	1.00	0.00	5.45	0.00
pro^2	0.10	1.11	0.05	2.22	0.03
grp	-0.03	0.97	0.01	-2.35	0.02
mcv	0.01	1.01	0.05	0.19	0.85
sua	-0.03	0.97	0.07	-0.37	0.71
bun	-0.03	0.98	0.01	-3.22	0.00
pro	-1.37	0.26	0.71	-1.93	0.05

Concordance	0.865
R-square	0.384

11.4 Assumed mortality rates

These are the 5-year mortality assumptions used for calculating standardised mortality ratios (rates are calculated for every year of age but this table just shows the rates at every 5-year interval). These rates should be interpreted as the probability that someone aged x at their last birthday will die in the next 5 years.

Age	Male	Female
20	0.014923	0.007811
25	0.010620	0.003718
30	0.009984	0.002435
35	0.013008	0.004005
40	0.019841	0.008456
45	0.030162	0.015601
50	0.043277	0.026252
55	0.062583	0.040909
60	0.090848	0.055654
65	0.135860	0.081825
70	0.201205	0.122949
75	0.288822	0.186747
80	0.398364	0.271715

11.5 Logistic regression coefficients

These are the coefficient estimates from the logistic regression of section 7.3.2.

Effect	Estimate	Std. Error	z	Pr(>|z|)
(Intercept)	-5.03	1.84	-2.73	0.006294
age	0.02	0.01	1.81	0.069805
sex2	-0.61	0.05	-11.44	<2.00E-16
I(age^2)	0.00	0.00	5.24	1.64E-07
rdw	1.14	0.14	8.35	<2.00E-16
I(rdw^2)	-0.03	0.00	-6.43	1.26E-10
wbc	0.03	0.01	3.34	0.000831
crp	0.20	0.04	5.29	1.20E-07
I(crp^2)	-0.01	0.00	-3.10	0.001969
cre	0.01	0.00	7.32	2.54E-13
I(cre^2)	0.00	0.00	-4.66	3.20E-06
alb	-3.77	0.76	-4.94	7.90E-07
I(alb^2)	0.40	0.10	4.15	3.39E-05
rbc	-0.45	0.05	-9.12	<2.00E-16
lmp	-0.02	0.00	-7.40	1.42E-13
mpv	0.01	0.02	0.61	0.543306

Null deviance: 18243 on 33398 degrees of freedom
Residual deviance: 12351 on 33383 degrees of freedom