COMPUTATIONAL BIOLOGY APPLICATIONS IN THE STUDY OF COMPLEX SYSTEMS

by

LAURA BRAVO MERODIO

A thesis submitted to the University of Birmingham for the degree of
DOCTOR OF PHILOSOPHY

Institute of Cancer and Genomic Sciences
Medical and Dental School
University of Birmingham
March 2023

UNIVERSITY OF
BIRMINGHAM

University of Birmingham Research Archive

e-theses repository

This unpublished thesis/dissertation is copyright of the author and/or third parties. The intellectual property rights of the author or third parties in respect of this work are as defined by The Copyright Designs and Patents Act 1988 or as modified by any successor legislation.

Any use made of information contained in this thesis/dissertation must be in accordance with that legislation and must be properly acknowledged. Further distribution or reproduction in any format is prohibited without the permission of the copyright holder.

Abstract

In biomedicine, the advent of digitalization and big improvements in computing power and high throughput technologies has yielded an unprecedented amount of data. To harness this data's full potential, research has become increasingly computational, with core tools of data science such as machine learning required to decipher patterns, help extract meaning and uncover underlying connections. In this work, we have explored key leading areas of computational research, from discovery science to translational medicine and complex system studies. First, a machine learning pipeline was developed to inform bioinformatic research. After validation, it was applied on a novel dataset generating insight onto the predictive power of immune features in ultra-early trauma injury, generating leads of possible biomarkers associated with the development of MultiOrgan Dysfunction. Then, in order to make our pipeline accessible, an interactive, simple, free and open-source supervised machine learning webtool was developed. Adapting this same framework to the realm of clinical translation, we then deployed two prognostic models for decision support in surgeries during the COVID-19 pandemic. This was done in collaboration with the NIHR surgical team. Lastly, we explored systems biology approaches by studying the complexity behind ageing and health to disease transitions. For this, we leveraged the UK Biobank data as a rich source of deeply phenotyped information. By calculating biological age (PhenoAge) longitudinally for nearly 400,000 participants, we identified four distinct ageing categories ranging from healthy to unhealthy ageing. These different trajectories were characterised by their chronic diseases and genetic makeup, revealing a strong association of metabolic dysfunction with unhealthier phenotypes and immune-related signals for healthier ones. Also, strong opposite-effect associations of longevity-related variants were found, with novel regulatory elements postulated as possible drivers of unhealthy phenotypes, opening new avenues for future study. Overall, our findings highlight the crucial role of computational methods in biomedicine and their potential to transform clinical practice.

Acknowledgments

First, I would like to express my gratitude to my supervisors Dr. Animesh Acharjee, Prof. Janet Lord, and Prof. Georgios Gkoutos. Without their guidance, support, and encouragement this PhD would not be possible. Animesh, I value incredibly your drive and “guide by example” principles and I am extremely grateful for all your mentorship advice. George, your constant support and good-humour is a source of calm that has uplifted me in this journey many times. Janet, you are a huge inspiration and I have really valued the opportunity of working with you.

I would also like to acknowledge the funding provided by the Wellcome Trust Four-Year PhD Studentship in Basic Science, Mechanism of Inflammatory Disease (MIDAS) Doctoral program under grant number 215182/Z/19/Z. Through this program, I have had the opportunity to engage in exciting research experiences and meet the most insightful cohort of scientists. I would especially like to celebrate meeting Ailbhe and Claire, friends who have made this journey abroad feel more like home.

Also, my time in Birmingham has been greatly enriched by all the friendship and support of the Spanish Society in the UK team. I am grateful for the joy and wonderful experiences shared and look forward to many more. Claudia, Simon, Jorge, Miguel, Cyntia, Asier let's continue the fun. Also, to all the new special people that joined this year, the future is exciting.

I am also thankful to my fellow PhD students Victor and Dom, and the entire Gkoutos group, for making me miss office days and lunch at 12. From our trip to Nottingham to bike rides and

pints, and even pandemic zooms, meeting you all has definitely been one of the key highlights of my time here. I would also like to acknowledge John, Luke, Andreas, Saisakul and all other wonderful senior leaders in the group, thank you for your support and friendship.

Now, I want to focus on the solid support team back home. I feel extremely lucky to count with the unwavering support of my school friends Elisa and Irene, Stone (right here with me now!), Maria, Carmen, Almu and many more. You are simply the best. Also uni friends, you inspire me everyday and have pushed me to dream bigger and brighter.

Finally, I want to express my heartfelt gratitude to my family and most notably to Carmen, Juan Carlos, and Alex, for being the main pillars in my life. Your encouragement, love, and belief in me is unmatched and I am truly blessed to have you.

List of Publications

Academic Journal Papers

1. **Bravo-Merodio, L,** Acharjee, A, Hazeldine, J, Bentley, C, Foster, M, Gkoutos, GV, & Lord, JM. (2019). Machine learning for the detection of early immunological markers as predictors of multi-organ dysfunction. *Scientific Data*, *6*(1), 1–10.

2. **Bravo-Merodio, L,** Williams, JA, Gkoutos, GV, & Acharjee, A. (2019). -omics biomarker identification pipeline for translational medicine. *Journal of Translational Medicine*, *17*(1), 1–10.

3. Aziz, F, Acharjee, A, Williams, JA, Russ, D, **Bravo-Merodio, L**, & Gkoutos, GV. (2020). Biomarker prioritisation and power estimation using ensemble gene regulatory network inference. *International Journal of Molecular Sciences*, *21*(21), 7886.

4. Aziz, F, Cardoso, VR, **Bravo-Merodio, L**, Russ, D, Pendleton, SC, Williams, JA, Acharjee, A, & Gkoutos, GV. (2021). Multimorbidity prediction using link prediction. *Scientific Reports*, *11*(1), 1–11.

5. Williams, JA, Russ, D, **Bravo-Merodio, L,** Cardoso, VR, Pendleton, SC, Aziz, F, Acharjee, A, & Gkoutos, GV. (2021). A causal web between chronotype and metabolic health traits. *Genes*, *12*(7), 1029.

6**. Bravo-Merodio, L,** & Collaborative, C. (2021). Machine learning risk prediction of mortality for patients undergoing surgery with perioperative sars-cov-2: The covidsurg mortality score. *The British Journal of Surgery*, *108*(11), 1274.

7. Russ, D, Williams, JA, Cardoso, VR, **Bravo-Merodio, L**, Pendleton, SC, Aziz, F, Acharjee, A, & Gkoutos, GV. (2022). Evaluating the detection ability of a range of epistasis

detection methods on simulated data for pure and impure epistatic models. *Plos One*, *17*(2), e0263390.

8. Bentley C, Hazeldine J, **Bravo-Merodio, L**, Taylor AE, Gilligan LC, Shaheen F, Acharjee A, Gkoutos G, Foster MA, Arlt W, Lord JM. (2023).The ultra-acute steroid response to traumatic injury: a cohort study. In *European Journal of Endocrinology*.

Book chapters

1. **Bravo-Merodio, L,** Acharjee, A, Russ, D, Bisht, V, Williams, JA, Tsaprouni, LG, & Gkoutos, GV. (2021). Translational biomarkers in the era of precision medicine. In *Advances in clinical chemistry* (Vol. 102, pp. 191–232). Elsevier.

Unpublished working papers

1. **Bravo-Merodio, L**, & Collaborative, C. (n.d.-). Risk of pulmonary complications in adults undergoing elective surgery: Development and validation of a prognostic model to support safe global recovery of elective surgery. In *Lancet Digital Health*.

2. **Bravo-Merodio, L,** Acharjee, A, Gkoutos, GV, & Lord, JM. (n.d.-). Longitudinal analysis of biological age (PhenoAge) in UKBiobank participants reveals key factors driving ageing trajectories. In *Nature Ageing.*

3. **Bravo-Merodio, L,** Williams, JA Gkoutos, GV & Magda. Exploring the combinatorial effects genetics, fronto-parietal networks, and education on cognitive outcomes.

Table of Contents

1 Chapter 1: General Introduction 1
1.1 Machine Learning 2
1.1.1 Model types 3
1.1.2 Model selection and assessment 5
1.1.2.1 Internal Validation: train-test split and resampling methods 7
1.1.2.2 External Validation 8
1.1.2.3 Discrimination and calibration 9
1.1.3 State of the art developments 11
1.1.3.1 Model explainability 11
1.1.3.2 Ethics and regulations 12
1.2 Systems Biology 13
1.2.1 Disease and multimorbidity 14
1.2.2 Describing healthy phenotypes 17
1.2.3 Health to disease transitions 17
1.3 Thesis aims and objectives 18
2 Chapter 2: Biomarker discovery and pipeline development 20
2.1 Background and chapter overview 21
2.2 Pipeline development 25
2.3 Pipeline validation 28
2.3.1 Lipidomics 28
2.3.2 Transcriptomics 30
2.4 Use case: Golden Hour dataset 35
2.5 Web app 42
2.5.1 Module 1: Visualisation 44
2.5.2 Module 2: Feature selection 44
2.5.3 Module 3: Final model building 45
2.5.4 Module 4: Analysis report 46
2.6 Conclusion and discussion 46
3 Chapter 3: Clinical translation and CovidSurg 49
3.1 Background and chapter overview 50
3.2 Machine learning risk prediction of mortality for patients undergoing surgery with perioperative SARS-COV-2: The CovidSurg mortality score 51
3.2.1 Summary 51
3.2.2 Methods and results 52
3.2.2.1 Data: Patient characteristics and outcomes 52
3.2.2.2 Feature selection 56
3.2.2.3 Model building and validation 58
3.3 Risk of pulmonary complications in adults undergoing elective surgery: Development and validation of a prognostic model to support safe global recovery of elective surgery (GlobalSurg-Pulmonary Score) 67
3.3.1 Summary 67
3.3.2 Methods and results 68

3.3.2.1 Data: Patient characteristics and outcomes 69
3.3.2.2 Model selection and building 73
3.3.2.3 Model validation 78
3.4 Conclusion and discussion 84
4 Chapter 4: Complex systems and ageing trajectories 88
4.1 Background and chapter overview 89
4.2 Summary 91
4.3 Methods 93
4.3.1 UK Biobank data 94
4.3.2 PhenoAge calculation 95
4.3.3 Disease studies: prevalence, co-occurrence calculations and correlation 97
4.3.4 Genome Wide Association Studies (GWAS) 97
4.3.5 Gene annotation 98
4.3.5.1 Proximity analysis 98
4.3.5.2 eQTLs 99
4.3.6 Functional enrichment 99
4.3.7 Association with ageing 100
4.4 Results 100
4.4.1 UKBB statistics and PhenoAge calculation. 100
4.4.2 Characterisation of ageing trajectory categories using lifestyle factors and health outcomes. 105
4.4.3 Characterisation of ageing trajectory categories using disease patterns 107
4.4.4 Genomic characterization of ageing trajectory categories. 112
4.4.5 Gene annotation 117
4.4.6 Functional enrichment 117
4.4.7 Ageing databases 119
4.5 Conclusion and discussion 121
5 Chapter 5: Discussion and thesis conclusions 129
5.1 Overview of major findings 130
5.2 Contributions to the literature 135
5.3 Limitations 136
5.4 Planned future work 138
5.4.1 Chapter 2 138
5.4.2 Chapter 3 139
5.4.3 Chapter 4 139
5.4.3.1 PheWAS 139
5.4.3.2 Disease sequences 140
5.4.3.3 Lifestyle and transitions 140
5.4.3.4 Knowledge graphs 141
5.5 Conclusions 142
Appendix 144
Appendix 1: Web app report of key results 145
Appendix 2: Read codes 149

Appendix 3: ICD-10 categorisation 153
Appendix 4: Moving window imputation 161
Appendix 5: Disease legend 162
Appendix 6: Summary Statistics 164
Appendix 7: Longevity associated variants 170
6 References **171**

List of Figures

Figure 1-1. Bias-variance trade-off 6
Figure 1-2. Confusion matrix (grey) and associated statistics 11
Figure 2-1. Machine learning pipeline structure 25
Figure 2-2. Final modelling results of lipids SM32.1, SM39.1 and SM36.2 30
Figure 2-3. Feature selection pipeline 33
Figure 2-4. Flowchart of the Golden Hour case-study 35
Figure 2-5. Univariate and multivariable performance of final selected features 38
Figure 2-6. Final model and selected immune markers 41
Figure 2-7. Diagram of all modules available in the RShiny supervised machine learning webapp 43
Figure 3-1. Workflow description 53
Figure 3-2. Variable importance and feature selection 57
Figure 3-3. Model performance and evaluation 59
Figure 3-4. Discrimination evaluation results 63
Figure 3-5. Calibration evaluation results 64
Figure 3-6. Tile plot 66
Figure 3-7. Workflow description 70
Figure 3-8. Comparison of LASSO and XGBoost algorithms during model selection 75
Figure 3-9. Discrimination and calibration performance for internal validation and external validation datasets 78
Figure 3-10. Subgroup analysis 80
Figure 3-11. Predicted score in the whole derivation dataset (SurgWeek) 83
Figure 4-1. Study overview and methodology with key steps highlighted 93
Figure 4-2. PhenoAge measurements, trajectories and categories 104
Figure 4-3. Frequency of the number of disease categories per participant 110
Figure 4-4. Relative Risk between disease categories for the four ageing phenotypes 111
Figure 4-5. GWAS results for the four ageing categories 114
Figure 4-6. Venn diagrams showing significant genetic associations in URU and HRH categories 116
Figure 4-7. FUMA gene enrichment results 118
Figure 4-8. Association of annotated genes with key ageing databases 120

List of Tables

Table 2-1. Selected lipid biomarkers 29
Table 3-1. Patient, disease and operation factors of patients included in the derivation and validation sets 55
Table 3-2. Top performing models 60
Table 3-3. Final logistic model coefficients 61
Table 3-4. Performance metrics of model probability output at different cut-off values in both derivation and validation cohorts 65
Table 3-5. Comparison of derivation and validation cohort datasets 72
Table 3-6. Model description 77
Table 3-7. Prognostic test accuracy in the validation set 82
Table 4-1. GP records data on the 9 markers extracted to calculate PhenoAge 102
Table 4-2. Lifestyle and health outcomes data for the different ageing categories 106
Table 4-3. Disease prevalence of 60 chronic disease categories 108

1 Chapter 1: General Introduction

Content from this chapter has been published:

> **Bravo-Merodio, L,** Acharjee, A, Russ, D, Bisht, V, Williams, JA, Tsaprouni, LG, & Gkoutos, GV. (2021). Translational biomarkers in the era of precision medicine. In *Advances in clinical chemistry* (Vol. 102, pp. 191–232). Elsevier.

From the start of the 21st century, high-throughput technologies have dramatically transformed biomedical and healthcare research by generating an immense volume of data. This surge in data has been met with simultaneous advancements in computational and statistical methodologies, such as machine learning (ML) and artificial intelligence (AI), bolstered by significant improvements in computing power (Berger, Peng, and Singh 2013). As a result, extracting meaningful patterns from these massive datasets has become increasingly feasible. In this chapter, we will discuss data science tools we have applied throughout the thesis, first presenting machine learning, which promises to revolutionize research, medicine and healthcare as we know it (Shah et al. 2019), and then continuing with a presentation of complex system studies. Complex systems focus on holistic approaches to decipher the complexity of human physiology, aiming to quantify health and demystify disease (Hood and Flores 2012).

1.1 Machine Learning

The term ‘machine learning’ is thought to have been first coined in 1956 to describe a subfield of artificial intelligence focused on building algorithms that ‘learn’ (Finlayson, Beam, and van Smeden 2023). The learning behind these algorithms is based on statistical models which improve from experience, through iterative exposure to data and without being explicitly programmed to do so (Kohavi and Provost, 1998).

In general terms, machine learning bridges both the fields of statistics and computer science and consists of models that are fit to data, learning underlying patterns. These models are then applied to new unseen data where they predict based on the patterns learnt. No a priori information is required, and a plethora of algorithms exist (Hastie, Friedman, and Tibshirani

2001). These algorithms can be used for three key types of analysis depending on the type of predictions performed: supervised, unsupervised, or semi supervised learning. In supervised learning, the goal is to predict the value of an outcome measure based on a number of input measures. Two types of supervised learning methods exist based on the type of data of the outcome, regression (if continuous i.e height) or classification (if discrete i.e presence or absence of disease)). In this work, we will explore supervised learning with Chapters 2 and 3 mainly focused on building supervised classification models. In unsupervised learning, there is no outcome measure, and the goal is to describe the associations and patterns among a set of input measures. These input measures are also known as independent variables, covariates, or features and outcomes as label, target, dependent variable or response.

1.1.1 Model types

As mentioned before many different ML algorithms exist. In this section we will do a brief overview of well-known algorithms we will apply in Chapters 2 and 3. These include: generalized linear models (GLMs), decision tress and neural nets.

Generalized linear models (GLMs) are a broad family of models that can be applied to classification and regression problems. Two well known types of GLMs are linear (Equation 1) and logistic regression (Equation 2) algorithms where, considering $x_1 \dots x_p$ as p number of predictor variables (features) and $h_\beta(x)$ as the model built, the aim is to match the predicted output $\hat{y}$ to the real outcome variable y as closely as possible by optimizing β, the parameters or weights parametrising the space mapping x (features) to y (outcome label). This will be achieved through the optimization of a loss function (Hastie, Friedman, and Tibshirani 2001).

$$h_{\beta}(x) \ = \hat{y} \ = \ \hat{\beta}_0 + \hat{\beta}_1 x_1 + \hat{\beta}_2 x_2 + \ldots \hat{\beta}_p x_p$$

Equation 1

$$h_{\beta}(x) = \hat{y} = \ g(\beta x) = \ \frac{1}{1+ e^{-\beta x}}$$

Equation 2

In Equation 2, $g(x)$ corresponds to the sigmoid function, bounding values between 0 and 1 and mimicking the discrete state we are predicting in classification scenarios.

Other examples included in this work are decision tree algorithms. Simple, easy to interpret and low-demanding computationally, they are built through recursive partitioning, splitting the feature space into a set of rectangles, and then fitting a simple model in each one (e.g measurements satisfying x < 5). Different splitting criteria amounts to different decision tree algorithms, with some popular derivations being ensemble methods such as Random Forest (RF) (Ho 1995) or gradient boosting decision trees such as eXtreme Gradient Boosting (XGBoost) (T. Chen and Guestrin 2016). In these decision tree derivations, multiple decision trees are combined to obtain a better predictive performance, with the key differences between them being the different ways in which they are merged. In Random Forest the merge occurs through averaging (if regression) or majority vote (if classification), whereas in boosting the merge is weighted, with those splits that predict well rewarded and those that do not, ‘punished’. XGBoost is a particularly popular gradient boosting decision tree, given it has been designed to scale beyond billions of examples but using far fewer resources than existing systems, demonstrating very good performance in ML problems. For a more detailed analysis on decision tree algorithms please refer to (Hastie, Tibshirani, and Friedman 2009).

One last algorithm type to mention is neural networks (NN), a term which has evolved to encompass a large class of models and learning methods all under the umbrella of Deep Learning (DL). In its simplest form (also known as single hidden layer back-propagation network), it refers to a non-linear model that consists of a two-stage regression or classification model and is typically represented by a network diagram. Although NNs, were first described in the 1950s in the context of the Rosenblatt´s perceptron (Rajkomar, Dean, and Kohane 2019), DL has become very popular in the last 10 years, with for instance convolutional neural nets (CNNs) being a key recent breakthrough (LeCun, Bengio, and Hinton 2015) and leading key improvements in the field of computer vision, image caption, speech recognition, natural language translation, robotics, or even self-driving cars.

1.1.2 Model selection and assessment

As seen above, a variety of algorithms exist, each with different characteristics and capabilities. To select the best for our data analysis, an understanding of which model would offer the best performance is required. To measure performance, we will assess the generalization error, a measure of how accurately and reliably the extracted patterns generalize to new observations, the ultimate goal in ML. In the case of supervised learning, this would amount to evaluating the capability of a model to accurately predict the outcome label in new data not used for training. In order to accurately evaluate model performance, two fundamental principles can help us guide our choice, the “No Free Lunch Theorem” (Wolpert and Macready 1997) and the bias-variance trade-off (Figure 1-1).

On one hand, the “No Free Lunch Theorem” (Wolpert and Macready 1997) acknowledges that no algorithm is inherently superior to any other, with different models performing best in different contexts. Therefore, when choosing models, a sound understanding of the data on which the model will operate and the circumstance of its intended use are required, with the

type of problem at hand, prior data distribution, amount of training data and computational resources being the key drivers of algorithmic choice. On the other hand, the bias-variance trade-off (Figure 1-1), helps us critically assess model performance and is a guiding principle for machine learning frameworks such as the train-test split, K-fold cross validation or bootstrapping which we have included throughout our work.

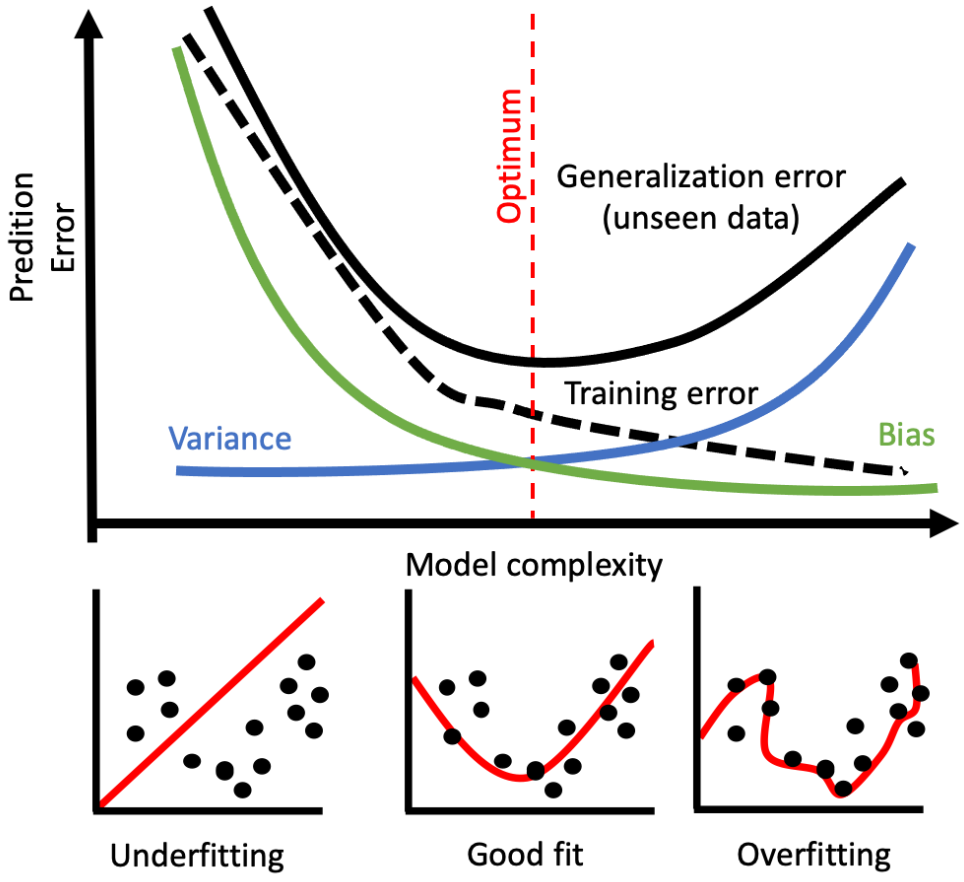


Figure 1-1. Bias-variance trade-off
Our objective when building a ML model is to minimize the generalization error (find the model found in the optimum point (red line)) In order to do so, we use the training error as reference (black line), but model complexity has to be acknowledged. Bias and variance define model complexity where, high bias models (green) yield high training and generalization errors in a phenomenon known as underfitting, whereas high variance models (blue) yield very low prediction errors but high generalization ones. This is termed overfitting, where the model is adapting itself too closely to the training data and therefore not generalizing well to new unseen one as, not only has it captured the underlying patterns we

are interested in but probably too all underlying noise. Below, polynomial models exemplifying a low variance/high bias model (low polynomial degree), an optimum model and a high variance/low bias model (high polynomial degree).

The best estimate of the generalization error we can have is the training error (predictive performance in the dataset used to fit the data) but this is not a perfect proxy given it does not fully account for model complexity. Model complexity is built from two metrics, bias and variance, where a high variance model fits the training data very strongly, leading to overfitting and high bias approximates data with a model that is too simple, leading to underfitting. In both cases, poor generalization performance on the new data will result. Therefore, in order to obtain an optimum model, we have to not only consider the training error as proxy of the generalization error, but acknowledge this extra model complexity by testing models in unseen data and critically appraise the change in performance. This is the reason behind internal validation methods such as train-test splits or resampling which we will now explore below:

1.1.2.1 Internal Validation: train-test split and resampling methods

Ideally, if we count with lots of data, we will split our data into chunks that the training does not see and leave one as this unseen independent dataset to measure performance in. Data would be fit in a training set, with any extra model selection phase done using extra unseen testing data. Then, after building this final model, the generalization error would be assessed in a third, unseen, independent validation set also known as external validation set. But, given data is a luxury, and in theory we would want to maximise its use for model building, resampling, or partitioning the data through bootstrapping and cross validation methods, can be used instead. In K-fold cross validation part of the available data is used to fit the model, and a different part to test it. More specifically, data is split into K roughly equal-sized parts and the model is fitted to the other K −1 parts of the data, evaluating the prediction error of

the fitted model when predicting the kth part of the data. This is done for k = 1, 2, . . . ,K and the K estimates of prediction error are combined. Regarding the bootstrap, it is a widely used tool for assessing statistical accuracy and consists on randomly sampling the data with replacement to create multiple datasets of the same size as the original data. Each dataset is used to train a model, and the results are combined to obtain a more accurate estimate of the model's performance.Once again, for a more detailed analysis refer to (Hastie, Tibshirani, and Friedman 2009).

Something key to consider in ML, is that this is a quite novel and very rapidly evolving field, where theory, with a sound mathematical foundation, lags behind its practical implementation (C. Zhang et al. 2016). As such, tools that are consolidated in practice like the above mentioned cross-validation have been recently questioned (Bates, Hastie, and Tibshirani 2022) with constant updates in methodology required. Another highly debated practice in ML is that of having proper external validation strategies as summarized in the next section.

1.1.2.2 External Validation

In high stakes industries such as healthcare, where model outcomes may influence decisions as important as diagnosis or treatment options, extra caution for model building is being demanded (Van Calster et al. 2023). These extra efforts include assessing this generalization error through internal validation and careful external validation (extra exteneral dataset) too. This is key in order to acknowledge possible “data shift’ issues and transportability of a model before recommended for widespread clinical implementation. Transportability is assessed by understanding why a machine-learning model might underperform when assessed in external validation set. As seen before, this might be due to issues with the model complexity itself (underfitting and overfitting) or because of a mismatch between the data set with which it was developed and the data on which it is deployed. For this, a proper assessment on the

model's sensitivity to changes in the distribution or statistical properties of the data, such as changes in the feature distributions, the sampling process, or the context in which the model is deployed is required. Such an analysis will help us understand too if the model has been trained on biased or unrepresentative data, which can limit its generalization ability to new data. Regardless, it is important to acknowledge too, that an exhaustive external validation is not possible as all datasets will have specific internal biases (Van Calster et al. 2023), therefore a key feature when deploying ML models lies in carefully monitoring the performance of the model in the real-world deployment context and update the model as necessary to account for changes in the data distribution or context. In Chapter 3, when exploring clinical validity of our prognostic models, addressing transportability becomes a critical part of our analysis.

1.1.2.3 Discrimination and calibration

Chapters 2 and 3 will mainly focus on supervised classification strategies, where we are predicting a categorical outcome (e.g disease yes or no). To measure performance, we use Area Under the Receiver Operating Characteristic curve (AUROC) as the main scoring rule, also known as the concordance statistic or c-statistic. AUROC reports on discrimination, which describes the models' ability to rank samples based on the outputed prediction, in other words, models should give higher prediction estimates for samples with the event than for samples without the event. This corresponds to plotting sensitivity versus 1- specificity for different thresholds (Figure 1-2). An AUROC of 0.5 indicates a completely random binary classifier, while an AUROC of 1.0 denotes perfect classifier performance (Fawcett 2006). Additionally, in our studies (Chapter 2 and Chapter 3), we also include classification performance at one or more probability thresholds through sensitivity, specificity, and (stratum-specific) likelihood ratios which can be extracted from a confusion matrix (Figure 1-2) . Many other scoring rules exist (e.g the popular logarithmic scoring rule) but these have

not been explored in this work. For a full review on performance evaluation metrics check (P.-H. C. Chen, Liu, and Peng 2019).

Another key metric to assess performance is calibration. Calibration measures the agreement between predicted and observed probabilities/risk. It is especially relevant in clinical risk prediction models and medical predictive analytics given that, as the model is used to inform patients, poorly calibrated risk estimates can lead to false expectations with patients and healthcare professionals. What we want to assess is that, among patients with an estimated risk of 10%, 10 in 100 have or develop the event. To evaluate calibration the key metrics used are: calibration in the large (slope and intercept) and the Brier Score.

The Brier Score, measures accuracy of probabilistic predictions without thresholding unlike AUROC, and measures the mean squared difference between the prediction probability assigned to the outcome and the actual outcome. The lower the Brier score the better the predictions are calibrated. Also, calibration can be visualized through smooth nonparametric calibration plots where the relation between the estimated risk (on the x-axis) and the observed proportion of events (y-axis) is plotted (Van Calster et al. 2019) and key measures such as calibration slope and intercept extracted. The calibration curve intercept measures the average predicted risk compared with the overall event rate and has a target value of 0, where negative values suggest overestimation and positive underestimation. The calibration slope evaluates the spread of the estimated risks and has a target value of 1. A slope with values less than 1 suggests that estimated risks are too extreme, i.e., too high for patients who are at high risk and too low for patients who are at low risk. A slope higher than 1 suggests the opposite, i.e., that risk estimates are too moderate.Extra information in calibration can be found (Huang et al. 2020).

	Predicted: Positive	Predicted: Negative	
Truth: Positive	True Positive (TP)	False Negative (FN) Type II Error	**Sensitivity** $\frac{TP}{(TP + FN)}$
Truth: Negative	False Positive (FP) Type I Error	True Negative (TN)	**Specificity** $\frac{TN}{(TN + FP)}$
	Precision/Positive Predicted Value (PPV) $\frac{TP}{(TP + FP)}$	**Negative Predictive Value (NPV)** $\frac{TN}{(TN + FN)}$	**Accuracy** $\frac{TP + TN}{(TP + TN + FP + FN)}$

Figure 1-2. Confusion matrix (grey) and associated statistics

1.1.3 State of the art developments

The fields of AI and ML are in constant development, given the high stakes, scope, and breadth of applications. As novel fields of research emerge, we have aimed to incorporate them into our work.

1.1.3.1 Model explainability

As seen before, there is a wide range of ML algorithms developed. One thing to consider when selecting models is the tradeoff between model interpretation and accuracy. For instance, interpreting large deep neural nets, such as the 17 billion parameter network recently released by Microsoft (‘Turing-NLG: A 17-Billion-Parameter Language Model by Microsoft’ 2020), has been impossible to date. These types of models are commonly known as “black boxes”. However, the performance of such models is at times far superior to other more explainable ones, given the vast amounts of data they are trained upon, rendering them extremely useful. Therefore, trying to understand its behaviour and explain the reasoning

behind its decision-making has opened up a new field of study termed “Explainable AI”, with even quantifications of interpretability to assess AI models proposed (Schmidt and Biessmann 2019). This field has yielded for instance work such as Google DeepMind´s eye-scans diagnosis model (Yim et al. 2020) which can accurately diagnose 50- sight threatening eye diseases whilst delivering a map highlighting features of disease it sees in the different eye tissues. Such development are needed to fuel adoption in high-risk scenarios such as healthcare, helping guide translation into actionable clinical models. A complete review on the state of the art of explainable AI can be found in (Doshi-Velez and Kim 2017). In our work, we have made a great emphasis on including feature importance metrics in our modelling strategies, adapting popular explainable AI methods such as Shapley Additive exPlanations (SHAP) (Rodríguez-Pérez and Bajorath 2020).

1.1.3.2 Ethics and regulations

Another leading topic is that of the underlying ethics of AI systems and guidance in reporting and regulations. Over recent years, tremendous progress has been made in the development of ML applications for healthcare. Regardless, given that this technology is so new and quickly evolving, risks are assessed in retrospective, with the development of proper regulation and assessment lagging behind. Questions such as patient privacy, anonymity retention, data security as well as development of fair, non-biased algorithms (Nordling 2019) present a grand challenge, with key questions regarding how to distinguish between data that is inherently biased still unresolved (A. Zhang et al. 2022). This is highly problematic, with pressing fears that AI can even widen the health gap globally (Blasimme and Vayena 2019) as biased datasets result in algorithms that mirror this bias, thereby preserving inequalities, cutting down diversity and producing flawed results (Leslie et al. 2021).

Regarding regulation, at the moment clinical algorithms are treated by the main regulatory agencies as medical devices (software as a medical device: SaMD) and hence not subject to the rigorous regulation that pharmaceutical companies or other healthcare incumbents have to go through. Luckily though, new standards are constantly being enforced (Collins and Moons 2019). Also, with increasingly lower barriers for adoption, many low quality ML models have been generated recently (Wynants et al. 2020), with high risk of bias through low sample sizes, poor reporting, and over-estimation of predictive performance. Adherence to better standards is urgently needed (Dhiman et al. 2022) with reporting guidelines being key to ascertain the quality of modelling strategies. Examples include SPIRIT-AI, CONSORT-AI checklists for reporting of clinical trials that involve AI methods (Ben-Israel et al. 2020) or TRIPOD and PROBAST guidelines to mitigate biased analysis (Mateen et al. 2020).

1.2 Systems Biology

In biology, research has become increasingly computational, with the birth of bioinformatics (Ouzounis and Valencia 2003) enabling a better understanding of the genetic makeup of diseases (Human Genome Project) (Gibbs 2020), fuelling precision medicine initiatives (Ashley 2016) and providing new insights into the molecular landscape of health and disease (Shen et al. 2023). This has been possible thanks to recent advances in high-throughput technologies and computational power that have allowed us to observe biological systems at molecular resolution, generating a wealth of different ‘omics’ data. ‘Omics’ are becoming essential for basic research, offering a glimpse along the central dogma of biology, with genomics profiling DNA, transcriptomics measuring transcripts or proteomics and metabolomics quantifying proteins and metabolites respectively. Having learnt about machine learning as a computational analysis tool behind understanding patterns in data, we then want to understand how to best study complex biomedical phenomenon. To do so, systems

biology perspectives are required, where biological systems are recognized as complex and requiring of the integration of different data layers and organizational levels for its full understanding (Kitano 2002). For this thesis, we have specifically focused on health to disease transitions as a complex system, considering multimorbidity and ageing as extreme phenotypic expressions (Schultze and Rosenstiel 2018; Davis, Tato, and Furman 2017; A. A. Cohen et al. 2022).

1.2.1 Disease and multimorbidity

Diseases represent non-random malfunctions in human biology, which can provide insights into particularly vulnerable pathways (Meizlish et al. 2021). To shed light on these malfunctions, -omics studies have played a crucial role. The emergence of Next Generation Sequencing (NGS) technologies and innovative bioinformatic approaches—such as Genome Wide-Association Studies (GWAS)— propelled by the Human Genome Project and the International HapMap Project (Gibbs et al. 2003), have unveiled new genetic associations and generated a wealth of novel disease-gene connections. In GWAS, millions of genetic variants are tested for an association with a single or a small number of related outcomes or traits (Altshuler, Daly, and Lander 2008), revealing that most associations of common diseases are with common variants (present in more than 1–5% of the population and with generally small effect sizes). But, these disease-associated variants explain only a small proportion of disease heritability (Manolio et al. 2009), with them being found mainly in regulatory regions, outside protein-coding exons (~90%) (Albert and Kruglyak 2015; B. Li and Ritchie 2021). Because of this, regulation of gene expression levels is now being seen as the primary biological mechanism through which genetic variants affect complex traits, with association studies of intermediate phenotypes such as chromatin marks, methylation status or transcript and protein abundance (L. Chen et al. 2016) hypothesised as key ways in which

to help prioritize likely causal variants and reveal functional effects and underlying mechanisms (Cheung et al. 2005; Lappalainen et al. 2013)). From these, expression quantitative trait loci (eQTL) studies, where a variant (i.e. the molecular quantitative trait loci (QTL)) influences mRNA levels has been the most studied (with 5-40% of functional variants found to colocalize with eQTLs (Connally et al. 2021)). The ImmVar project, one of the first examples, mapped the genetic architecture of the adaptive and innate immunologic transcriptomes, (De Jager et al. 2015; Fairfax and Knight 2014), showing a large effect of genetic variants on gene expression in immune cells. Most importantly though, it showed too the importance of cell-type specific and context dependent associations (Yao et al. 2020), with for instance, a local eQTL for NOD2 found to involve a single nucleotide polymorphism (SNP) previously associated with leprosy only after IFNβ exposure (Lee et al. 2014).

Integrating all this molecular QTLs and intermediate regulatory information, is therefore essential to try and properly connect the causal chain spanning DNA variant to trait variation (Hormozdiari et al. 2018), with essential tools such as fine mapping and gene-prioritization methods (e.g Mendelian Randomization (MR), transcription-wide association studies (TWAS) or colocalization) now part of every functional genomics analysis and extensively reviewed elsewhere (B. Li and Ritchie 2021). These analysis are only made possible thanks to key initiative such as GTEx (GTEx Consortium 2020) or BLUEPRINT that make available repositories of comprehensive QTL annotations for a large number of human tissues. Others, such as FUMA (Watanabe et al. 2017) and Qtilizer (Munz et al. 2020), have made use of them to integrate and comprehensively annotate GWAS results, and others, such as (Kerimov et al. 2021) have focused on including more specialized disease-relevant cell types as well as functional studies, given, as previously mentioned, that the precise actions of genes are dependent on timing and functional context (Connally et al. 2021). For the future, the inclusion of single-cell RNA sequencing (van der Wijst et al. 2018; Jagadeesh et al. 2022)

information is proving essential to drive discoveries, with the combination of state of the art modelling such as deep learning, aiding in the quest (Novakovsky et al. 2022). In the work highlighted in Chapter 4 we will apply GWAS to better understand the genetic makeup of ageing trajectories as well as perform gene annotation with eQTLs and use FUMA for gene enrichment analysis.

One last tool which we have harnessed in our research and which is hypothesised to help identify pathways especially vulnerable to malfunction is the study of multimorbidity. Defined as the co-occurrence of at least two diseases in an individual, it has recently become a priority for global research ('Multimorbidity: A Priority for Global Health Research | The Academy of Medical Sciences') with it being associated to a high mortality rate, increased healthcare utilization with increased healthcare expenses and overall affecting functioning and quality of life (Chowdhury et al. 2023). Also, recent studies have revealed that some diseases co-occur more often than expected by chance (Hidalgo et al. 2009), pointing at possible shared molecular mechanisms or common risk factor triggers such as lifestyle or age, with systemic chronic inflammation postulated as the enabling molecular mediator (Furman et al. 2019). Although a genetic basis for multimorbidity was first studied through the Human Disease Network (HDN) exploring common genetic origins of Mendelian diseases (Goh et al. 2007). by studying diseases together using GWAS, pleiotropic risk variants (genetic locus that affect more than one trait) that are not otherwise found in single-disease studies can be detected (Ellinghaus et al. 2016). This may be due to these variants granting a general disease risk and so only apparent in multi-disease studies instead of single disease ones (Hu, Thomas, and Brunak 2016; Blair et al. 2013). Therefore, disease co-occurrences can help elucidate the underlying network biology through shared and multifunctional genes and pathways.

1.2.2 Describing healthy phenotypes

Exploring health to disease transitions requires of a clear signature of what a healthy phenotype entails. This has been possible through the analysis of multi-omics data of healthy cohorts through time, with examples including the following longitudinal studies: P100 cohort (Price et al. 2017), the Swedish CArdioPulmonary bioImage Study (SCAPIS) (Tebani et al. 2020) or the integrated personal omics profiling (iPOP) project (Schüssler-Fiorenza Rose et al. 2019)). As key conclusions, all studies demonstrated that a stable and unique molecular profile is found in healthy individuals, with “high variation between individuals across different molecular readouts but a low baseline variation intra-individually”. This same conclusion has been reached when profiling the immune system in depth (Alpert et al. 2019), with claims such as “relatively conserved homeostatic states” at the individual level. Such insights help pave the way towards precision medicine, where measuring individual longitudinal profiles postulates as the best way to make actionable decisions in healthcare.

1.2.3 Health to disease transitions

Longitudinal multi-omics integration in the previous studies, also allowed for the study of health to disease transition patterns. For instance, the iPOP study (Schüssler-Fiorenza Rose et al. 2019)) integrated information of 109 “deep-profiled” individuals, including information on microbiome (Zhou et al. 2019) and ageing (Ahadi et al. 2020) too. A follow-up of 3 years resulted in 49 major health discoveries. Amongst these discoveries we can highlight two. Firstly, they were able to detect a lymphoma before disease symptom onset (Schüssler-Fiorenza Rose et al. 2019). Secondly, they were able to track the transition from health to disease of a participant, following genomic information revealing an increased risk for type II diabetes mellitus. This participant developed said disease following a rhinovirus

infection during the course of the study. In other examples such as (Zimmer et al. 2021), the integration of information through Multi-Omics Factor Analysis (MOFA) described four key archetypes or health states in a wellness cohort defined by three major axis of variation: age, gender and general health. Also, in (Ahadi et al. 2020), 18 million data points for proteomics, metabolomics, transcriptomics, microbiome, genomics and clinical marker data were analysed to create a panel of 87 significant molecules for ageing patterns defining different health trajectories. Four enriched ageing patterns or “ageotypes” were derived from key molecular pathways that changed over time in a given individual. These were pathways affecting: metabolic, immune, liver or kidney pathways. Liver and kidney dysregulations have also been found in recent studies associating multimorbidity and metabolic studies of healthspan, with impaired glucose homeostasis, low-grade inflammation, and lipoprotein metabolism (Pietzner et al. 2021) as other key markers.

1.3 Thesis aims and objectives

Powerful analytical methodologies have been developed to study these ‘big datasets’ and identify underlying patterns. In this work, we are going to focus on the use of these novel methodologies, looking at computational approaches from the systems biology perspective, where biological phenomenon is recognized as a complex system that requires of the integration of these different data layers and organizational levels for its full understanding (Kitano 2002). With this, we will explore the use of machine learning as key computational framework in biological research (Chapter 2) and medicine (Chapter 3), developing tools for its adoption. Finally, in Chapter 4, we will explore the power of computational biology to tackle complex biological questions such as health to disease transitions, focusing on ageing and multimorbidity as driving phenotypes. This thesis therefore aims to:

- Conduct interdisciplinary research combining computational biology, machine learning, and epidemiology to advance biomedical knowledge and address pressing health challenges
- Develop open-source web applications and pipelines to disseminate research findings and make computational tools more accessible to the broader scientific community
- Utilize large-scale datasets, such as the UK Biobank, to study ageing trajectories and health-to-disease transitions and identify risk factors and key biological mechanisms associated with healthy and unhealthy ageing

2 Chapter 2: Biomarker discovery and pipeline development

Content from this chapter has been published:

Bravo-Merodio, L, Williams, JA, Gkoutos, GV, & Acharjee, A. (2019). -omics biomarker identification pipeline for translational medicine. *Journal of Translational Medicine*, *17*(1), 1–10.

Bravo-Merodio, L, Acharjee, A, Hazeldine, J, Bentley, C, Foster, M, Gkoutos, GV, & Lord, JM. (2019). Machine learning for the detection of early immunological markers as predictors of multi-organ dysfunction. *Scientific Data*, *6*(1), 1–10.

2.1 Background and chapter overview

With the development of high throughput technologies, we can now observe biological systems at molecular resolution, generating a wealth of different 'omics' data. The problem now lies in the growing gap between the accumulation of big data — and researchers' knowledge about how to use it effectively (Barone, Williams, and Micklos 2017). To democratize access and enable efficient, reproducible, and reliable analysis of biomedical big data, analytical pipelines with the computational steps required to turn raw data into scientific knowledge have been created with varying levels of complexity and resources. Examples range from accessible bioinformatic scripts such as (Xia and Wishart 2016) and (A. Acharjee et al. 2013)'s modelling of quantitative metabolomics datasets, to more convoluted professional frameworks suitable to streamline large-scale biological analysis. Such infrastructures include Firmiana derived by (Feng et al. 2017) or (Satagopam et al. 2016)'s work which count with modular workflows and accommodate production pipelines with advanced features for data storage and sharing. For a detailed review on bioinformatic workflows please refer to (Leipzig 2017; Canuel et al. 2015).

In this work, we will explore the use of computational tools in biology through the development of a machine learning pipeline adapted for research. As explained previously, machine learning models can be broadly subdivided into supervised and unsupervised algorithms depending on the type of prediction addressed. Here, we have focused on supervised learning and although AI and ML platforms already exist (i.e. Amazon SageMakers or Azure Machine Learning), making these algorithms accessible, most resources still require a high expertise for an appropriate use and are not specifically tailored to biological data.

Typically, biomedical datasets can consist of thousands of data points and numerous associated features, with datasets being highly collinear and noisy. One strategy that we

have followed in our work to deal with these challenges is to apply regularized models, that is, models which have some additional constraints imposed on them so that they yield models with desirable properties. One such property is sparsity, where the model uses only a fraction of the features available to it when making predictions. In our work, we have used The Least Absolute Shrinkage and Selection Operator (LASSO) (Tibshirani 1994) and its successor, the Elastic Net (EN) (Zou and Hastie 2005). These algorithms are very popular (Culos et al. 2020), given they make few assumptions on the data at hand (the features just have to be real-valued) and are therefore applicable to a wide variety of data settings. Also, their predictions are sparse linear combinations of the input features, making them easy to interpret and their model coefficients can be solved in a computationally efficient manner given they result from a convex minimization problem ('Extending the Reach of the Lasso and Elastic Net Penalties: Methodology and Practice - ProQuest' n.d.). A brief explanation of regularization algorithms would be as follows:

Model fitting in linear regression (Equation 1), produces the vector of estimated regression coefficients through ordinary least squares (OLS), with the objective function as the minimum of the residual sum of the squares (RSS) equation (Equation 3). In other words, the values minimizing the RSS function are the estimated regression coefficients (β^).

$$Residual\ sum\ of\ squares(RSS) = \sum_{i=1}^{N} \left(y_i - x_i^T \beta\hat{}\right)^2$$

Equation 3

In regularization methods, an extra term is added (Equation 4), and so the new objective function to minimize becomes:

$$RSS(\beta) + p\lambda(\beta)$$

Equation 4

Here p is a function to penalize and λ forms the penalty/regularization parameter. The penalty function λ controls the trade-off between likelihood and penalty and so influences the variables to be selected. The higher the value of λ, the fewer number of variables are selected and vice versa. The differences between regularization methods lie in the different functions p they penalize. In LASSO, the penalty is applied to the sum of the absolute values of the regression coefficients (L1 norm). Mathematically, we can write this as:

$$\underset{\beta \in R^p}{minimize} \frac{1}{2} \|y - X\beta\|_2^2 + \lambda \|\beta\|_1$$

Equation 5

The left side of the equation is the normal least squares criterion, whereas the right is the penalized sum of the absolute values of the regression coefficients.

Elastic Net creates a sparse model too, as well as theoretically improving its performance with respect to LASSO in datasets with more features than samples ($p \gg n$ problems), by allowing high collinear groups of features through grouped selection. EN is created by including the L2 norm of the coefficients (sum of the squares) to minimize:

$$\underset{\beta \in R^p}{minimize} \frac{1}{2} \|y - X\beta\|_2^2 + \lambda_1 \|\beta\|_1 + \lambda_2 \|\beta\|_2^2$$

Equation 6

This combination of LASSO (Equation 5) and EN (Equation 6) methods comprises the backbone of the computational pipeline presented below. To prove its use in the realm of biological research, we have validated it with three different previously published -omics datasets (Bravo-Merodio, Williams, et al. 2019) and then show-cased its use with a novel dataset involving immune predictors (Bravo-Merodio, Acharjee, et al. 2019). We have then

continued our work by presenting this pipeline as an online supervised machine learning framework to facilitate its uptake. This tool has been based on previous open source initiatives such as the still under development MLStudio (Krispin [2017] 2022), data_exploration app (‘Data_Exploration’ n.d.) or Vicent Nijs’ radiant (‘Radiant - A Shiny App for Statistics and Machine Learning’ n.d.). These apps offer intuitive and guided machine learning analysis but are not specifically tailored to biological data and are less focused on providing interpretable results. Also recently, more complex software alternatives such as SIMON (A. Tomic et al. 2021) HTPmod (D. Chen et al. 2018) or any open-source tools developed by Saez-Rodriguez group, e.g FUNKI (Hernansaiz-Ballesteros et al. 2021), have been developed, exemplifying the need for such tools in the bioinformatics community. Regardless, some of these platforms require setups such as Docker which, for non-experts, might act as main limitation for adoption, especially if interested in just a quick and simple analysis, which we aim to supply.

As a summary, in this chapter, we will explore the use of computational tools in biology through three pieces of work. First, we will describe the development of a machine learning pipeline adapted for research and validated with three different previously published -omics datasets and then show-case its use with a novel dataset. We then, present it too as an online supervised machine learning framework.

2.2 Pipeline development

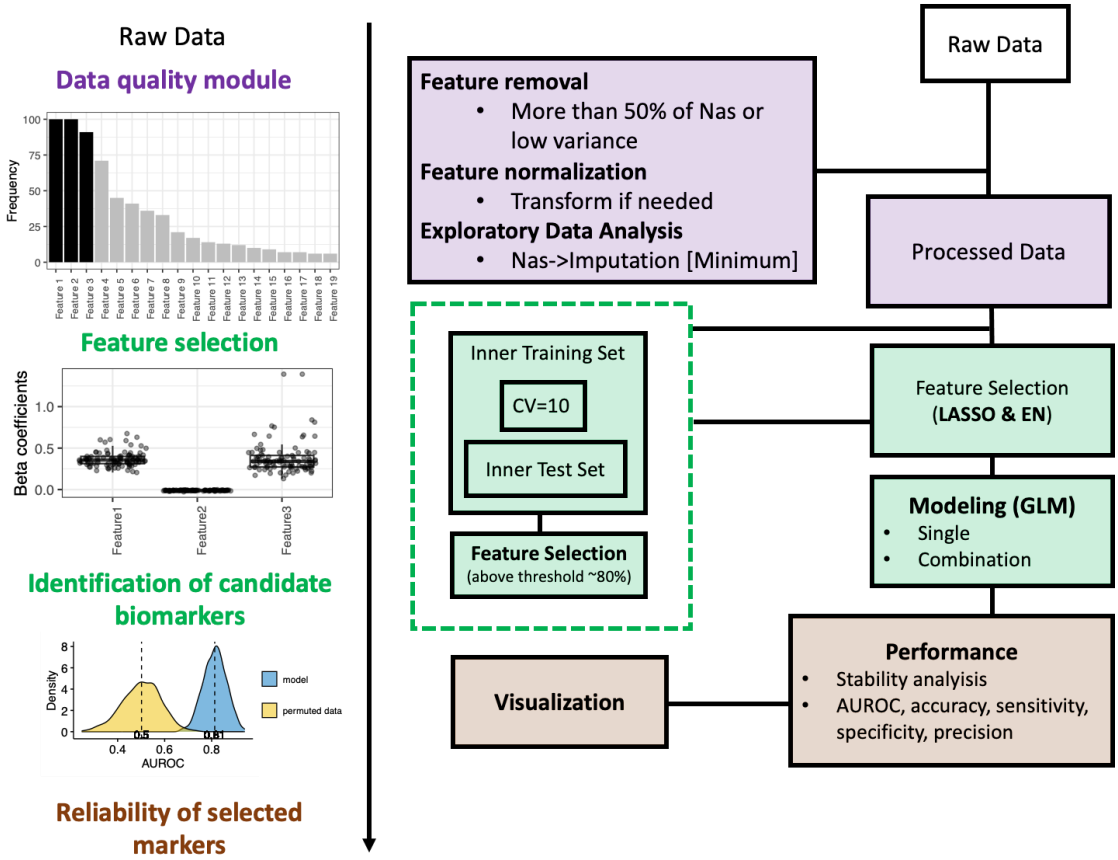


Figure 2-1. Machine learning pipeline structure
Schematic of the workflow/pipeline developed in (Bravo-Merodio, Williams, et al. 2019). It consists of three main modules highlighted in three different colours. Purple) Data quality module offering different options depending on the type of data introduced e.g. microarray and generic/other data. Green) Cross validated feature selection through regularization algorithms LASSO and EN and model building. Brown) Performance assessment, stability analysis and results visualization.

We developed a pipeline (Bravo-Merodio, Williams, et al. 2019) that includes two different supervised classification techniques based on regularization methods to identify biomarkers from -omics or other high dimension clinical datasets. A summary of the different modules configuring it can be seen in Figure 2-1. The pipeline is presented as an editable RMarkdown script found in https://github.com/jaw-bioinf/Biomarker_Identification/. First, the

input data with outcome information and features to model is pre-processed, with standardization and missing values evaluated as well as any extra changes implemented needed by end users. For example, optional filtering steps to reduce dimensionality include many-to-one probe-to-gene mapping in transcriptomics data or the use of external tools or expert biological insight to exclude certain features. In all cases, the input for the 'feature selection' segment of the pipeline is a matrix (file or data frame) of features (key predictors of outcome studied i.e potential biomarkers) and samples along with a target or outcome variable. After a random division of samples into training and test sets, LASSO and EN algorithm as feature selection methods are applied to identify the most highly associated features with respect to the outcome, representing potential biomarker candidates. In order to apply the algorithms for biologically relevant feature selection, we need to optimize the penalty parameter associated with each of the methods in an unbiased way. To achieve this, the pipeline divides samples randomly into a training set composed of 75% of the total number of the samples and a test set consisting of the remaining 25% samples. To ensure optimal training in real-world datasets, all data splits retained class balances of the target variable, so each split reflected a proportion of the target observed in the underlying dataset. We note that class balancing measures, such as boosting or under sampling, are not used to artificially balance training/testing data in each split (outer loop set). Then we apply a 10-fold cross validation on the training set (inner loop set) aiming to have an optimised penalty parameter that can be plugged into the LASSO and EN models. Our pipeline then iterates model creation 100 times (bootstraps) and selects the features that appear more than 80 times in the analysis, as these we deem to be the more significant for the classification model (Figure 2-1). In posterior iterations of the pipeline, this threshold will be automatically adjusted to select those features that appear more than the mean upper quantile of the total frequencies. Also, based on the number of samples available, more bootstraps will be performed too. Then, in order to better understand the relationship between the features

selected and the outcome variable analysed, a display of the weight (β^ coefficients) distribution per model and a box plot of the class differences per feature is generated (Figure 2-1). Selected features are then considered potential candidate biomarkers and final generalized linear model (GLM) fitted with their performance evaluated both alone and in combination through a stability analysis. A stability analysis consists of comparing the devised model with a random model created by permuting the label features and so testing against an empirical null distribution. Both the real model and permutation tests are produced by sampling 1000 random training and test sets and performance is assessed through AUROC, accuracy, sensitivity, precision and specificity plotted as density plots alongside their means and standard deviation (Figure 2-1).

All analyses were performed in R statistical computing (R version 3.4.3) environment with the R packages needed specified and the necessary software dependencies described in the README file located in the repository. All analyses can be performed on a standard PC environment with the run time increasing with larger datasets. After running, it compiles a PDF report containing both all code generated and figures produced. Figures generated include analogues to those found in Figure 2-1. Importantly, a list of selected features (genes, metabolites, etc.) is also printed in the PDF report which can be easily copied so as to be used as input for pathway and ontology enrichment analyses. An example of the report produced has been included in Appendix 1. For the purposes of our validation studies, pathway analysis and ontology enrichment were performed with the EnrichR tool with default settings (analysed on Sept 7, 2018) (E. Y. Chen et al. 2013).

2.3 Pipeline validation

2.3.1 Lipidomics

To assess the performance of our framework, we employed previously published data. First we analysed a lipidomics dataset, available online as supplementary material from (Animesh Acharjee et al. 2017). The lipidomics data was generated from The Cambridge Baby Growth Study (CBGS), a prospective observational birth cohort. From the CBGS cohort we used three different datasets, namely CBGS-1, CBGS-2 and POPS (Pregnancy Outcome Prediction Study) and our objective was to identify nutritional lipid biomarkers able to classify formula, human or a mix of human and formula milk nutrition in babies. All data was obtained from dried blood spots and generated with direct infusion high-resolution mass spectrometry (HRMS). For more details about the processes related to the data generation as well as sample information please check (Animesh Acharjee et al. 2017; Prentice et al. 2015).

In their study, (Animesh Acharjee et al. 2017) identified relevant lipid biomarkers that were able to predict infant feeding patterns based on a combination of supervised Random Forest (a non-linear machine learning algorithm based on an ensemble of decision trees) and iterative backward elimination. In our analysis, by applying our pipeline to the three sets of data and performing three different classification studies (i. human milk vs mixed feeding, ii.formula milk vs mixed feeding, iii. human vs formula milk), two types of relevant lipids; four Sphingomyelin and two Phosphocholine were found as potential biomarkers, with three matching those reported.

Lipids	CBGS1	CBGS2	POPS
SM(39:1)	**C, B, A**	**C,A**	**C**
SM(32:1)	C, B, A	C,A	C
SM(36:1)	C, B		
SM(36:2)	**C, B,A**	**C,B,A**	
PC(34:2)	C,A	C,B	C
PC(35:2)	**C, B**	**C,A**	**C**

Table 2-1. Selected lipid biomarkers
These lipids were found to be able to distinguish different feeding patterns in infants. Three different classification models (A,B,,C)* were generated per dataset (CBGS1, CBGS2, POPS). In bold we have those lipids detected in (Animesh Acharjee et al. 2017).
*Classification A: Human milk vs. mixed (formula milk and human milk combined) feeding*Classification B: Formula milk vs. mixed (formula milk and human milk combined) feeding *Classification C: Human milk vs. formula milk

An in-depth study of the final performance results of dataset CBGS1 can be seen in Figure 2-2 where lipids SM(32:1), SM(39:1) and SM(36:2) (selected at all different classification instances as seen in Table 2-1) make up a final logistic regression/GLM model able to classify participants having formula milk versus mixed with a significant 0.83 AUROC following our stability analysis (Figure 2-2 A). Also, direct visualization of the selected lipids in a box plot, allows for a clear display of the differing values of this feature in different feeding choices justifying its selection and inclusion in the classification model.

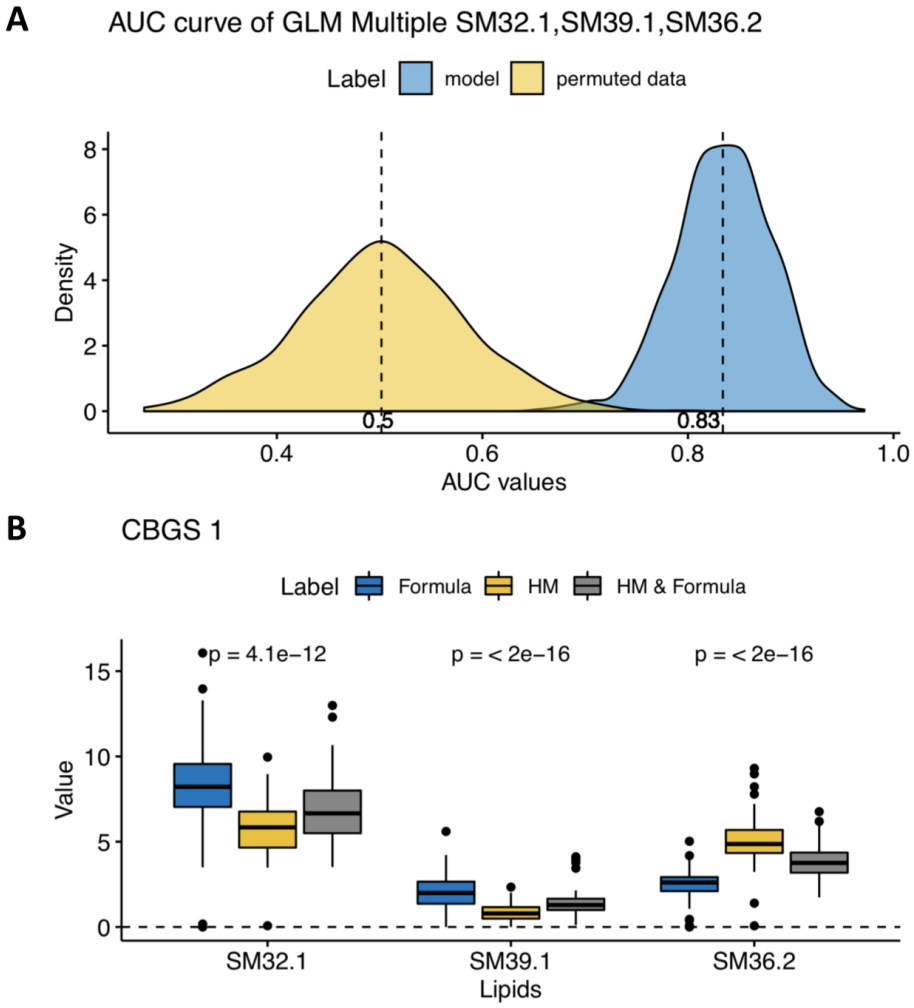


Figure 2-2. Final modelling results of lipids SM32.1, SM39.1 and SM36.2
A) Stability analysis of the GLM multivariable model comprising the lipids selected in the CBGS1 cohort for formula milk versus mixed nutrition classification. 1000 models were generated with both real and permutated labels, with performance assessed using AUROC displaying a difference of mean AUROC 0.83 in real data and 0.5 AUROC in permuted data (i.e random distribution)

2.3.2 Transcriptomics

For transcriptomic data, two datasets from the Gene Expression Omnibus (GEO) were tested. GSE15471,a pancreatic ductal adenocarcinoma (PDAC) microarray expression dataset (n = 36 control, n = 36 cases) (Rajamani and Bhasin 2016) and GSE15061 microarray expression data from a three-cohort study of acute myeloid leukaemia (AML) cell lines with n = 404 AML and n = 138 control samples. A third transitional cohort of MDS

samples was excluded (Mills et al. 2009). In each case, the Robust Multichip Average standardized Affymetrix Human Genome U133 Plus 2.0 data submitted by the authors was taken as input along with class information indicating case or control condition. Data can be accessed at https://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSE15471 and https://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSE15061 respectively. Pre-processing of microarray data included; (a) taking the median of duplicate probes across all conditions to yield one unique probe per experiment, (b) collapsing rows of probes belonging to identical genes and taking maximum expressed probe via the WGCNA R package, and (c) testing features (samples) for low variance via the caret R package and removing those with near-zero variance among all genes. The resulting numerical matrix of normalized gene expression values was used as input for each experiment, yielding 22,880 genes and 542 samples for the AML dataset, and an equal number of genes and 78 samples for the pancreatic cancer dataset. Normalized data matrices produced both for validating the reproducibility of our pipeline as well as the ones used as example input, are available in our project github repository (https://github.com/jaw-bioinf/Biomarker_Identification/).

The studies reporting this data performed slightly different classification studies in their research, with (Rajamani and Bhasin 2016) integrating epigenetic and regulatory information (miRNAs) in their feature selection procedure and (Mills et al. 2009) including too a transitional cohort of myelodysplastic syndrome (MDS) samples in its prognostic classification model. Therefore, we could not directly map our results and so opted for validation via external pathway and ontology gene set enrichment tools.

In the pancreatic cancer dataset GSE15471, the top features selected included the following 20 genes: SULF1, COL10A1, MIR34AHG, INHBA, COL8A1, FN1, THBS2, NOX4, NTM,

RASAL2, ADAMTS12, CAPG, CTHRC1, FAP, VCAN, SLPI, WISP1, LTBP1, GPRC5A, TIMP1 (Figure 2-3 A), which were found to be often discussed in pancreatic cancer literature (Ma et al. 2018). Also, our biological pathway analysis revealed a class of biomarkers enriched in several known cancer pathways. After stringent multiple testing correction, the following pathways were identified as being enriched:

- Senescence and Autophagy in Cancer
- Integrated Pancreatic Cancer Pathway
- TGF Beta Signalling Pathway (mouse)
- miRNA targets in ECM and membrane receptors
- TGF Beta Receptor Signalling (human)

The full results of our pipeline, including feature ranking graphs and AUROC, sensitivity, specificity, and accuracy scores of each variable, as well as all the executed code are included in the above mentioned github repository.

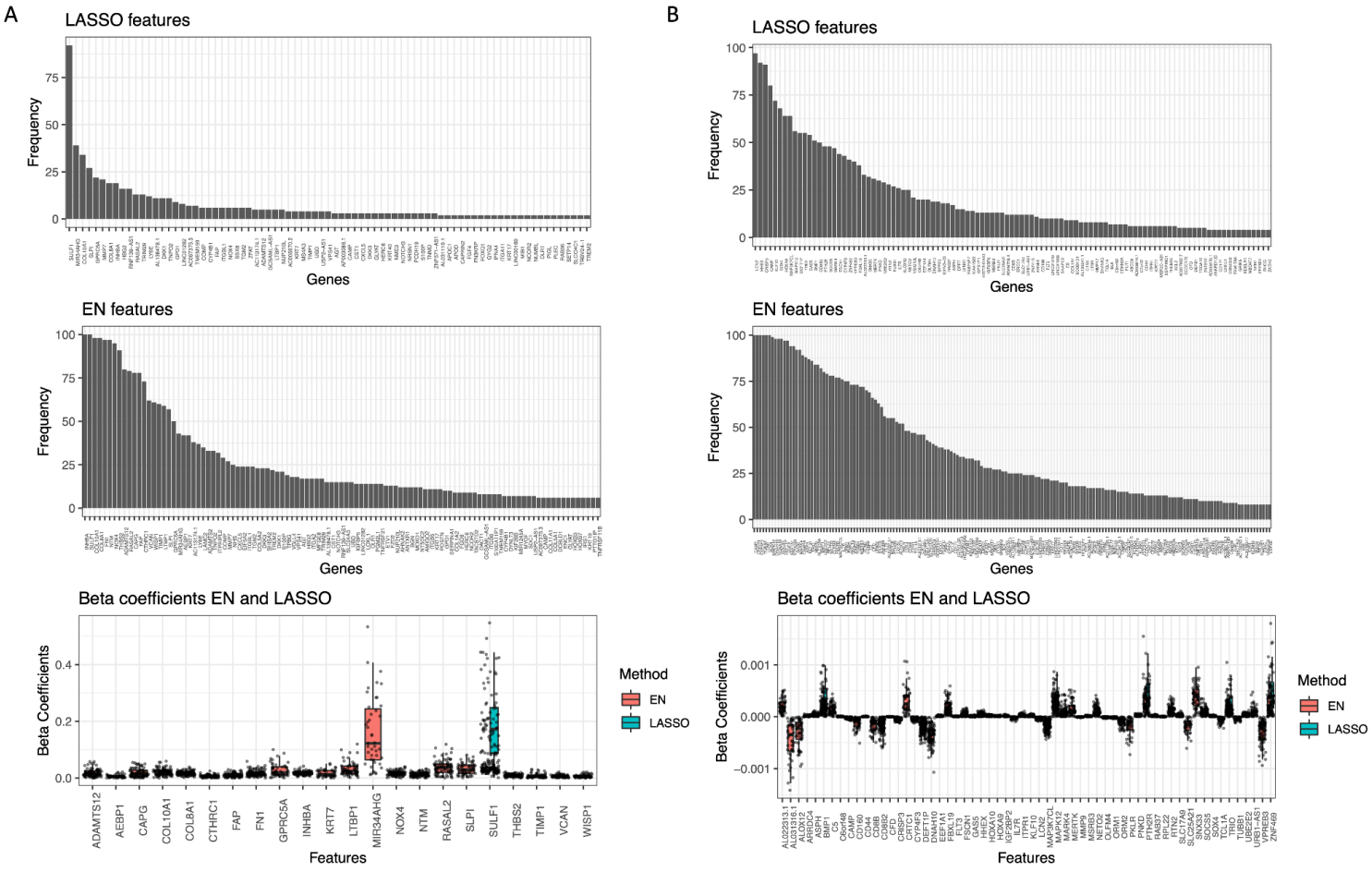


Figure 2-3. Feature selection pipeline

A) Pancreatic cancer data set. B) AML data set. From top to down, feature selection using LASSO or EN. Beta coefficients of the selected genes and the models they were selected with. For LASSO and EN, only the genes with frequency above 3 and 7 were plotted in the histogram respectively.

The pipeline was also employed against an acute myeloid leukaemia (AML) cohort (GSE15061). This analysis revealed sixty genes contributing significantly to AML (Figure 2-3 B). This gene set interacts with several AML-associated transcription factors, including NKX2-3, HOXA7, and MYB. The analysis of genes active in cell lines available in the Cancer Cell Line Encyclopedia (Barretina et al. 2012) revealed that our derived gene set is significantly enriched in multiple haematopoietic and lymphoid tissue lines. Additionally, investigation of the presence of our depicted genes in biological pathways, annotated in the KEGG database (Kanehisa and Goto 2000) confirmed their known AML-gene associations mediated by the ‘Hematopoietic stem cell lineage’ and ‘Transcriptional misregulation in cancer’ pathways. Further results from these analyses are presented in (Bravo-Merodio, Williams, et al. 2019).

2.4 Use case: Golden Hour dataset

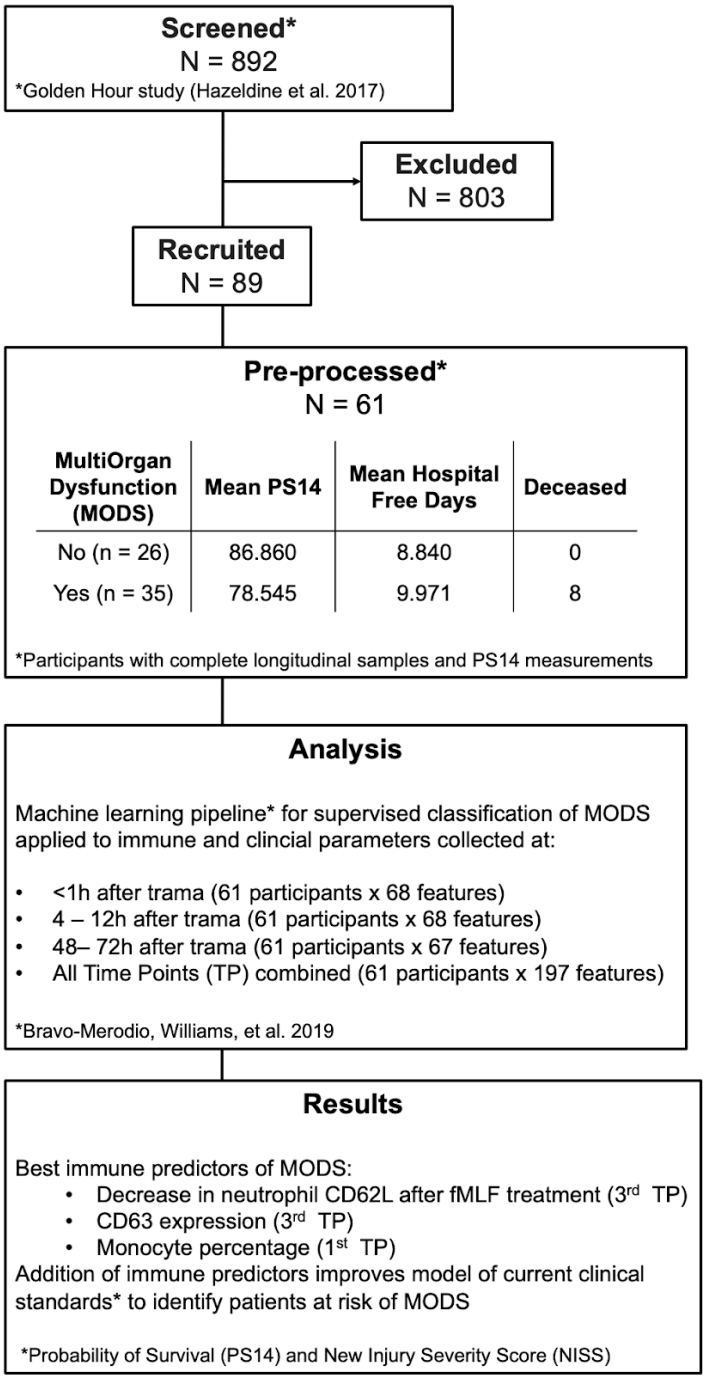


Figure 2-4. Flowchart of the Golden Hour case-study

The immune response to trauma was analysed in 89 adult trauma patients at three different time points (<1 h, 4–12 h, 48–72 h). From these 89 patients, 61 were selected for a more thorough analysis given their complete datasets and availability of main outcome scoring. Analysis description and key results are here highlighted, with the full published work accessible at (Bravo-Merodio, Acharjee, et al. 2019)

We then applied our pipeline to a novel dataset from the Golden Hour study (Hazeldine et al. 2017) focused on the early immune response to major trauma. This dataset is unique in characterising the immune signature of patients during the very early response to trauma given that logically, with the abrupt nature of a traumatic injury, knowledge of the physiological response after trauma has largely been based on post-admission to hospital blood samples. However, studies such as (Hazeldine et al. 2017; Naumann et al. 2017; Xiao et al. 2011) have recently suggested that by then the body's immune response might already be set towards either a helpful or dysregulated response influencing subsequent positive or negative patient outcomes. Therefore, the predictive value of immune information immediately after severe injury (before 2 hours) might help provide optimal care for patients, encouraging interventions such as early total care, damage control orthopaedic surgery (DOC) or guiding immunomodulatory therapies (Huber-Lang, Lambris, and Ward 2018; Spruijt, Visser, and Leenen 2010). Moreover, the immediate immune signature has been shown to hold predictive information for outcomes such as MultiOrgan Dysfunction (MODS) and mortality unaccounted for in later time points (Cabrera et al. 2017).

A summary of the work (Bravo-Merodio, Acharjee, et al. 2019) can be seen in Figure 2-4.
Here, the previously devised pipeline (Figure 2-1) was applied on immune and clinical markers describing the immune response to trauma in 61 patients at three different post-injury time points (ultra-early (<=1 h), 4–12 h, 48–72 h) to classify participants developing MODS. MODS, a progressive failure of several independent organ systems, occurs in 30% to 40% of trauma patients (Shepherd, Cole, and Brohi 2017) with nearly 50% of those in the Intensive Care Unit (ICU) dying as a result (Mayr et al. 2006). The immunological variables assessed belonged to the following categories: neutrophil functional analysis (response to formyl-methionine-leucine-phenylalanine (fMLF)), monocyte response to lipopolysaccharides (LPS) stimulation, haematological, cfDNA analysis and serum

analysis (cytokines and cortisol). Results of the key identified features per time point and in combination can be seen in Figure 2-5 . As expected, current clinical scoring variables such as New Injury Severity Score (NISS) and Probability of Survival (PS14) are predominantly selected features at all time points given they aggregate complex information such as gravity of the wound to estimate the extent of trauma damage.

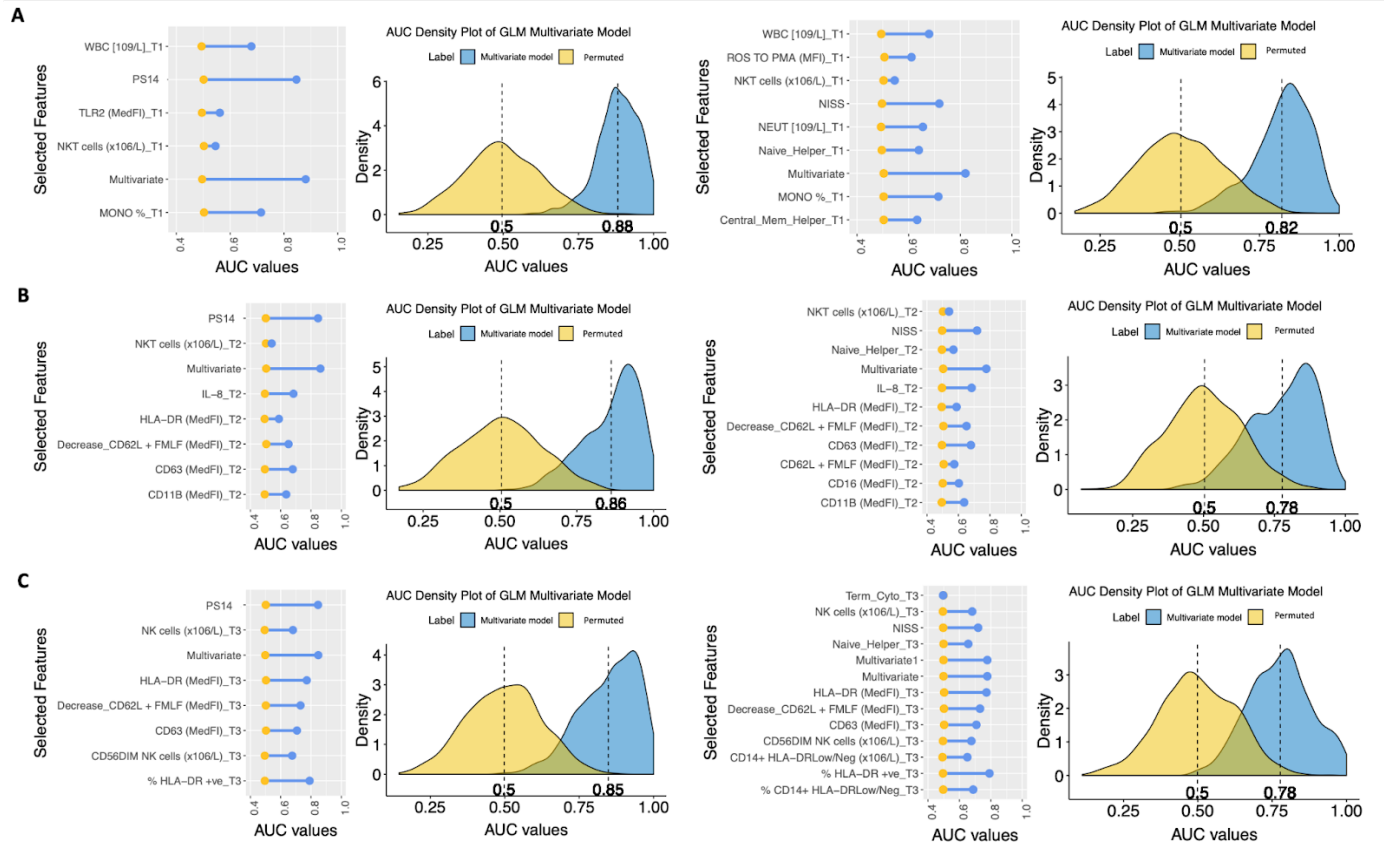


Figure 2-5. Univariate and multivariable performance of final selected features

The AUC performance values for each time point (a) <1 h, (b) 4–12 h, (c) 48–72 h with either PS14 (left) or NISS (right) included. These are a result of 1000 GLM models of both permuted (yellow) and normal/real data (blue). On the left hand side of each figure a dumbell plot for each of the selected features per time point and the combination of all under “Multivariate” is presented. Here we present the selected features for each time point, and so those that correspond to the first quantile of frequencies of the 100 sparse models created in the LASSO and EN feature selection models. On the right hand side, a density plot of the AUC values of the 1000 permuted and non permuted GLM models of the selected features per time point is depicted. The average AUC value is presented in bold.

We then repeated the analysis using data integrated from the three different time points and applied a more stringent threshold for feature selection. The final model (Figure 2-6 (Left)) follows from an understanding that MODS development necessitates a multivariable approach that is able to reproduce heterogeneous and complex interactions characterising the complex nature of the acute immune response to injury. The final selected features were not significantly correlated and showed both biological significance and clinical relevance given the low cost and short time taken to perform their analysis. By selecting the features irrespective of time points, we are selecting the overall highest associated changes with MODS development. With this approach, a further understanding of how the response develops might be reflected in this combination of time points, with multifaceted relations between components and kinetics revealed.

More specifically, this multivariable combination consisted of three features related to neutrophil function: the third time point both neutrophil decrease in CD62L expression after fMLF treatment and CD63 expression and the first time point monocyte percentage (Figure 2-6 (Right)). This is unsurprising as it is well understood that traumatic injury leads to marked alterations in the phenotype, function and life-span of circulating neutrophils with significant changes in the composition of the circulating neutrophil pool (Manson et al. 2016). More specifically, L-selectin (CD62L) a receptor that facilitates tethering of neutrophils to the endothelium in inflammation is shown to be less prominently shed in MODS patients than in non-MODs at all time points suggesting a level of immunoparesis. Similar phenomenons have been previously reported in other receptors such as fMLF-induced expression of FcγRII (CD32) (Hietbrink et al. 2009). Reduced expression of CD63, a marker of primary granule exocytosis on the surface of neutrophils, is a well known marker for neutrophil activation. We hypothesize that this apparently less active peripheral pool of neutrophils at time point three (48 to 72h) in patients who developed MODS is due to those activated CD63 expressing neutrophils migrating into tissues/organs where they would cause damage and potentially

contribute to the pathogenesis of MODS. Finally, it is likely that the significantly lower monocyte percentage associated with the development of MODS in the first time point (<1 h) is simply a consequence of the higher frequency of neutrophils and immature granulocytes. By just using immune predictors, an AUC value of 0.874 ± 0.064 was achieved. The prediction ability of the model was then further improved with the addition of NISS and PS14 clinical scores, where AUC values of 0.917 ± 0.064 AUC and 0.922 ± 0.054 were obtained respectively. All three models, as seen in Figure 2-6 (Left) were statistically significant with respect to pure randomness as exemplified by permuted data. By comparing with their univariate model performance, we can further establish a 13 and 8% increase in the predictive capability of NISS and PS14 scores. Also, the discovery of relevant biological descriptors offers the possibility of potentially identifying clinical targets that may be used for intervention. All code for this work can be found at InFlamUOB/NSDTrauma.

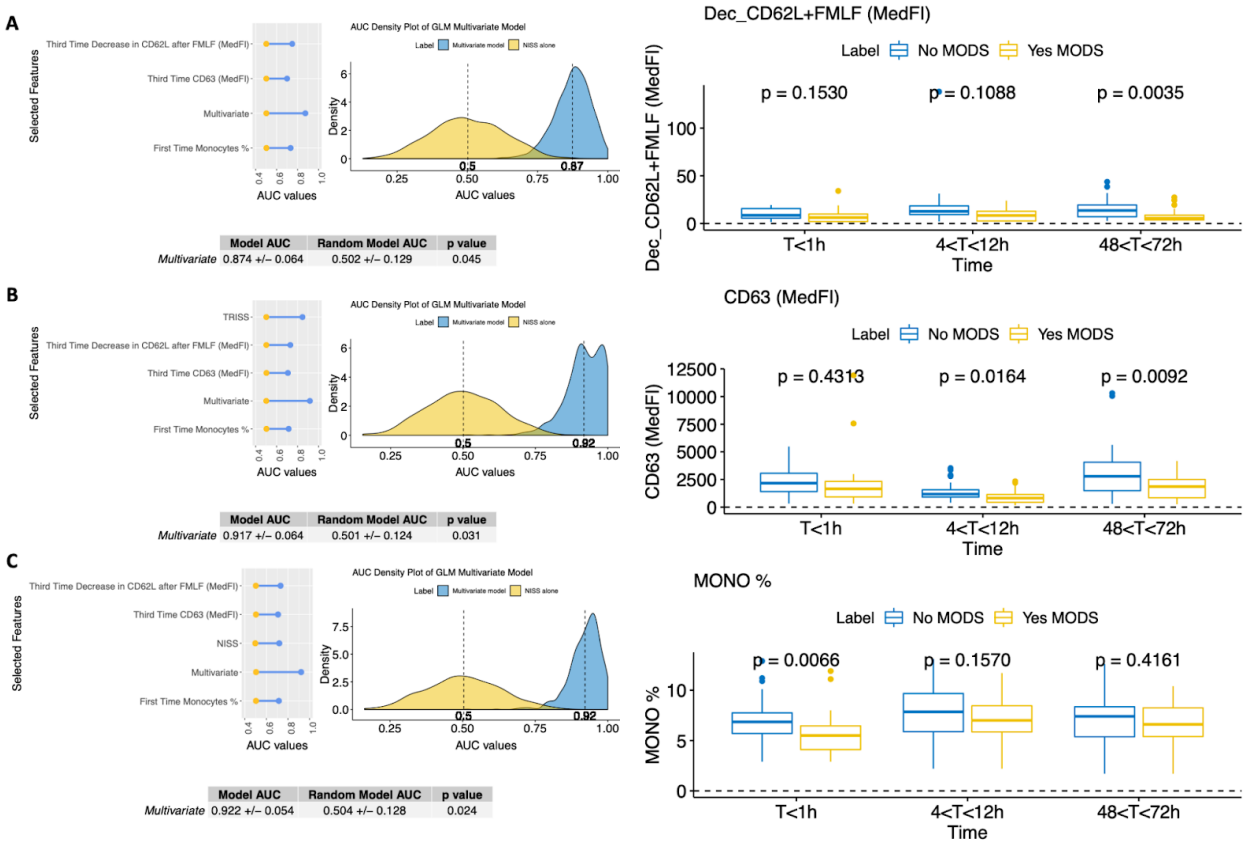


Figure 2-6. Final model and selected immune markers
Left. Univariate and multivariable performance (AUC values) of the selected features for the integrated time points in dumbbell and density plot of (a) the selected features (b) the selected features plus PS14 (c) the selected features plus NISS. Confidence interval and p-value information of the multivariable model is also included. P-value was evaluated by assessing how many times a permuted value's AUC was above the non-permuted/real AUC mean. Right. Boxplot of values of the selected immune features comparing MODS and non MODS participants per each time point.

2.5 Web app

Having devised and validated our pipeline of work, we now aimed to make it more user friendly and accessible. Therefore, we have developed an interactive, simple, free and open-source supervised machine learning webtool using RShiny (Chang et al. 2023) to facilitate its use. Our developed infrastructure (Figure 2-7) consists of an intuitive interface with four modules mapping to a great extent those described in Figure 2-1: an exploratory data analysis visualization module, a module that performs feature selection using the same regularization algorithms in (Bravo-Merodio, Williams, et al. 2019; Bravo-Merodio, Acharjee, et al. 2019) as well as others, a third module for model building with the selected features and their combinations and a summary of results module. As key differences, feature importance and more machine learning algorithms have been introduced (GLM, RF, Decision Trees (DT), Boosted Trees (BT) or regularized boosted trees), using the tidymodels framework (Kuhn and Wickham 2020). Performance of each model with each different predictor combination and algorithm can be visualized and compared and a descriptive RMarkdown report can be downloaded summarizing all main results and offering a guided and easy interpretation. An example can be found in Appendix 1.

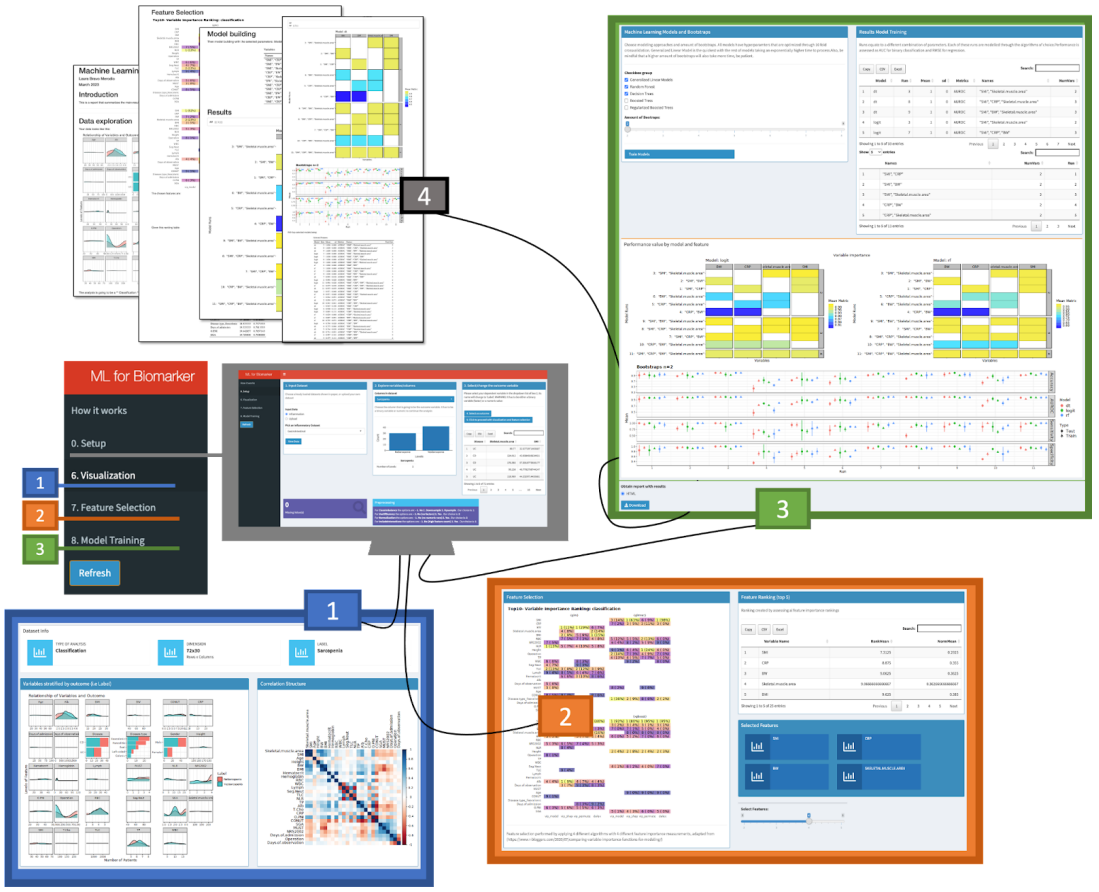


Figure 2-7. Diagram of all modules available in the RShiny supervised machine learning webapp.
(1) Following data input and preprocessing, data can be visualized and features of interest selected. The user can then perform features selection through different algorithms combining both non-linear and linear modelling strategies (2). After determining the predictors to build the final model with, a selection of machine learning models can be applied to model the data (3). A report with a summary of all results can be downloaded in the end with a guided interpretation of all visualizations (4).

The latest version of this pipeline can be reproduced through the code found in our repository (InFlamUOB/MachLearnWebApp). A still in development online version is also available at https://inflam.shinyapps.io/MachLearnBiomarker2/. For a previous version of the pipeline with detailed step-by-step instructions please refer to: https://youtu.be/WsOwZ7kIFEQ, also featured in the “How it works” widget of the interface. A more detailed overview of the webtool is as follows:

2.5.1 Module 1: Visualisation

A numbered box indicates to first select a dataset to start with the analysis. The dataset can be either uploaded (.csv file) or publicly available biomedical example datasets (Bamba et al. 2017; Kou et al. 2019; Souza et al. 2017) can be used. The number of columns/features or samples is not limited, although the larger the dataset the longer the analysis takes. The selected dataset can then be visualised, both as a table or explored column by column with added summary data such as count for categorical variables. Upon selection of the outcome label, the data will change the variable’s name selected to “Label” and the user will obtain a summary of the automatically generated preprocessing steps based on the data characteristics. Supervised machine learning can be performed for both continuous (regression) or binary categorical data (classification). General characteristics of the data selected such as names, dimensions and output variables are printed alongside a simple correlation plot (corrr package) and a density plot with variables stratified by the outcome selected.

2.5.2 Module 2: Feature selection

Our aim was to create a tool targeted to the life science industry’s needs, with conclusive visualisation and results upon which actionable decisions can be taken and useful biological insight generated. For this we have wanted to acknowledge the common need to reduce the feature dimension space previously mentioned and focus on potential biomarkers of interest

through feature selection. First, the regularization algorithms used in the previous work are applied through GLMNET (Friedman, Hastie, and Tibshirani 2010), as well as three others (GLM, RANGER (Random Forest) and the boosted tree algorithm XGBoost) using the tidymodels package in R (Kuhn and Wickham 2020) and following (ElHabr 2020)'s work. The features that appear to be more important for each model are then selected through different feature importance methods including model-agnostic algorithms such as the explainable AI package DALEX (Biecek 2018), Shapley values (Sundararajan and Najmi 2020) and permutation related variable importance or model-specific algorithms such as model related variable importance. Given that there is no algorithm able to fit all data perfectly ("No Free Lunch Theorem"(Belkin et al. 2019)), testing more than one algorithm and more than one approach to interpret feature importance provides us with the possibility of acknowledging many more possible underlying relationships. For each model, the data is resampled and hyperparameters optimized. The top 9 ranking features per model can be visualized in a heatmap and those scoring high importance in the most models are selected and followed through to final model building.

2.5.3 Module 3: Final model building

Models are built based on a combination of the top features selected (user can select from 1 to the 5 top ranked predictors). These features are then combined with each other in pairs, trios as well as all together, generating for example 11 different multivariable models if 4 predictors are selected. Five different algorithms are then offered as options, allowing for the comparison of results from different ML algorithms and tailoring to the specific users needs, with ones being simpler and interpretable (i.e GLM) and others of higher complexity and requiring extra memory resources (i.e XGBoost). Moreover, different number of bootstraps (random splits of the dataset into 25% testing and 75% training) can be selected. All feature

combinations are then modelled based on the specified algorithm as well as the specified number of bootstraps. Results are then presented in a ranking table, a dot plot showing all accuracy, sensitivity, specificity and AUROC results for both training and testing splits, as well as a tile plot showing the AUROC values by model and feature.

2.5.4 Module 4: Analysis report

The results can then be downloaded through an RMarkdown report for all main results (plots and tables). A report example is detailed in Appendix 1.

2.6 Conclusion and discussion

Biological research has become increasingly computational. In this chapter we have presented a machine learning pipeline for biomarker identification through three different pieces of work. First, we have developed the workflow and validated it, then applied it to a novel case of study and finally worked on improving it, adapting it to allow its dissemination. To accomplish this, we have harnessed the power of open source software using R (R Core Team 2020) and RStudio (RStudio Team 2020) which develop infrastructure such as RShiny (Chang et al. 2023) and RMarkdown to allow for the seamless sharing of workflows for the larger bioinformatics community and empowering researchers to conduct reproducible analyses at scale. Our framework harnesses these developments to cater the provision of such tools to non-experts allowing for an easy, intuitive and simple preliminary analysis of data that can aid in future biomedical/clinical translation.

In this work we have mainly focused on the development of pipelines and workflows given workflow systems are rapidly becoming the workhorses of modern bioinformatics. With our in-house developed pipeline, we have then been able to analyse the Golden Hour dataset (Hazeldine et al. 2017), unique for its ultra-early immune information upon trauma, unravelling signatures of patients with worse outcomes. Albeit a low sample size (61

participants), the robustness of our pipeline, with stability analysis and previous validation studies (Bravo-Merodio, Williams, et al. 2019) points at our identified potential biomarkers (i.e decrease in neutrophil CD62L after fMLF treatment, CD63 expression and monocyte percentage) as interesting leads to follow-up in future studies (Bravo-Merodio, Acharjee, et al. 2019).

In an attempt to further improve our workflow, our last version of our machine learning pipeline is adapted into a webapp and incorporates a broader range of machine learning algorithms, including XGBoost, considered one of the best performing algorithms (X. R. Gao et al. 2023). It also focuses on explainability of the underlying models, understanding that biological and healthcare research requires of a high degree of interpretability, where being able to unravel mechanisms of interest, translate results and make actionable decisions is essential (Rudin 2019; Caruana et al. 2015). Examples include the use of heatmaps showing the contribution of specific variables (Appendix 1), or feature importance algorithms incorporating the latest developments on explainable AI.

For the future, further work on the webapp is required. First, more work on software robustness would be needed with a special effort required to create a stable software package compliant with standard FAIR principles (findable, accessible, interoperable and reusable) (Barker et al. 2022). This might be achieved through the adaptation to well-known workflow software tools such as Snakemake, or Galaxy (Reiter et al. 2021). Also, a special emphasis on unit testing, scalability, interoperability and parallelization in external servers would be needed. Regarding content, with a quickly expanding ML universe, novel algorithms to adapt to the most recent developments would be of interest too, with for example the inclusion of multiclass classification analysis or other metrics for model comparison such as maximum likelihood. These improvements would align with current state

of the art examples such as PyCaret, a low-code machine learning library in Python that automates machine learning workflows (Ali 2020).

3 Chapter 3: Clinical translation and CovidSurg

Content from this chapter has been published:

Bravo-Merodio, L, & Collaborative, C. (2021). Machine learning risk prediction of mortality for patients undergoing surgery with perioperative Sars-Cov-2: The CovidSurg mortality score. *The British Journal of Surgery*, *108*(11), 1274.

Or is undergoing review:

Bravo-Merodio, L, & Collaborative, C. (n.d.-). Risk of pulmonary complications in adults undergoing elective surgery: Development and validation of a prognostic model to support safe global recovery of elective surgery. In *Lancet Digital Health*.

3.1 Background and chapter overview

Medicine is inherently a decision-based process. But the decisions on diagnosis or prognosis are extremely complex and prone to inherent bias and errors. With machine learning, large amounts of heterogeneous data are transformed into statistical patterns able to capture relevant information, helping guide patient care. Regardless, actual deployments of ML systems in the clinic are scarce with the need for robust, trustworthy, explainable, and transparent systems still presenting challenges (Rajkomar, Dean, and Kohane 2019).

In this chapter, we showcase our work building clinical decision-support systems through two different clinical prediction models. During the COVID-19 pandemic, the pipeline presented in Chapter 2 was adapted to explore the impact of COVID-19 in surgical patients and services. These studies were done alongside the CovidSurg Collaborative (‘CovidSurg – Globalsurg’ n.d.), a NIHR Global Surgery Research Unit built out of the need to derive evidence-based guidelines for surgical practice. During the pandemic, they were focused on exploring how to restart surgery in a safely manner and collected international, multicentre, observational cohorts of patients with SARS-CoV-2 infection who had surgery at different hospitals and countries (COVIDSurg Collaborative 2020; J. Glasbey et al. 2021; STARSurg Collaborative and COVIDSurg Collaborative 2021; Collaborative and Collaborative 2021). Given that guiding patient risk stratification for surgery was essential with tens of millions of operations cancelled due to excessive postoperative pulmonary complications (~51.2%) and mortality rates (~23.8%) in patients with perioperative SARS-CoV-2 infection (COVIDSurg Collaborative 2020), in this work, we have focused on building two different prognostic models aimed for:

1. Prediction of postoperative mortality risk in patients with perioperative SARS-CoV-2 infection (CovidSurg Mortality Score) (COVIDSurg Collaborative 2021)

2. Prediction of postoperative pulmonary complications in adults undergoing elective surgery (GlobalSurg-Pulmonary Score).

3.2 Machine learning risk prediction of mortality for patients undergoing surgery with perioperative SARS-COV-2: The CovidSurg mortality score

3.2.1 Summary

At the time of the study, perioperative SARS-CoV-2 infection was a great challenge with rates of infection in elective surgery patients ranging from 1 to 9% (Kibbe 2020; J. C. Glasbey et al. 2021), vaccination not yet available and preoperative screening increasing the numbers of SARS-CoV-2-positive patients. Therefore, to inform consent and shared decision-making, a robust, globally applicable score to predict individualized mortality risk for patients with perioperative SARS-CoV-2 infection was created (COVIDSurg Collaborative 2021).

To do so we studied an international, multicentre (756 hospitals from 69 different countries), prospective, cohort study (COVIDSurg Cohort Study) which included patients who were diagnosed with SARS-CoV-2 in the 7 days before or the 30 days after surgery. Patients undergoing any type of surgery were included from 1 February 2020 to 31 July 2020. Full study methodology has been published previously (COVIDSurg Collaborative 2020).

Machine learning methods were used to analyse the COVIDSurg Cohort Study dataset, in a similar way to that described in Chapter 2. First, sixteen patient and operative variables were entered into the analysis and based on time, data were split into a derivation set, where all analysis and modelling were performed, and a validation set, where the final model was assessed. Five features (age, preoperative respiratory support, ASA grade, specialty and revised cardiac risk index (RCRI) score) were selected following variable importance measurements from both linear and non-linear modelling. These features were then combined into 26 different predictor sets, which were fitted through three different algorithms

(logistic regression, decision trees and random forest), generating a total of 78 different models. These models were tuned with a 10-fold cross-validation, fitted in the 75% split of the derivation set and assessed in the remaining 25%. In order to ascertain the model's stability, this training and testing split was randomly repeated 100 times (bootstraps). The best performing model (logistic regression of predictors: age, ASA grade, RCRI score and preoperative respiratory support) had a performance of 0.80, 95% CI (0.77 – 0.83) AUROC and calibration of 0.084 (Brier score), 0.078 (calibration-in-the-large) and 1.08 (slope). The model was then deployed alongside risk scores for individual prediction and can be accessed through either http://covidsurgrisk.app/ or https://inflam.shinyapps.io/RiskAppLukeAugust2/.

3.2.2 Methods and results

3.2.2.1 Data: Patient characteristics and outcomes

From 10,029 patients in the cohort study, data from 8492 patients entered the machine learning processes (Figure 3-1).

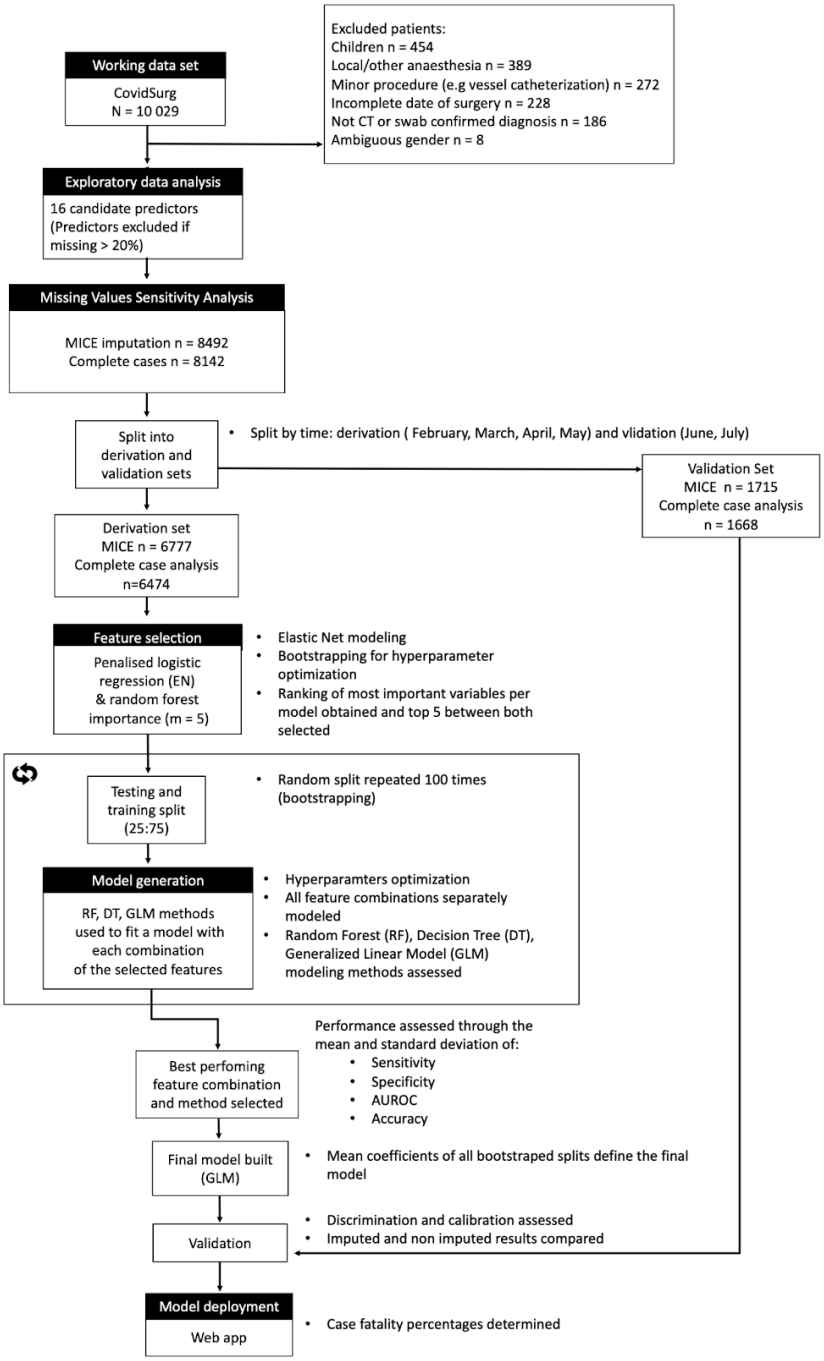


Figure 3-1. Workflow description
Cohort study, patient inclusion and model derivation and validation description (MICE, multiple imputation by chained equations; EN, elastic net)

The most common surgical procedures were abdominal (40.%, 3446 of 8492 patients), orthopaedic (33.8%, 2867 of 8492) and head and neck surgery (9.8%, 835 of 8492). Emergency surgery accounted for 80.8% (6862 of 8492 patients) and benign disease for

57.3% (4868 of 8492). Of the 8492 patients, data from 6777 patients who had surgery from February to May 2020 (726 hospitals, 65 countries) were entered into the derivation set and data from 1715 who had surgery from June to July 2020 (296 hospitals, 53 countries) were entered into the validation set. Mortality rates were 17.2% overall (1462 of 8492 patients), 18.9%(1279 of 6777) in the derivation set and 10.7% (183 of 1715) in the validation set (Table 3-1). This decrease in overall postoperative mortality is associated with a change in the characteristics of patients undergoing surgery between the two periods, with fewer older and higher ASA grades patients operated in during the validation set period. Whilst this indicates that surgeons have started to select lower-risk patients for surgery, the validation set mortality rate of 10.7% is still very high, indicating that a more structured and reproducible approach to risk-assessment would still improve patient selection.

From the 16 variables studied 15 only continued in our study (Figure 3-1), given C-Reactive Protein had to be excluded due to exceeding our missing data threshold (≥20%). The remaining features are widely available before surgery and globally (even in resource limited environments without the need for additional tests). A complete case analysis is presented in this research with the extra multiple imputation analysis carried out available in (COVIDSurg Collaborative 2021).

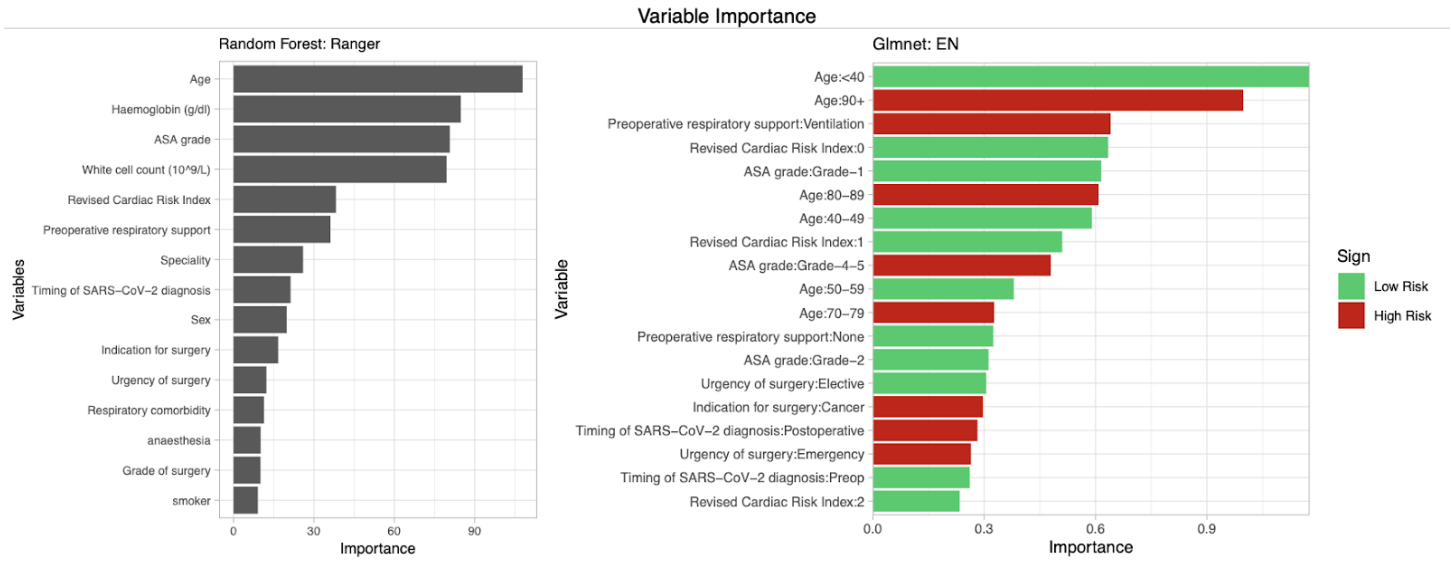
		Derivation cohort			Validation cohort		
Factor	**Level**	**Alive N=5498**	**Dead N=1279**	**P-val**	**Alive N=1532**	**Dead N=183**	**P-val**
Age	18-39 years	1124 (20.4%)	45 (3.52%)	<0.001	718 (46.9%)	29 (15.8%)	<0.001
	40-49 years	563 (10.2%)	62 (4.85%)		215 (14.0%)	15 (8.20%)	
	50-59 years	756 (13.8%)	120 (9.38%)		211 (13.8%)	35 (19.1%)	
	60-69 years	921 (16.8%)	221 (17.3%)		170 (11.1%)	36 (19.7%)	
	70-79 years	1025 (18.6%)	342 (26.7%)		110 (7.18%)	36 (19.7%)	
	80-89 years	832 (15.1%)	337 (26.3%)		95 (6.20%)	25 (13.7%)	
	≥90 years	277 (5.04%)	152 (11.9%)		13 (0.85%)	7 (3.83%)	
Sex	Female	2794 (50.8%)	528 (41.3%)	<0.001	764 (49.9%)	75 (41.0%)	0.028
	Male	2703 (49.2%)	749 (58.7%)		768 (50.1%)	108 (59.0%)	
C-reactive protein	mg/L	33.0 [7.00;102]	49.0 [12.0;137]	<0.001	20.0 [5.00;71.2]	94.5 [24.0;162]	<0.001
Grade of surgery	Major	4023 (73.7%)	1020 (80.3%)	<0.001	1077 (70.6%)	135 (75.0%)	0.250
	Minor	1435 (26.3%)	251 (19.7%)		449 (29.4%)	45 (25.0%)	
Anaesthesia	General	4123 (75.0%)	995 (77.8%)	0.039	1036 (67.6%)	147 (80.3%)	0.001
	Regional only	1375 (25.0%)	284 (22.2%)		496 (32.4%)	36 (19.7%)	
Specialty	Abdominal surgery	2236 (41.0%)	430 (33.8%)	<0.001	684 (44.8%)	96 (53.3%)	<0.001
	Cardiothoracic surgery	217 (3.98%)	72 (5.66%)		48 (3.15%)	6 (3.33%)	
	Head and neck surgery	579 (10.6%)	152 (12.0%)		83 (5.44%)	21 (11.7%)	
	Obstetric	448 (8.21%)	15 (1.18%)		324 (21.2%)	3 (1.67%)	
	Orthopaedic surgery	1870 (34.3%)	591 (46.5%)		352 (23.1%)	54 (30.0%)	
	Other surgery	108 (1.98%)	11 (0.87%)		35 (2.29%)	0 (0.00%)	
Haemoglobin (g/dl)		119 [101;135]	111 [96.0;130]	<0.001	123 [108;139]	106 [90.0;125]	<0.001
White cell count (10^9/L)		9.41 [6.90;13.1]	9.70 [7.00;13.9]	0.087	10.0 [7.00;14.0]	13.2 [8.67;18.3]	<0.001
Preoperative respiratory support	None	3595 (66.4%)	662 (52.2%)	<0.001	1188 (78.0%)	75 (41.0%)	<0.001
	Supplemental oxygen	1244 (23.0%)	383 (30.2%)		256 (16.8%)	65 (35.5%)	
	Ventilation	578 (10.7%)	224 (17.7%)		79 (5.19%)	43 (23.5%)	
Indication for surgery	Benign	3065 (55.8%)	623 (48.7%)	<0.001	1049 (68.5%)	131 (71.6%)	0.624
	Cancer	917 (16.7%)	210 (16.4%)		179 (11.7%)	21 (11.5%)	
	Trauma	1514 (27.5%)	445 (34.8%)		304 (19.8%)	31 (16.9%)	
Urgency of surgery	Elective	1132 (20.6%)	176 (13.8%)	<0.001	304 (19.8%)	17 (9.29%)	0.001
	Emergency	4365 (79.4%)	1103 (86.2%)		1228 (80.2%)	166 (90.7%)	
Timing of SARS-CoV-2 diagnosis	Postoperative	2645 (48.1%)	769 (60.1%)	<0.001	370 (24.2%)	67 (36.6%)	<0.001
	Preoperative	2852 (51.9%)	510 (39.9%)		1162 (75.8%)	116 (63.4%)	
Current smoker	Non-smoker	4922 (89.5%)	1180 (92.4%)	0.002	1404 (91.7%)	160 (87.4%)	0.073
	Smoker	575 (10.5%)	97 (7.60%)		127 (8.30%)	23 (12.6%)	
Respiratory comorbidity	No respiratory disease	4710 (85.7%)	1033 (80.9%)	<0.001	1426 (93.1%)	156 (85.2%)	<0.001
	Respiratory disease	787 (14.3%)	244 (19.1%)		105 (6.86%)	27 (14.8%)	
Revised Cardiac Risk Index	0	1734 (31.5%)	305 (23.9%)	<0.001	600 (39.2%)	27 (14.8%)	<0.001
	1	2481 (45.1%)	471 (36.9%)		680 (44.4%)	86 (47.0%)	
	2	889 (16.2%)	320 (25.1%)		196 (12.8%)	50 (27.3%)	
	3	294 (5.35%)	121 (9.48%)		40 (2.61%)	15 (8.20%)	
	≥4	99 (1.80%)	60 (4.70%)		15 (0.98%)	5 (2.73%)	
ASA grade	Grade 1	624 (11.4%)	26 (2.05%)	<0.001	409 (26.7%)	7 (3.83%)	<0.001
	Grade 2	1801 (33.0%)	210 (16.5%)		614 (40.1%)	38 (20.8%)	
	Grade 3	2219 (40.6%)	616 (48.5%)		372 (24.3%)	79 (43.2%)	
	Grades 4-5	821 (15.0%)	417 (32.9%)		137 (8.94%)	59 (32.2%)	

Table 3-1. Patient, disease and operation factors of patients included in the derivation and validation sets. Univariate analysis and p-values obtained through the function

compareGroups package where all continuous variables (non-normal as assessed by the Shapiro-Wilk test) are described through the 25th and 75th percentiles. Categorical variables are shown with % and p-values were calculated depending on the type of variable, with Kruskall-Wallis for continuous and chi-squared permutation or fisher test for categorical, depending on the number of classes. Percentages given as proportion of column total. Differences between grouped data was tested with chi-squared test. ASA: American Society of Anaesthesiologists.

3.2.2.2 Feature selection

In the same way as performed in Chapter 2, we now proceeded with feature selection (Figure 3-1). Feature selection from the candidate predictors aimed to optimize the predictor group by finding the smallest independent set of variables with the greatest association to the outcome, to ensure both best performance and minimize overfitting. The methods used were elastic net (Zou and Hastie 2005) using the package glmnet and random forest from the package ranger, with variable importance evaluated through the vip package and Gini index respectively. As mentioned previously, elastic net generates sparse linear models whereas random forest, an ensemble of decision trees, gives insight on non-linear interactions. These methods were applied on the derivation set, with hyperparameter tuning using 10- and 5-fold resamplings across a regular grid search using the R packages' tidymodels and recipes. By applying these different variable selection methods, we are acknowledging different possibilities of feature interaction criteria ensuring no important patterns are missed out Figure 3-2.



Feature Ranking

Variable Name	Top rank in model	Freq of appearance
Age	1-LASSOEN	8
ASA grade	3-RF	3
Revised Cardiac Risk Index	4-LASSOEN	3
Preoperative respiratory support	3-LASSOEN	2
Speciality	7-RF	2
Haemoglobin (g/dl)	2-RF	1
White cell count (10^9/L)	4-RF	1

Figure 3-2. Variable importance and feature selection

Plot of the impurity importance ranking of random forest ranger model and importance ranking of EN models. Importance was calculated through the vip and ranger gini index parameters for each model respectively and ranked. For all 7 top ranking features of each model, frequency of appearance was assessed, and results shown in B. For EN, the reported coefficients are the highest non-sparse coefficients generated by the linear model and the relationship with the outcome was included and coded here as low risk for negatively associated with mortality and high risk for highly associated with mortality. ASA Grade (America Society of Anaesthesiologists Grade).

The final set of features was determined by merging the feature selection importance ranking of both methods. The criteria used consisted of accounting for both the feature's ranking and frequency of appearance. These features were then combined with each other, generating different models with which to find the most informative set of predictors. A maximum of 5 features were selected as threshold, given this number offered the best compromise of computing time required and possible information loss. The five features taken forward into model building were age, preoperative respiratory support, ASA grade, specialty and revised cardiac risk index (RCRI) score (Figure 3-2. B).

3.2.2.3 Model building and validation

The combination of these 5 selected features yields 26 different multivariable sets of predictors to evaluate in three different modelling approaches (78 final models per bootstrap). To perform model building, the derivation set was split into a 25% testing and 75% training partition (Figure 3-1). This process was bootstrapped (or randomly repeated) 100 times and for each of the bootstraps, all 26 feature combinations were fitted using three different algorithms (logistic regression, decision trees and random forest). The first two algorithms were chosen because of their relative simplicity for interpretation and well-known use in medical settings as well as their contrasting methodology given they focus on different types of interactions (linear vs non-linear) (Rosenfeld et al. 2020). The third model, random forest, was chosen given it can be considered a complex extension of decision trees and so acts as a useful reference for oversimplification in model performance assessment. A visual summary of all the operational algorithms performed can be seen in (Figure3-3A), with the evaluation of all feature combinations (runs 1-26) of the three different modelling methods at both training and testing partitions. Hyperparameters in each of these models were also tuned through grid search and 10-fold cross validation, with results available in the supplementary figures of the paper (COVIDSurg Collaborative 2021).

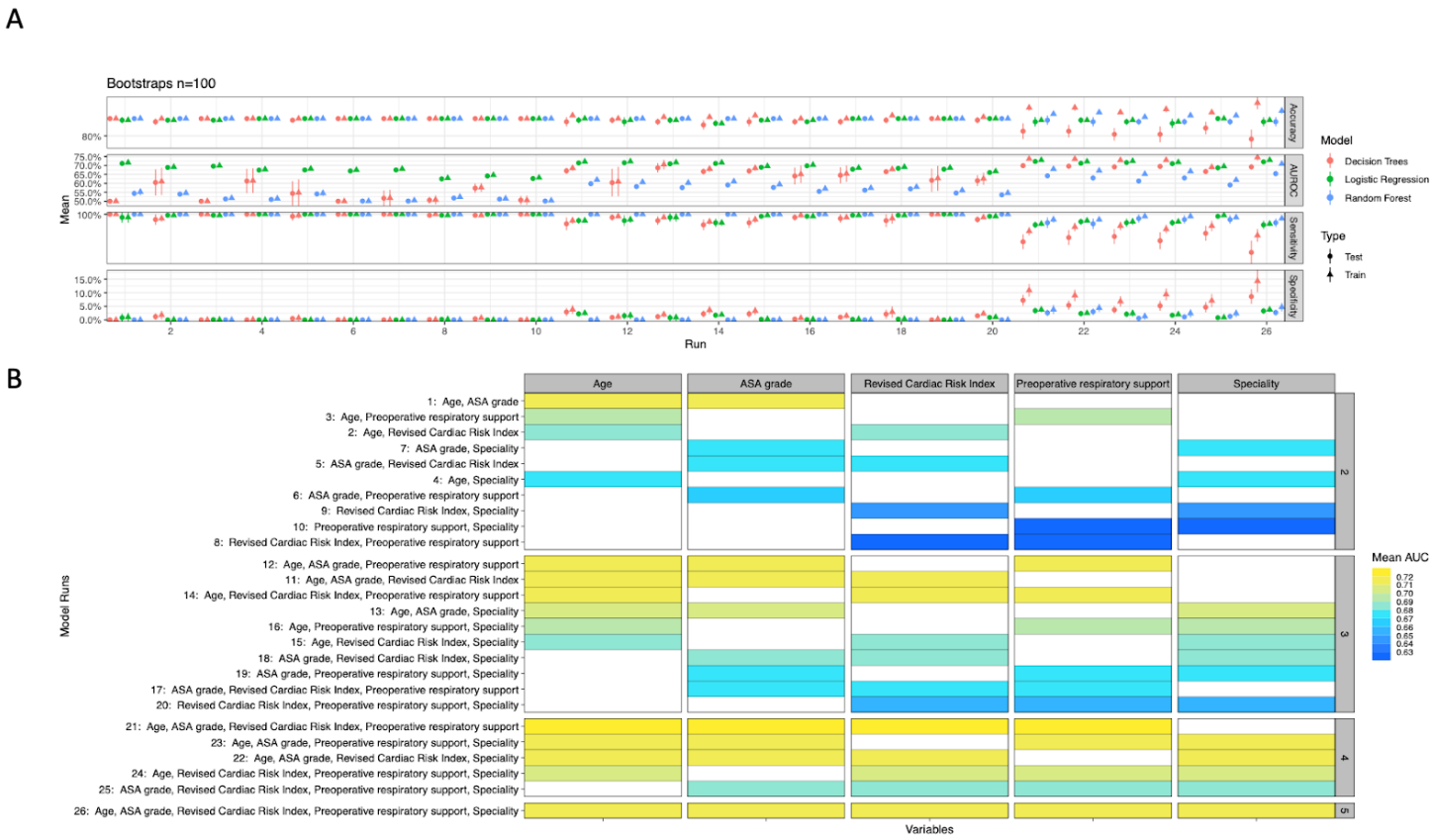


Figure 3-3. Model performance and evaluation
A) Mean and standard deviation of accuracy, AUROC, sensitivity and specificity of 10 fold cross-validated 78 models (26 feature combinations in 3 different algorithms) for 100 bootstraps in both train and test set. B) Tile map of the performance results (AUROC) from the logistic regression modelling with different runs divided by each of the features. Colour coded according to performance values (mean AUROC of 100 bootstraps)

For the final decision on the best model combination, performance was evaluated through the mean AUROC value of the 100 bootstraps**.** The standard deviation of our results was also considered in order to assess the stability of our models, as well as sensitivity, specificity and accuracy. We decided a priori to choose the simplest model (less number of independent features) with the highest and most robust performance. Also, to guide model choice, we provided a visual overview of the performance associated with particular features in all logistic regression models where for instance, models without age can be seen to perform worse (e.g runs 8, 11, 25). Complete modelling results can be found in the supplementary data of (COVIDSurg Collaborative 2021) with the top 15 performing models found in Table 3-2.

Model	Run	Mean AUROC	SD AUROC	Variable names	Length
Logistic Regression	21	0.722	0.011	Age, ASA grade, RCRI score, Preoperative respiratory support	4
Logistic Regression	26	0.72	0.011	Age, ASA grade, RCRI score, Preoperative respiratory support, Specialty	5
Logistic Regression	23	0.716	0.012	Age, ASA grade, Preoperative respiratory support, Specialty	4
Logistic Regression	12	0.715	0.012	Age, ASA grade, Preoperative respiratory support	3
Logistic Regression	11	0.714	0.013	Age, ASA grade, RCRI score	3
Logistic Regression	22	0.712	0.012	Age, ASA grade, RCRI score, Specialty	4
Logistic Regression	14	0.711	0.01	Age, RCRI score, Preoperative respiratory support	3
Logistic Regression	1	0.711	0.013	Age, ASA grade	2
Logistic Regression	24	0.71	0.01	Age, RCRI score, Preoperative respiratory support, Specialty	4
Logistic Regression	13	0.709	0.013	Age, ASA grade, Specialty	3
Decision Trees	21	0.699	0.014	Age, ASA grade, RCRI score, Preoperative respiratory support	4
Logistic Regression	16	0.698	0.011	Age, Preoperative respiratory support, Specialty	3
Decision Trees	22	0.696	0.013	Age, ASA grade, RCRI score, Specialty	4
Logistic Regression	3	0.696	0.012	Age, Preoperative respiratory support	2
Decision Trees	24	0.695	0.013	Age, RCRI score, Preoperative respiratory support, Specialty	4

Table 3-2. Top performing models
Top 15 models (combination of features and machine learning method), ranked by mean AUROC, evaluated in the test partition of the derivation. The best performing model as evaluated by the highest AUROC was then selected as the final model (Run 21, made up of logistic regression of age ASA grade, RCRI score and preoperative respiratory support).

AUROC: Area under the receiving operator characteristic curve; ASA: American Society of Anaesthesiologists

A logistic regression model with four features (age, ASA grade, RCRI score and preoperative respiratory support) was finally selected as our model of choice and our final model was then built by averaging the coefficients of the 100 models produced (Teschendorff 2019; X. Chen et al. 2020) (Table 3-3) and applied to both the whole derivation and validation sets.

Variables	**Mean**	**Standard Deviation**	**Odds**
Intercept	-3.906	0.158	0.020
Age: 40-49 years	0.677	0.138	1.968
Age: 50-59 years	0.875	0.111	2.400
Age: 60-69 years	1.313	0.106	3.719
Age: 70-79 years	1.623	0.106	5.068
Age: 80-89 years	1.869	0.106	6.481
Age: ≥90 years	2.215	0.118	9.165
ASA grade: Grade 2	0.320	0.132	1.377
ASA grade: Grade 3	0.705	0.133	2.025
ASA grade: Grades 4-5	1.236	0.139	3.440
RCRI score: 1	0.154	0.058	1.167
RCRI score: 2	0.456	0.063	1.5787
RCRI score: 3	0.509	0.079	1.663
RCRI score: ≥4	0.738	0.099	2.091
Preoperative respiratory support: oxygen	0.273	0.043	1.313
Preoperative respiratory support: ventilated	0.727	0.055	2.070

Table 3-3. Final logistic model coefficients
Obtained by averaging the coefficients generated in each bootstrap (100) for the top performing model (run 21, i.e., logistic regression of features age, ASA grade, RCRI score and preoperative respiratory support) ASA: American Society of Anaesthesiologists; RCRI: Revised Cardiac Risk Index

Performance was assessed through both discrimination (AUROC) and calibration (Brier score, slope, and calibration-in-the-large (Gupta et al. 2021; Huang et al. 2020; van Walraven 2014; Steyerberg and Vergouwe 2014)), with the AUROC for the final score being 0.73 (95%

CI: 0.71 to 0.74) in the derivation set and 0.80 (95% CI:0.77 to 0.83) in the validation set (Figure 3-4). Calibration in the validation set is shown in Figure 3-5 (Brier score 0.084, calibration-in-the-large 0.078, slope 1).

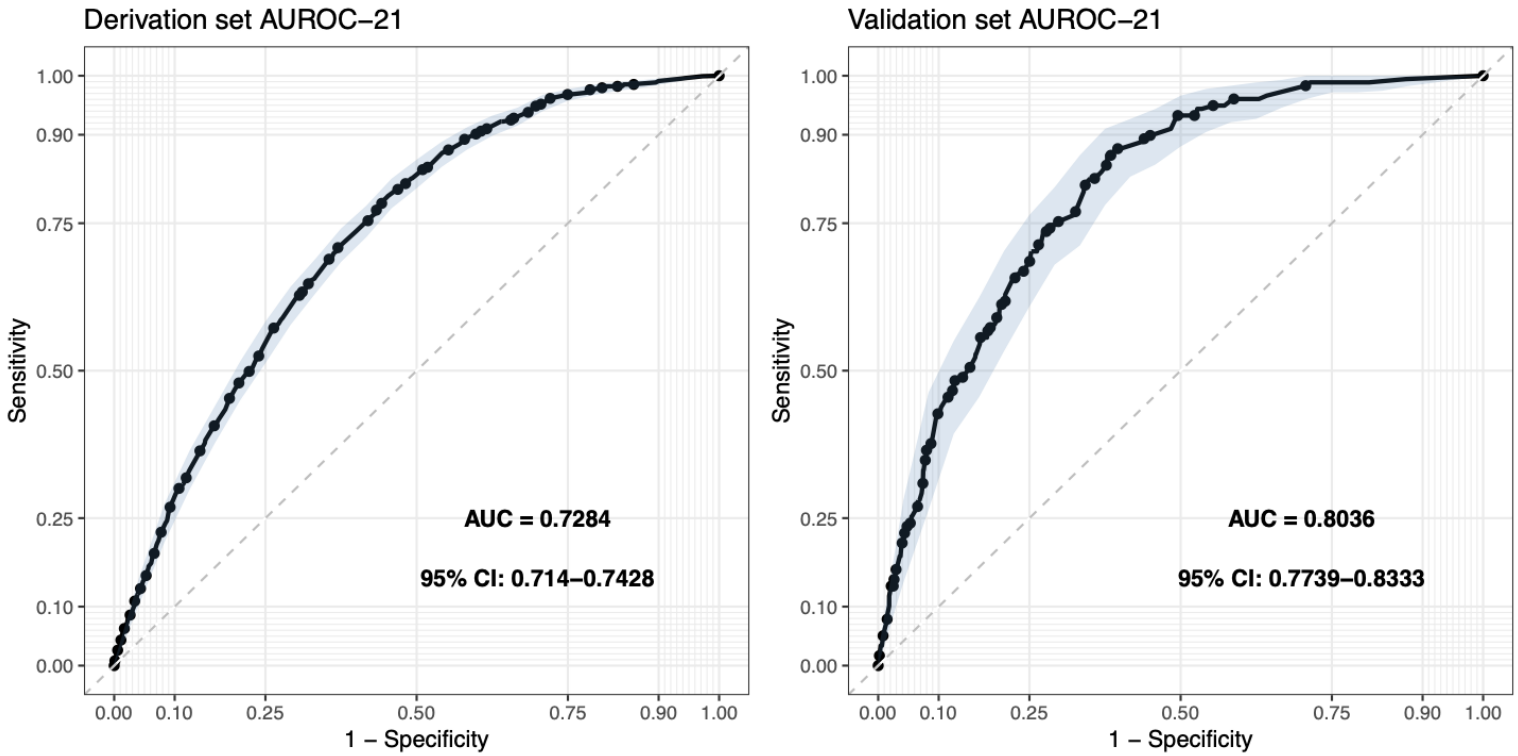
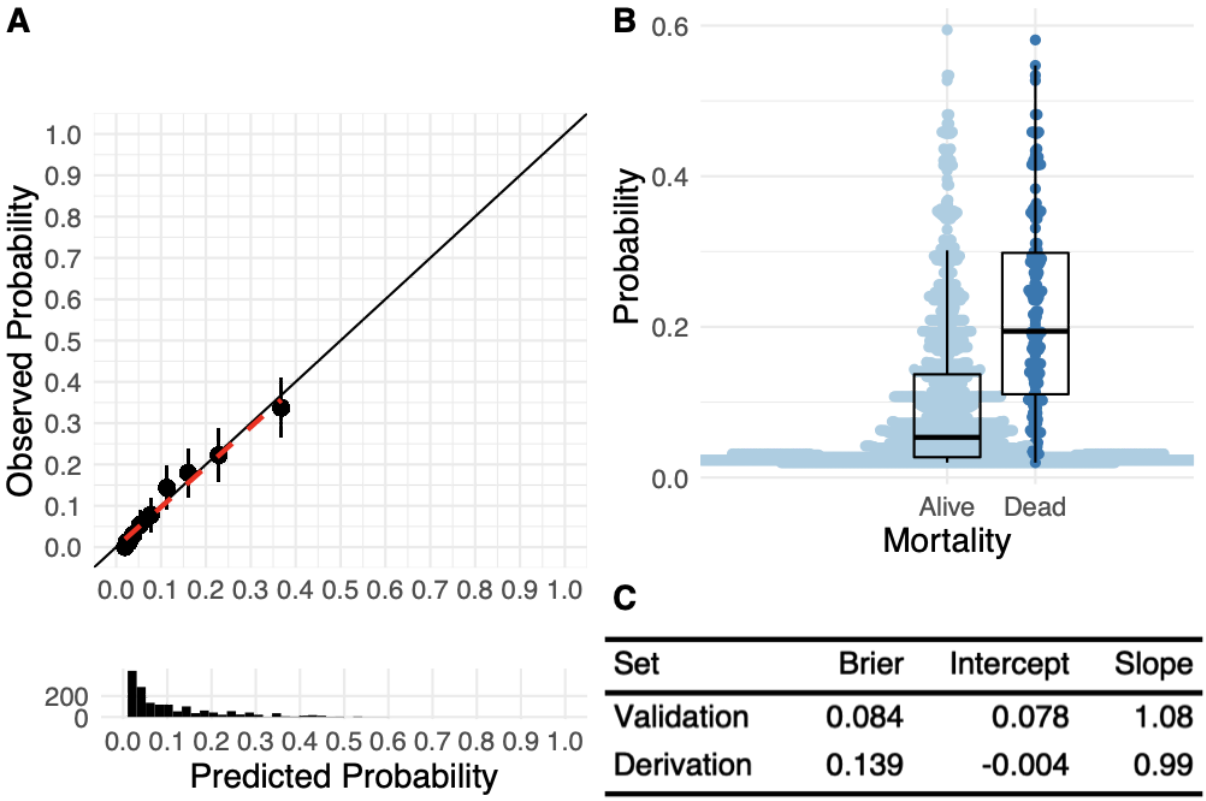


Figure 3-4. Discrimination evaluation results
After generating the final model made up by the averaged coefficients of the bootstraps of run 21, it was evaluated in both the derivation set as a whole and the validation set. Results are depicted with AUROC and confidence intervals generated through the pROC and plotROC packages where 95% CI were computed with default 2000 stratified bootstrap replicates.



C

Set	Brier	Intercept	Slope
Validation	0.084	0.078	1.08
Derivation	0.139	-0.004	0.99

Figure 3-5. Calibration evaluation results
A) Calibration plot for the predicted probabilities in the validation set. In red a loess smoothed line linking calibration at bins of 10 with confidence intervals calculated. 45 degree black angle line shows the perfect calibration plot (slope 1 , intercept/ calibration-in-the-large 0). Below we have a histogram of probability estimates. B) Boxplot and beeswarm plot of predicted probabilities in validation set, stratified by outcome. C) Assessment of calibration in both validation and derivation sets (Brier score, slope and intercept/ calibration-in-the-large evaluated through val.prob function from rms package)

Finally, in order to establish the different risk cut-off values, case fatality ratios at different model probability output's thresholds were assessed in the validation set and checked in the derivation set (Knight et al. 2020) (Table 3-4). These thresholds (~ 2%, ~10%, ~30%, and ~45% mortality), were influenced by distribution of the data and the clinical importance of various cut-offs. The whole dataset, merging both validation and derivation sets was then used to report the final case fatality percentages.

Derivation										
Cut-off Value	**Patients (%)**	**TP**	**TN**	**FP**	**FN**	**Sensitivity (%)**	**Specificity (%)**	**PPV (%)**	**NPV (%)**	**Mortality (%)**
≤0.05	1031 (15.93)	1202	1008	4241	23	98.1	19.2	22.1	97.8	2.2
≤0.2	3683 (56.89)	861	3319	1930	364	70.3	63.2	30.8	90.1	9.9
≥0.2	2791 (43.11)	861	3319	1930	364	70.3	63.2	30.8	90.1	30.8
≥0.5	85 (1.31)	40	5204	45	1185	3.3	99.1	47.1	81.4	47.1
Validation										
≤0.05	723 (43.35)	166	711	779	12	93.3	47.7	17.6	98.3	1.7
≤0.2	1366 (81.89)	88	1276	214	90	49.4	85.6	29.1	93.4	6.6
≥0.2	302 (18.11)	88	1276	214	90	49.4	85.6	29.1	93.4	29.1
≥0.5	8 (0.48)	4	1486	4	174	2.2	99.7	50.0	89.5	50.0

Table 3-4. Performance metrics of model probability output at different cut-off values in both derivation and validation cohorts.
TP: True positive; TN: True negative; FP: False positive; FN: False negative; PPV: Positive predictive value; NPV: Negative predictive value. Mortality = estimated absolute mortality risk for all patients in group.

From the model and risk cut off we then derived this simple visualization (Figure 3-6) to aid in clinical decision making, including it in the COVIDSurg Mortality Score interface, where the model has been deployed online too (accessible through either http://covidsurgrisk.app/ or https://inflam.shinyapps.io/RiskAppLukeAugust2/) as a risk calculator/stratification tool. Given the objective is to inform surgeons, patients, and healthcare decision makers to reach a better understanding of risk when treating surgical patients during the COVID-19 pandemic, a public tutorial to present it was performed too and is accessible via (https://www.youtube.com/watch?v=a0McrK-NnE)

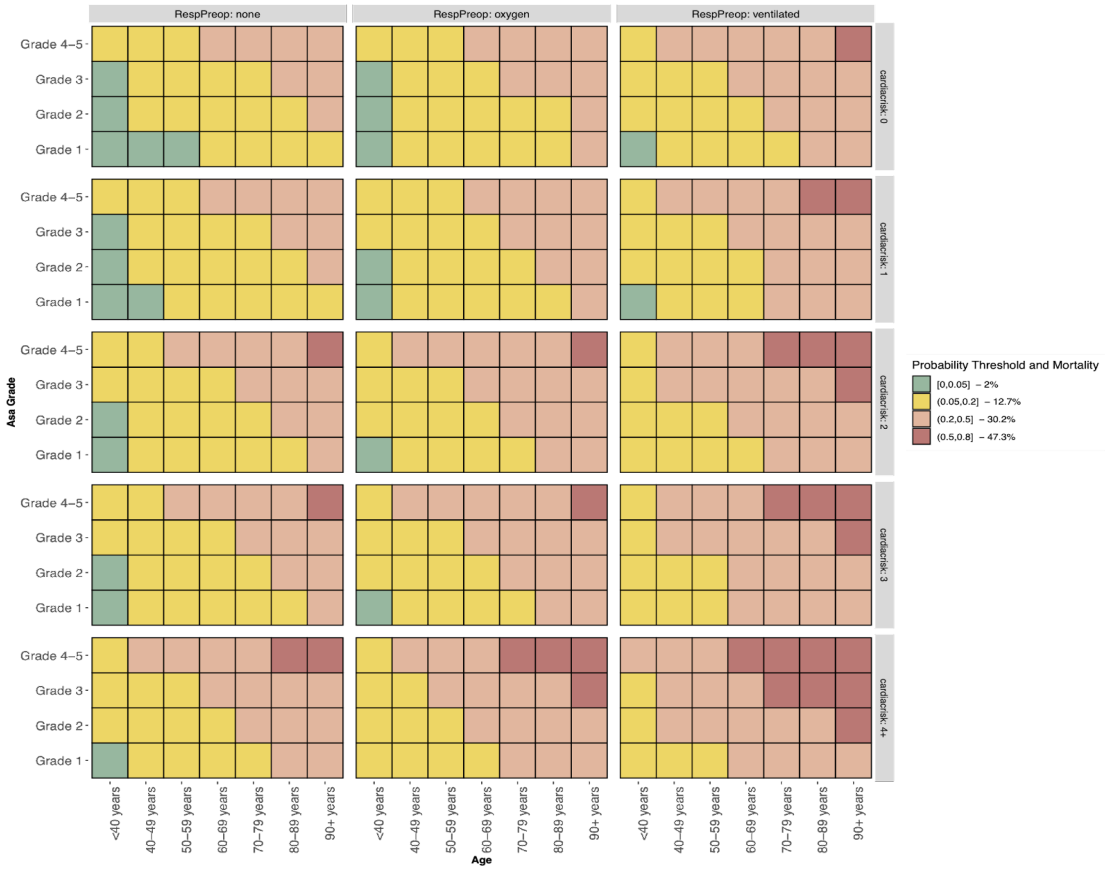


Figure 3-6. Tile plot

Tile plot of Table 3-4 cut-off values for the complete cases dataset (n=8492), whereby case fatality rates are associated to a range of probability outcomes, grouping patients into a color coded description of risk. Probability outcome range is from 0 to 0.8 and are split into 4 groups corresponding to case fatality ratios of approximately 2%, 10 %, 30% and 50% as seen in all three derivation, validation and complete datasets . ASA: America Society of Anaesthesiologists; RCRI: Revised Cardiac Risk Index

This prediction model is reported in alignment with TRIPOD (Moons et al. 2015; Jalali, DiGennaro, and Sridhar 2020) and PROBAST (Wolff et al. 2019) guidelines (supplementary data of (COVIDSurg Collaborative 2021)). All analysis was performed in R Foundation for Statistical Computing V4.0 (Vienna, Austria) version 4.00 and RStudio through the University of Birmingham's BlueBEAR High Performance Computing service using rslurm, drake and all packages mentioned above. To aid transparency in research, all scripts and code used have been uploaded to github (https://github.com/InFlamUOB/CovidSurg) with data available upon reasonable request. The code reported contains an Rmarkdown example (.html) as well as figures demonstrating the analysis pipeline.

3.3 Risk of pulmonary complications in adults undergoing elective surgery: Development and validation of a prognostic model to support safe global recovery of elective surgery (GlobalSurg-Pulmonary Score)

3.3.1 Summary

Pulmonary complications are a common and lethal sequela of surgery and as elective surgery is upscaled to address a growing global backlog during pandemic recovery, they pose a major threat to patient safety (COVIDSurg Collaborative 2022; Simoes, Bhangu, and CovidSurg Collaborative 2020). More specifically, pulmonary complications have been implicated in 37% of postoperative deaths in the pre-pandemic period, and as many as 66% of deaths during the pandemic due to additional risks of perioperative SARS-CoV-2 infection (STARSurg Collaborative and COVIDSurg Collaborative 2021). Several prognostic risk models for postoperative pulmonary complications have been reported in the literature (Canet et al. 2010; 2015) but, in a recent systematic review, few existing models have been found to be externally validated with none demonstrating acceptable performance (AUROC>0.7) (STARSurg Collaborative and TASMAN Collaborative 2022). Also, all risk

scores that reported external validation used at least one complex variable that may not be available in the outpatient clinic at the time of decision for elective surgery, or in resource-limited environments. No existing scores reflect additional risk to patients during local outbreaks of SARS-CoV-2 for patients undergoing surgery (Wynants et al. 2020).

In this study we aimed to derive and validate a novel prognostic model that can be used before elective surgery to estimate the risk of postoperative pulmonary complications at 30-days after surgery and support resource allocation and prioritisation during pandemic recovery. For this, 13 candidate predictor variables in data from 86,231 patients were used (Collaborative and Collaborative 2021). To build the model, both LASSO, a regularized linear model and XGBoost were tested. Given they both displayed a similar discrimination performance (AUROC 0.786, 95% CI:0.774-0.798 versus 0.785, 95% CI:0.772-0.797 respectively), the LASSO model was selected given it is more explainable and simpler.The final GlobalSurg-Pulmonary Score included ten predictor variables and demonstrated good discrimination and calibration upon internal validation (AUROC 0.7729, 95%CI 0.7509-0.7948; Brier score: 0.020, calibration-in-the-large: 0.072, slope: 0.930). The model was then externally validated through two different datasets (CovidSurg-Cance and RECON) (COVIDSurg Collaborative 2020; J. Glasbey et al. 2021; STARSurg Collaborative and COVIDSurg Collaborative 2021; Collaborative and Collaborative 2021), with an acceptable performance in CovidSurg-Cancer (AUROC 0.746, 95% CI: 0.733-0.760; Brier score: 0.036, CITL: 0.324, slope: 1.002), but with some miscalibration in RECON data (AUROC: 0.690, 95% CI 0.663-0.719; Brier score: 0.045, CITL: 0.711, slope: 0.875). The model was deployed to an online web application: https://surgicalrisk.app with the hope to inform surgical consent, resource allocation, and hospital level prioritisation.

3.3.2 Methods and results

3.3.2.1 Data: Patient characteristics and outcomes

For this study, data from three international prospective cohort studies on elective surgeries in adult patients (≥18 years) across all operation and disease types was used (in total 126,406 patients from 119 countries and 1158 hospitals). All these datasets have been explored previously, with information on ethics, approvals, data governance and data quality already published (COVIDSurg Collaborative 2020; J. Glasbey et al. 2021; STARSurg Collaborative and COVIDSurg Collaborative 2021; Collaborative and Collaborative 2021).

The first dataset, GlobalSurg-CovidSurg week study (global, October 2020, NCT04509986, n=87, 863))(Collaborative and Collaborative 2021), was used for model development. The other two, CovidSurg-Cancer (Worldwide, February to August 2020, NCT04384926, n=31, 205)) (J. Glasbey et al. 2021), and RECON (UK and Australasia, January to October 2019, n=7, 342) (STARSurg Collaborative 2020) were used for external validation. For all datasets, the overall rate of missing data was very low (<0.1% overall and <2% for all candidate predictor variables) and appeared to be missing at random, so a complete case analysis was conducted. The complete case analysis included 86,231 patients in model development, and 30,488 from CovidSurg-Cancer and 6789 from RECON in external validation (Figure 3-7). The adequacy of our sample size was tested using an estimate of the predicted AUROC (0.7) and prevalence (3.0%) from a contemporary systematic review (STARSurg Collaborative and TASMAN Collaborative 2022), where we fulfilled the estimated minimum sample size requirement of 21,299 patients, and 639 events (Riley et al. 2019).

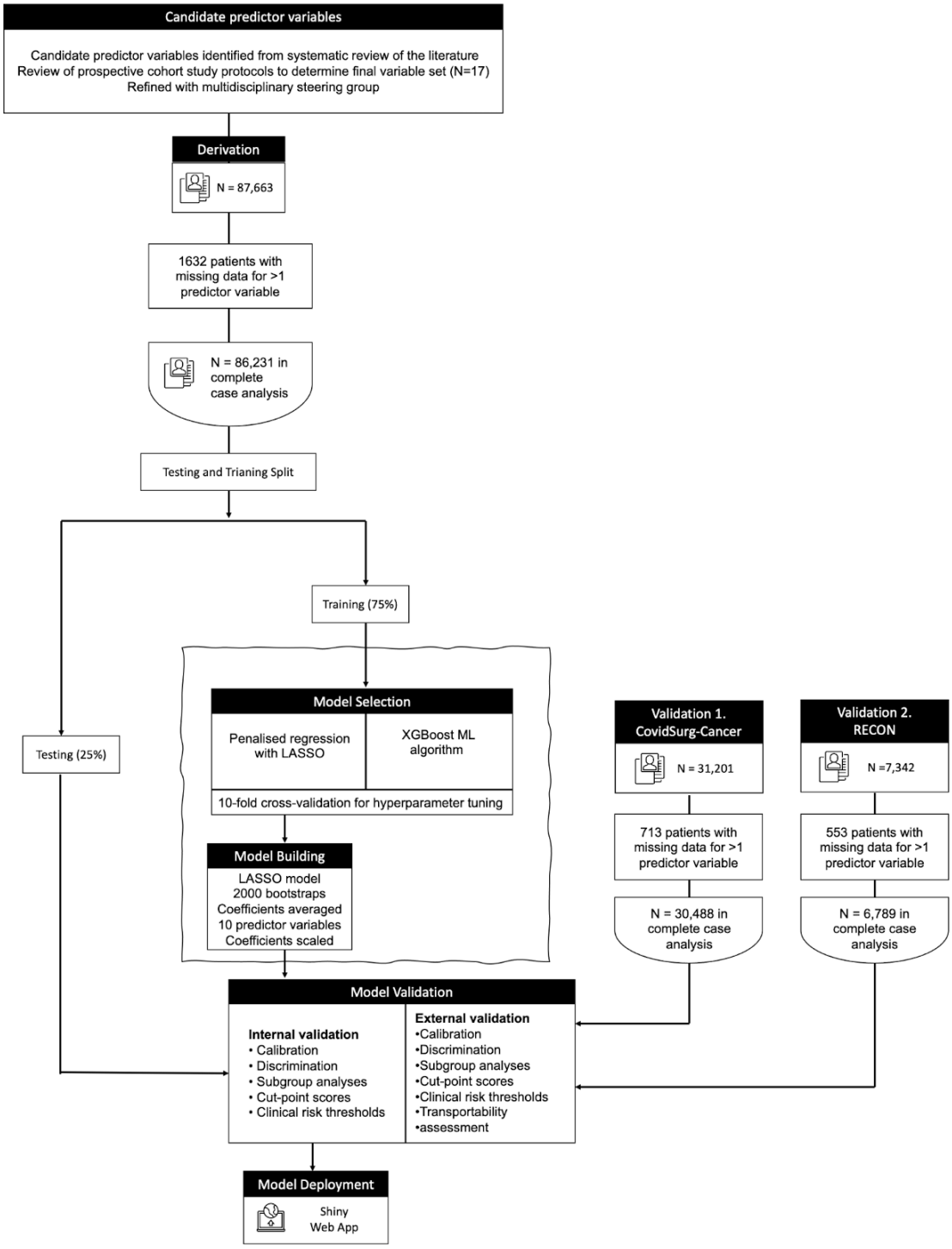


Figure 3-7. Workflow description
Cohort study, patient inclusion and model derivation and validation description

The primary outcome measure was postoperative pulmonary complications (PPC) within 30-days of surgery (with day 0 as the day of surgery). This was defined as a composite of

pneumonia, acute respiratory distress syndrome and unplanned use of invasive or non-invasive mechanical ventilation.Candidate predictor variables were identified in a systematic review of existing risk scores and refined with multidisciplinary clinical input from surgery, anaesthesia and critical care experts. In order to improve the adoption of the risk stratification tool, we planned to only include predictor variables that are available globally without the need for additional tests, and that were relevant to resource-limited environments (STARSurg Collaborative and TASMAN Collaborative 2022). A summary of all variables and their distribution per dataset can be seen in Table 3-5, where we can see that the overall rates of pulmonary complications were 2.0% in the derivation cohort, and 3.9% (CovidSurg-Cancer) and 4.7% (RECON) in the validation cohorts.

Features	Levels	Derivation GlobalSurg/ CovidSurg-Week N=86231	Validation CovidSurg-Cancer N=30488	Validation RECON N=6789
Patient factors				
Age	18-29 years	9106 (10.6%)	636 (2.1%)	276 (4.1%)
	30-39 years	12176 (14.1%)	1708 (5.6%)	555 (8.2%)
	40-49 years	13272 (15.4%)	3744 (12.3%)	983 (14.5%)
	50-59 years	15703 (18.2%)	6697 (22.0%)	1466 (21.6%)
	60-69 years	16665 (19.3%)	8345 (27.4%)	1623 (23.9%)
	70-79 years	13742 (15.9%)	6953 (22.8%)	1390 (20.5%)
	80-89 years	5110 (5.9%)	2265 (7.4%)	484 (7.1%)
	≥90 years	457 (0.5%)	140 (0.5%)	12 (0.2%)
Sex	Female	47219 (54.8%)	18325 (60.1%)	3905 (57.5%)
	Male	39012 (45.2%)	12163 (39.9%)	2884 (42.5%)
ASA grade	Grade 1	21667 (25.1%)	6431 (21.1%)	1048 (15.4%)
	Grade 2	43099 (50.0%)	15752 (51.7%)	3903 (57.5%)
	Grade 3	19057 (22.1%)	7912 (26.0%)	1733 (25.5%)
	Grade 4/5	2408 (2.8%)	393 (1.3%)	105 (1.5%)
RCRI	0	38541 (44.7%)	9244 (30.3%)	4112 (60.6%)
	1	37688 (43.7%)	15893 (52.1%)	2210 (32.6%)
	2	7595 (8.8%)	4284 (14.1%)	448 (6.6%)
	≥3	2407 (2.8%)	1067 (3.5%)	19 (0.3%)
Pre-existing respiratory disease	No	77747 (90.2%)	27278 (89.5%)	5611 (82.6%)
	Yes	8484 (9.8%)	3210 (10.5%)	1178 (17.4%)
Current smoker	No	72830 (84.5%)	27307 (89.6%)	5852 (86.2%)
	Yes	13401 (15.5%)	3181 (10.4%)	937 (13.8%)
Disease factors				
Compartment	Abdominopelvic	43761 (50.7%)	16816 (55.2%)	6789 (100.0%)
	Head & Neck	8228 (9.5%)	3823 (12.5%)	N/A
	Limb	12514 (14.5%)	196 (0.6%)	N/A
	Other	17516 (20.3%)	7225 (23.7%)	N/A
	Thoracic	4212 (4.9%)	2428 (8.0%)	N/A
Indication	Benign	64822 (75.2%)	N/A	3884 (57.2%)
	Cancer	21409 (24.8%)	30488 (100.0%)[$]	2905 (42.8%)
Preoperative SARSCoV2 test	Negative	61339 (71.1%)	14128 (46.3%)	6789 (100.0%)*
	Not performed	24892 (28.9%)	16360 (53.7%)	N/A
Operation factors				
Anaesthesia type	General	63235 (73.3%)	28767 (94.4%)	6789 (100.0%)
	Regional/local	22996 (26.7%)	1721 (5.6%)	N/A
Operation grade	Major	53061 (61.5%)	24435 (80.1%)	6789 (100.0%)
	Minor	33170 (38.5%)	6053 (19.9%)	N/A
Location factors				
Hospital type	COVID19 free surgical pathway	71809 (83.3%)	8097 (26.6%)	6789 (100.0%)
	No defined pathway	14422 (16.7%)	22391 (73.4%)	N/A
Country income	High income	59284 (68.8%)	24362 (79.9%)	6789 (100.0%)
	Upper middle income	14614 (16.9%)	3385 (11.1%)	N/A
	Low middle income	12333 (14.3%)	2741 (9.0%)	N/A
Community SARS-CoV-2 risk	High	57229 (66.4%)	16767 (55.0%)	N/A
	Low	29002 (33.6%)	13721 (45.0%)	6789 (100.0%)*
Country income * SARS-CoV-2 risk	High/High	48375 (56.1%)	14811 (48.6%)	N/A
	High/Low	10909 (12.7%)	9551 (31.3%)	6789 (100.0%)*
	Uppermiddle/High	8339 (9.7%)	1309 (4.3%)	N/A
	Uppermiddle/Low	6275 (7.3%)	2076 (6.8%)	N/A
	Low middle/High	515 (0.6%)	647 (2.1%)	N/A

	Low middle/Low	11818 (13.7%)	2094 (6.9%)	N/A
Primary outcome measure				
Postoperative pulmonary complication	Negative	84450 (97.9%)	29303 (96.1%)	6471 (95.3%)
	Positive	1781 (2.1%)	1185 (3.9%)	318 (4.7%)

Table 3-5. Comparison of derivation and validation cohort datasets.
Assessment of differences in case-mix and outcomes across the three cohorts In the derivation cohort, 2.1% (1781/86231) of patients had a pulmonary complication. The 30-day PPC rate was higher in both the CovidSurg-Cancer (3.9%, 1185/30492) and RECON (4.7%, 318/6789) validation studies. The postoperative mortality rate was (0.7%, 601/86231, 3 missing) in the derivation cohort, (1.05%, 320/30492, 24 missing) in CovidSurg-Cancer and (0.4%, 27/6789, 223 missing) in RECON.

3.3.2.2 Model selection and building

Following previous work (Knight et al. 2020), we decided to just test two different algorithms to model this data (Figure 3-7). The first model consisted of LASSO, the penalised logistic regression model explained in Chapter 2, where logistic regression with an extra penalty term shrinks covariate coefficients towards zero, allowing for the generation of sparse models and concurrently performing feature selection (Tibshirani 1994). The second model applied was gradient boosting decision trees (XGBoost), which is commonly recognised as a best-in-class machine learning algorithm given it can acknowledge more complex interactions and so is hypothesized to obtain better predictive capabilities (Knight et al. 2020). However, because of its complexity, the ability to explain and understand predictive reasoning is hindered (‘black box phenomenon’) (Caruana et al. 2015; The Lancet Respiratory Medicine 2018)). Hyperparameters of learning rate, maximum tree depth, loss reduction and minimum number were tuned, with trees set to 1000. The derivation set was split into a random 75% (training) and 25% (testing) split, and both models were fit through 10-fold cross-validation and parameter tuning using tune race analysis of variance (ANOVA) and grid search in the training set. Confidence intervals were assessed from the cross validated models using the cvAUC package (LeDell, Petersen, and Laan 2022). Similar discrimination (Figure 3-8. A) was observed between both LASSO (AUROC: 0.786, 95% CI 0.774-0.798) and XGBoost

(AUROC 0.785, 95% CI 0.772-0.797) models with variable importance observed in both modelling approaches summarised in (Figure 3-8. B). Given the comparable performance, more explainable nature and increased simplicity of LASSO, we proceeded to develop this model as the GlobalSurg-Pulmonary score.

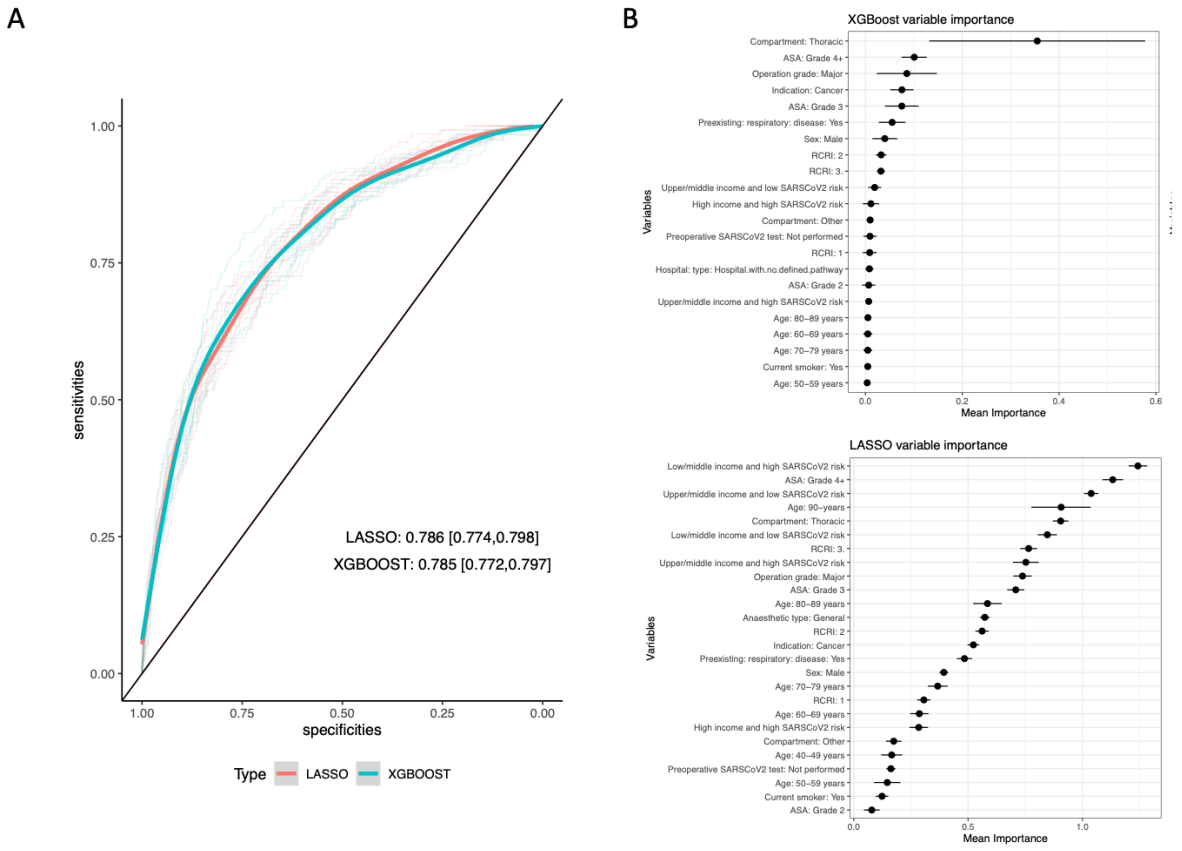


Figure 3-8. Comparison of LASSO and XGBoost algorithms during model selection

A)ROC curves with confidence intervals assessed from the cross validated models using the cvAUC package. B)Feature Importance assessment performed through the VIP package and averaged in each resample. In LASSO, the absolute value of the estimated coefficient is used as a proxy measure and in XGBoost, importance is determined based on specific decision trees and tree ensemble particularities (i.e improvement upon partition with specific variables)

Following the same methodology used for the CovidSurg mortality score (Table 3-3), after resampling with 2000 bootstraps, beta coefficients were averaged and two variables with zero coefficients were dropped (X. Chen et al. 2020; Teschendorff 2019). The penalised regression coefficients and scaling for the pragmatic clinical scoring system, which included 10 predictor variables, are summarised in Table 3-6. For clinical application, individual point scores for each risk factor level are summed to give an overall point score out of a maximum of 23.

Features	Levels	No PPC	PPC	OR (univariable)	Mean coefficients (LASSO)	OR (LASSO)	Point score
Age	18-29 years	6751 (10.7)	66 (5.0)	1.01 (0.73-1.40)	0.019 (0.012-0.025)	1.019	0
	30-39 years	9030 (14.3)	87 (6.6)	-			
	40-49 years	9822 (15.5)	154 (11.6)	1.63 (1.25-2.13)	0.176 (0.17-0.182)	1.192	1
	50-59 years	11571 (18.3)	227 (17.1)	2.04 (1.59-2.62)	0.158 (0.153-0.163)	1.171	0
	60-69 years	12122 (19.1)	335 (25.3)	2.87 (2.27-3.66)	0.299 (0.293-0.304)	1.348	1
	70-79 years	10035 (15.8)	309 (23.3)	3.20 (2.53-4.08)	0.382 (0.376-0.388)	1.465	1
	80-89 years	3686 (5.8)	133 (10.0)	3.75 (2.86-4.93)	0.594 (0.587-0.601)	1.811	2
	90+ years	331 (0.5)	14 (1.1)	4.39 (2.37-7.55)	0.882 (0.867-0.897)	2.416	3
Sex	Female	34908 (55.1)	535 (40.4)	-			
	Male	28440 (44.9)	790 (59.6)	1.81 (1.62-2.03)	0.393 (0.39-0.395)	1.481	1
ASA grade	Grade 1	16031 (25.3)	143 (10.8)	-			
	Grade 2	31963 (50.5)	443 (33.4)	1.55 (1.29-1.88)	0.086 (0.082-0.09)	1.09	0
	Grade 3	13712 (21.6)	570 (43.0)	4.66 (3.89-5.62)	0.715 (0.71-0.719)	2.044	2
	Grade 4+	1642 (2.6)	169 (12.8)	11.54 (9.19-14.51)	1.138 (1.131-1.144)	3.12	3
RCRI	0	28632 (45.2)	263 (19.8)	-			
	1	27661 (43.7)	610 (46.0)	2.40 (2.08-2.78)	0.311 (0.307-0.315)	1.365	1
	2	5415 (8.5)	303 (22.9)	6.09 (5.15-7.21)	0.566 (0.561-0.571)	1.761	2
	3+	1640 (2.6)	149 (11.2)	9.89 (8.03-12.15)	0.769 (0.763-0.775)	2.158	2
Pre-existing respiratory disease	No	57217 (90.3)	1052 (79.4)	-			
	Yes	6131 (9.7)	273 (20.6)	2.42 (2.11-2.77)	0.483 (0.48-0.486)	1.621	1
Current smoker	No	53592 (84.6)	1071 (80.8)	-			
	Yes	9756 (15.4)	254 (19.2)	1.30 (1.13-1.49)	0.125 (0.122-0.129)	1.134	0
Compartment	Abdominopelvic	32281 (51.0)	649 (49.0)	-			
	Head & Neck	6043 (9.5)	110 (8.3)	0.91 (0.73-1.11)	0.073 (0.068-0.078)	1.076	0
	Limb	9208 (14.5)	111 (8.4)	0.60 (0.49-0.73)	0.054 (0.049-0.059)	1.056	0
	Other	13010 (20.5)	141 (10.6)	0.54 (0.45-0.65)	-0.173 (-0.178–0.168)	0.841	-1
	Thoracic	2806 (4.4)	314 (23.7)	5.57 (4.83-6.40)	0.904 (0.901-0.908)	2.471	3
Indication	Benign	47878 (75.6)	755 (57.0)	-			
	Cancer	15470 (24.4)	570 (43.0)	2.34 (2.09-2.61)	0.523 (0.521-0.526)	1.688	2
Preoperative SARSCoV2 test	Negative	45047 (71.1)	960 (72.5)	-			
	Not performed	18301 (28.9)	365 (27.5)	0.94 (0.83-1.06)	0.164 (0.161-0.167)	1.179	0
Anaesthetic type	Regional/local	17067 (26.9)	138 (10.4)	-			
	General	46281 (73.1)	1187 (89.6)	3.17 (2.67-3.80)	0.576 (0.572-0.58)	1.779	2
Operation grade	Minor	24750 (39.1)	197 (14.9)	-			
	Major	38598 (60.9)	1128 (85.1)	3.67 (3.16-4.29)	0.744 (0.74-0.747)	2.104	2
Hospital type	COVID-19 free surgical pathway	52788 (83.3)	1077 (81.3)	-			
	Hospital with no defined pathway	10560 (16.7)	248 (18.7)	1.15 (1.00-1.32)	-0.03 (-0.034–0.027)	0.97	0
Country income*SARSCoV2 risk	High/Low	8055 (12.7)	123 (9.3)	-			
	High/High	35563 (56.1)	705 (53.2)	1.30 (1.07-1.58)	0.289 (0.284-0.293)	1.334	1
	Low middle/High	379 (0.6)	18 (1.4)	3.11 (1.82-5.02)	1.22 (1.208-1.232)	3.387	4
	Low middle/Low	8675 (13.7)	183 (13.8)	1.38 (1.10-1.74)	0.851 (0.846-0.857)	2.343	3
	Uppermiddle/High	6121 (9.7)	142 (10.7)	1.52 (1.19-1.94)	0.758 (0.753-0.764)	2.134	2
	Uppermiddle/Low	4555 (7.2)	154 (11.6)	2.21 (1.74-2.82)	1.039 (1.033-1.044)	2.825	3
Intercept					-6.836 (-6.844–6.827)	0.001	0

Table 3-6. Model description. Results after 2000 bootstraps and averaged coefficients. To generate the final scaled calculator score, mean coefficients were multiplied by 3 and rounded. OR=odds ratio.PPC=Postoperative pulmonary complication

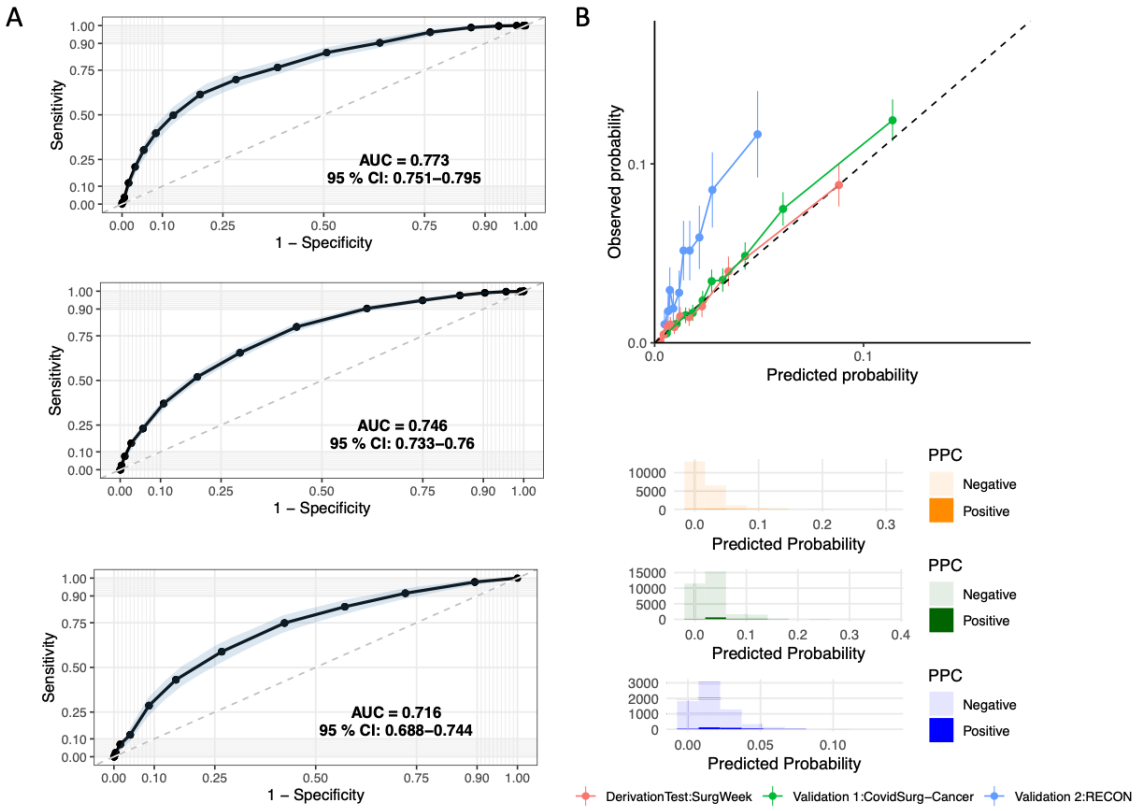


Figure 3-9. Discrimination and calibration performance for internal validation and external validation datasets.
A) Receiver operating characteristic (ROC) curves for model evaluation, confidence intervals generated through the pROC and plotROC packages where 95% CI were computed with default 2000 stratified bootstrap replicates. Upper: DerivationTest-SurgWeek, middle:CovidSurg-Cancer, bottom: RECON. B) Calibration plot generated in bins of 10 with confidence intervals calculated. 45 degree black angle dashed line shows the perfect calibration plot (slope 1,intercept/ calibration-in-the-large 0). Below we have a histogram of probability estimates (log-odds).

3.3.2.3 Model validation

The model was both internally and externally validated. In internal validation, the discrimination of the GlobalSurg-Pulmonary score was acceptable in the testing split of the derivation cohort (AUROC 0.7729, 95% CI 0.7509-0.7948) (Figure 3-9A Top), with Brier score of 0.020, calibration in the large of -0.072 and slope of 0.930(Figure 3-9B).

For external validation, two large, international prospective datasets where used (CovidSurg-Cancer and RECON).The discrimination of the GlobalSurg-Pulmonary score was moderate in both the CovidSurg-Cancer (AUROC: 0.746, 95% CI 0.733-0.760)) (Figure 3-9A Middle) and RECON (AUROC: 0.691, 95% CI 0.663-0.719) (Figure 3-9A Bottom) datasets . Calibration again was acceptable in CovidSurg-Cancer (Brier score: 0.036, CITL: 0.324, slope: 1.002) but with some miscalibration in RECON data (Brier score: 0.045, CITL: 0.711, slope: 0.875) demonstrating systematic underestimation of risk (Figure 3-9B). Because of this we continued exploring performance (interpreting external validation results), in specific subgroups of high clinical importance: (1) major versus minor operation grade, low (ASA grade 1-2) versus high (ASA grade $\geq$3) risk, and high versus low-middle country income (including UMIC, LMIC and LIC). Upon subgroup analysis, discrimination and calibration remained acceptable across all pre-defined groups (>0.7 AUROC)(Figure 3-10).

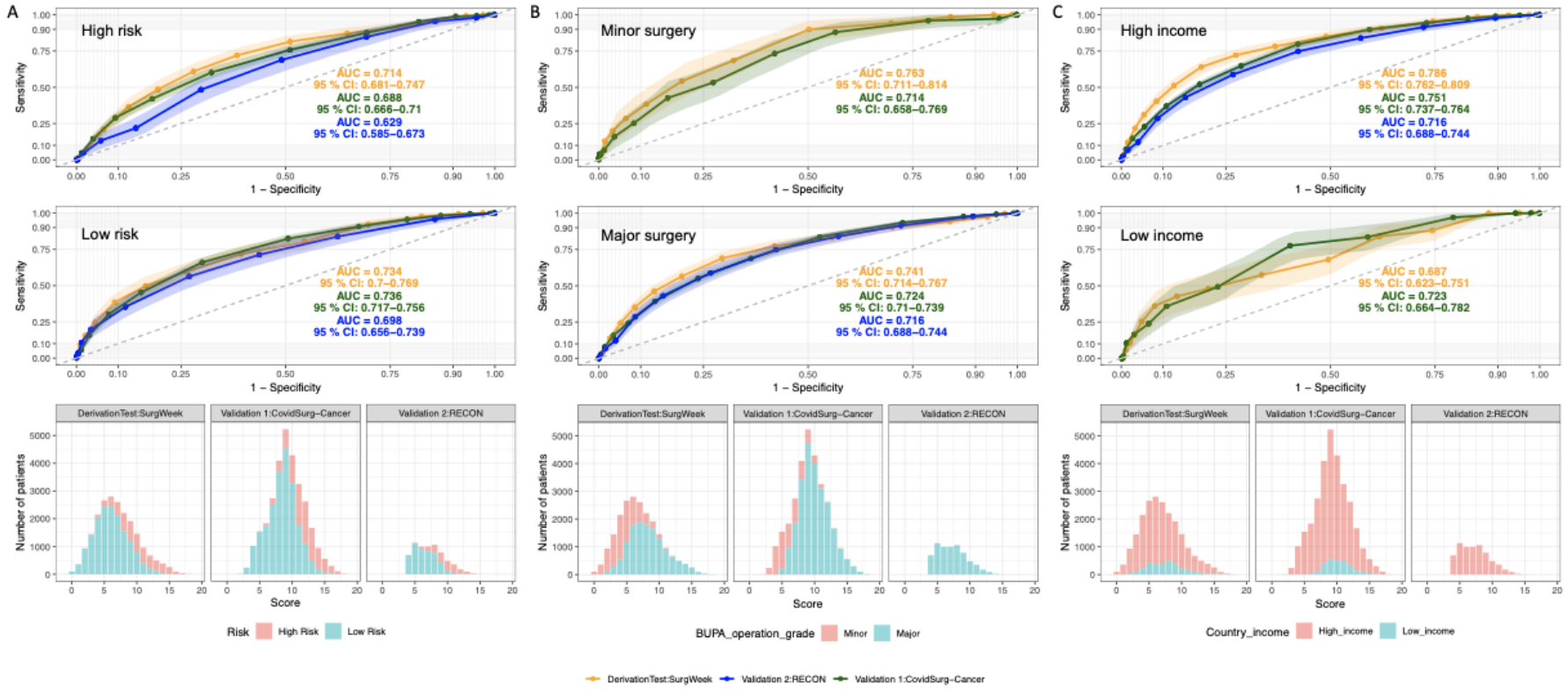


Figure 3-10. Subgroup analysis.

The model was run again on just selected categories (A) major versus minor operation grade, (B) low (ASA grade 1-2) versus high (ASA grade >3) risk, and (C) high versus low-middle country income (including UMIC, LMIC and LIC). RECON had no coded participants as minor surgeries or low income patients. Score distribution for all patients can be seen in the histogram below, per dataset.

To aid in the interpretation of the external validation results, we explored whether reproducibility or transportability was being evaluated (Debray et al. 2015). We identified that CovidSurg-Cancer had similar mix of patient and operations to the development set but included only patients with a malignant indication for surgery. RECON included only patients undergoing major abdominopelvic surgery under general anaesthesia in high income countries with no (imputed as 'low') community SARS-CoV-2 risk. Therefore, the spread of the linear predictor (log-odds predictor) was narrower in RECON (Figure 3-9B) and more like the derivation set in CovidSurg-Cancer (Figure 3-9B). Comparing the calibration and discrimination of the derivation and validation models (Figure 3-9). we concluded that CovidSurg-Cancer appeared to evaluate (statistical) reproducibility of the GlobalSurg-Pulmonary score whereas RECON appeared to favour (clinical) transportability.

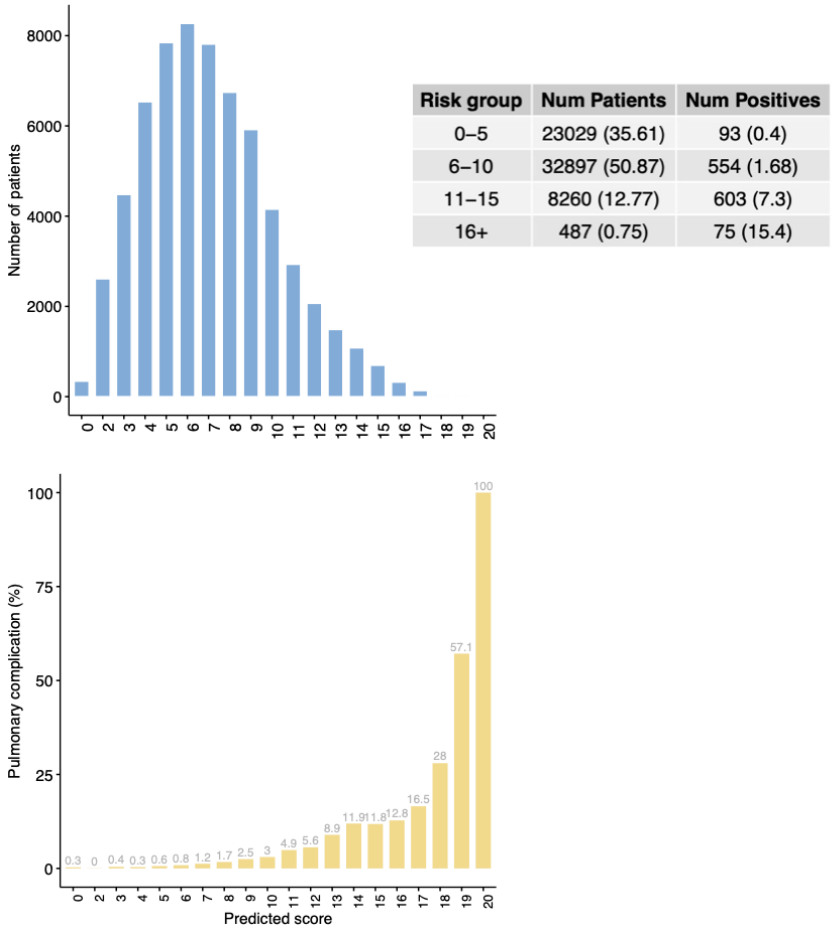
Finally, prognostic test accuracy metrics across a range of 'cut-off' point scores to 'rule-in' or 'rule-out' PPC in the derivation set are presented in Table 3-7. A cut-point score of $\leq$5 (N=23029, 35.6% of patients, PPC rate 0.4%) ruled out PPC with a sensitivity of 93.05%, specificity of 35.82%, positive predictive value of 2.96 and negative predictive value of 99.95.

.

Cut-off point score	Patients N (%)	TP	TN	FP	FN	Sensitivity (%)	Specificity (%)	PPV (%)	NPV (%)	Pulmonary complications (%)
'Rule out' pulmonary complications										
≤5	23029 (35.61)	1232	22556	40412	92	93.05	35.82	2.96	99.59	0.4
≤7	39111 (60.48)	1066	38472	24496	258	80.51	61.1	4.17	99.33	0.66
≤9	51773 (80.05)	804	50872	12096	520	60.73	80.79	6.23	98.99	1
≤11	58858 (91.01)	535	57688	5280	789	40.41	91.61	9.2	98.65	1.34
≤13	62412 (96.5)	286	60993	1975	1038	21.6	96.86	12.65	98.33	1.66
≤15	64186 (99.25)	75	62556	412	1249	5.66	99.35	15.4	98.04	1.95
≤17	64640 (99.95)	12	62947	21	1312	0.91	99.97	36.36	97.96	2.03
≤19	64672 (100)	1	62968	0	1323	0.08	100	100	97.94	2.05
'Rule in' pulmonary complications										
≥5	49489 (76.52)	1281	14760	48208	43	96.75	23.44	2.59	99.71	2.59
≥7	33374 (51.6)	1162	30756	32212	162	87.76	48.84	3.48	99.48	3.48
≥9	18818 (29.1)	951	45101	17867	373	71.83	71.63	5.05	99.18	5.05
≥11	678	54899	8069	646	51.21	87.19	7.75	98.84	7.75	678
≥13	3748 (5.8)	419	59639	3329	905	31.65	94.71	11.18	98.51	11.18
≥15	1181 (1.83)	157	61944	1024	1167	11.86	98.37	13.29	98.15	13.29
≥17	166 (0.26)	34	62836	132	1290	2.57	99.79	20.48	97.99	20.48
≥19	8 (0.01)	5	62965	3	1319	0.38	100	62.5	97.95	62.5

Table 3-7. Prognostic test accuracy in the validation set.
TP=True positive. TN=True negative. FP=False positive. FN=False negative. PPV=Positive predictive value. NPV=Negative predictive value.

Using Table 3-7, four risk strata Figure 3-11 were defined clinically as low risk (0-5 score, PPC rate in derivation set: 0.4% (N=23029 patients), intermediate risk (6-10 score, PPC rate: 1.68% (N=32897 patients), high risk (11-15 score, PPC rate 7.30% (N=8260 patients)) and very high risk (≥16 score, PPC rate 15.4% (N=487 patients). This visualization and the final model (GlobalSurg-Pulmonary Score calculator) was deployed online at: www.surgicalrisk.app.



Risk group	Num Patients	Num Positives
0–5	23029 (35.61)	93 (0.4)
6–10	32897 (50.87)	554 (1.68)
11–15	8260 (12.77)	603 (7.3)
16+	487 (0.75)	75 (15.4)

Figure 3-11. Predicted score in the whole derivation dataset (SurgWeek)
Top) Bar plot of all participants given their predicted scores . Bottom) Cumulative percentage of pulmonary complications with increasing scores. Table) Description of the four risk groups

This study was reported with reference to Transparent Reporting of a multivariable prediction model for Individual Prognosis or Diagnosis (TRIPOD) guidelines (Moons et al. 2015; Jalali, DiGennaro, and Sridhar 2020) .Two independent statisticians assessed risk of bias in the study using the PROBAST tool (Wolff et al. 2019) with the independent PROBAST assessment indicated that the study was at low risk of bias and there were low concerns about applicability. All analysis was performed in R Foundation for Statistical Computing V4.02 (Vienna, Austria) and RStudio using tidyverse, cvAUC, tidymodels, workflowsets, finetune, pROC, rms, glmnet, dcurves and shiny packages. To ensure transparency in research, all scripts and code used have been uploaded to github (https://github.com/gkoutos-group/globalsurg-pulmonary-score). Data sharing is available upon request after satisfactory completion of a data sharing agreement.

3.4 Conclusion and discussion

These studies define two novel risk stratification tools (CovidSurg mortality score: http://covidsurgrisk.app/) and (GlobalSurg-Pulmonary score: https://surgicalrisk.app) to estimate risk of mortality and of pulmonary complications in adults undergoing surgery during the SARS-CoV-2 pandemic. With this work we have aimed to contribute to the global response and efforts during the SARS-CoV-2 pandemic, with redeployment into pressing work exploring these unique datasets of global surgery response data. By studying these datasets with the machine learning pipeline and tools developed in Chapter 2, we have delved deeper into the use of machine learning in healthcare and navigated the specific needs of the industry, gaining a better understanding of what clinical translation entails.

Both studies count with similar methodology and criteria, focused on simple and resource-light predictor variables which are readily available to clinicians across all settings

and resources. Also, both sets of features can be calculated remotely (e.g., from electronic health records or during telemedicine consultation), with no additional tests or equipment required (in comparison to preoperative Sp02 and serum haemoglobin in other models (e.g ARISCAT score (Canet et al. 2010)) and pre-operatively (in comparison to duration of surgery and emergency procedure in others (Canet et al. 2010)).

As key strength, extra emphasis has been made on the model's validation strategies, complying with both internal validation standards through train-test split and bootstrapping (Jalali, DiGennaro, and Sridhar 2020) and external validation. External validation is a contested claim in clinical prediction models (Van Calster et al. 2023), with models argued to never being truly validated due to expected heterogeneity in model performance between locations and settings, and over time. Because of this, we have focused on quantifying this heterogeneity through subgroup analysis and transportability assays in well-powered, multi-country validation sets (Debray et al. 2015) as seen in the GlobalSurg-Pulmonary score work. For the mortality score, we were hindered by no other international, multispecialty data sets for SARS-CoV-2 surgical patients and so further external validation was not possible. Regardless, model reliability was assessed by using a temporally distinct validation cohort, in compliance with current standards (TRIPOD (Moons et al. 2015)). Also, having an international, multicentre database makes our data more generalisable and we plan to regularly monitor model performance and update it to ensure that prediction models stay up-to-date to support medical decision-making.

During the pandemic, a big surge of prognostic models happened, with most found to have a high or unclear bias with poor methodologies implemented (Wynants et al. 2020) . To the best of our knowledge, both our studies were designed and reported following best practice guidelines (TRIPOD and PROBAST) and have been shown to be at a low risk of bias.

Regardless, we acknowledge there are some limitations to our work. For instance, all clinical predictor variables were coded following standardized medical criteria and so collected categorically, a simplification that may reduce our statistical power (P, Dg, and W 2006) and may explain a degree of miscalibration in RECON data in the GlobalSurg-Pulmonary score work. More specifically, miscalibration may be due to unmeasured differences in case-mix (e.g., balance of major and complex-major surgery within the single ‘major surgery’ predictor variable), differences in practice during the pandemic (e.g., proactive risk modification) or system-level factors such as access to care and health behaviours in lower resource settings (e.g., barriers to travel to hospital after discharge, so less likely to have PPC observed). It is worth mentioning too, that regarding missing data, we have used complete case analysis in both our studies. In the CovidSurg mortality score, imputation was assessed, with comparable results achieved. In the GlobalSurg-Pulmonary score work, with a <0.1% overall missingness, we did not consider it necessary, with clinical and simulation studies demonstrating little impact of multiple imputation methods where there are very low rates of missing data (Jakobsen et al. 2017) and high sample sizes such as ours. Finally, more models and pre-processing steps could have been implemented, including using methods for managing class imbalance (i.e., over/under sampling or Synthetic Minority Over- Sampling Technique (SMOTE)). However, these are criticised in clinical datasets due to risk of miscalibration (strong overestimation of probability to belong to the minority class), and we opted not to adopt these here (van den Goorbergh et al. 2022).

As final comment, this work was particularly rewarding due to the close collaboration with the CovidSurg and GlobalSurg teams, surgeons with whom we co-developed both models and tailored the tools to meet their specific needs. Key insights include the necessity for transparency concerning the system's behaviour and limitations, both in terms of statistical performance and recognizing situations where the system's predictions are inapplicable.

Moreover, emphasizing model interpretability and deployment was crucial, as decisions to postpone or continue surgery are complex, multifactorial, and require in-depth shared decision-making between the patient and the perioperative multidisciplinary team. Consequently, we have not pre-specified how the scores should be used in decision-making but presented a range of 'cut-off' point scores for 'ruling in' or 'ruling out' risk that clinicians can adopt during implementation.Recent studies underscore the importance of such close collaborations, with one study illustrating a failed ML model in a clinical setting where providing mortality risk information did not influence clinical decision-making or improve patient outcomes with heart failure (Ahmad et al. 2022) due to a lack of adoption. Ultimately, ML models do not exist in isolation and must be seamlessly integrated into a workflow designed to drive behaviour change.

4 Chapter 4: Complex systems and ageing trajectories

Content from this chapter will appear:

Bravo-Merodio, L, Acharjee, A, Gkoutos, GV, & Lord, JM. (n.d.-). Longitudinal analysis of biological age (PhenoAge) in UKBiobank participants reveals key factors driving ageing trajectories. In *Nature Ageing*.

4.1 Background and chapter overview

In the previous chapters we have explored the use of computational algorithms to find underlying patterns in biological and medical datasets. These datasets have been small, thoroughly curated and related to mostly a single type of data with a clear hypothesis to model. Here, we would like to extend our analysis to another side of computational biology, one involving the study of complex biological systems and requiring the integration of multiple data types. More specifically, taking advantage of longitudinal, deeply phenotyped data resources such as UKBiobank (UKBB), we will investigate ageing (A. A. Cohen et al. 2022) and health to disease transition as the epitome of complex systems. To do so, we will first build different ageing phenotypes based on their health trajectories and then proceed to its characterization through health outcomes, disease patterns and genetic makeup.

Age is the greatest single risk factor for many diseases such as cancer, diabetes, cardiovascular disease, sarcopenia or dementia (Niccoli and Partridge, 2012), non-communicable, chronic diseases responsible for more than 70% of deaths and disability worldwide (Allen et al. 2023). Recently, the field of geroscience has accelerated rapidly, with important discoveries in the identification of biological underpinnings for the age-dependency of most chronic diseases (Sierra 2016) pushing the field forwards. Such association have been possible thanks to being able to define the genetic make-up of diseases through genome-wide association studies (GWAS) (Burton et al., 2007). As mentioned in the introduction, these studies have revealed that complex traits are driven by high polygenicity, where a high number (e.g. around 10,000 for variants for human height (Yengo et al., 2022)) of small effect common variants (Minor Allele Frequency (MAF) > 1%) account for most of the heritability (Boyle, Li, and Pritchard., 2017; Uffelmann et al., 2021). A key element, is that these common genetic variants tend to be non-coding, outside protein-coding exons (~90%) (Albert and Kruglyak, 2015; B. Li and Ritchie, 2021) and mostly located in genomic regulatory

regions within relevant cell types (Farh et al., 2015). As such, many common variants are highly pleiotropic in nature, with shared loci in different diseases and traits revealing shared mechanisms and pathways. This is the case for age-dependent chronic diseases, where recent studies have provided proof on common underlying genetic mechanisms (Dönertaş et al., 2021; Baghdadi et al., 2022). These genetic links, in turn, have then also been found (Fraser et al. 2022).to be associated with the “Hallmarks of Ageing” pathways (López-Otín et al. 2013; 2022), which represent the key molecular, cellular, and systemic drivers of the ageing process and now too, possible representations of common pathogenic drivers of chronic diseases (Kennedy et al., 2014).

One key hallmark known to be highly associated with the development of chronic and age-related diseases is inflammation, with the increase in higher basal inflammatory status characteristic of old age known as “inflammaging” (Franceschi et al. 2000). Recent research points at most diseases having an inflammatory component in their aetiology (Furman et al. 2019), with those properly classified as chronic inflammatory diseases (CIDs) (Schultze and Rosenstiel, 2018) found to have a significant overlap in their pathophysiology via common aberrant inflammatory signal transduction pathways as exemplified by common genetics (Ellinghaus et al. 2016). This pervasive presence of inflammation in diseases is of no surprise, with the inflammatory response being an integral part of animal biology and most biological processes requiring of inflammatory control (Kotas and Medzhitov 2015; Medzhitov 2021). Also, other hallmarks such as altered nutrient sensing, mitochondrial dysfunction, cell senescence and stem cell exhaustion have also been notably associated with disease development, with most of them being highly interrelated (Kirkwood 2005; Kennedy et al. 2014b; López-Otín et al. 2013; 2022).

As a means to better characterize the heterogeneous patterns of existing ageing phenotypes (Alpert et al. 2019; C.-L. Kuo et al. 2021), defining the sequences of disease developed (i.e

multimorbidity signature) has been postulated. In old age, higher incidence of multimorbidity (Fabbri et al. 2015) goes hand in hand with certain disease patterns seen to occur more frequently than expected by chance (eg. cardiovascular/metabolic clusters, neuropsychiatric or depression/anxiety clusters (Chowdhury et al. 2023)), pointing at possible common disrupted pathways. To ascertain these links between aging and disease development, longitudinal studies are required (Jensen et al. 2014), with promising examples already unveiling possible mechanistic insights (Alpert et al. 2019; P.-L. Kuo et al. 2022; X. Li et al. 2020).

4.2 Summary

In this study, we explored the concept of biological age as a definition of health status, with deviations from chronological age representing age-accelerated or age-decelerated phenotypes (Rutledge, Oh, and Wyss-Coray 2022). Previous estimates of biological age (also termed ageing clocks) have been created through modelling of biomarker data with longitudinal mortality or morbidity outcomes. These clocks derive from a range of biomarkers, including: routine clinical blood parameters as used in the PhenoAge (Levine et al. 2018) and GrimAge (Lu et al. 2019) clocks; immune parameters in iAge (Sayed et al. 2021) and Imm-Age (Alpert et al. 2019); DNA methylation for the epigenetic clocks (Horvath 2013; Hannum et al. 2013); and various omics based clocks using metabolite (van den Akker et al. 2020), proteomic (Tanaka et al. 2018; Lehallier et al. 2019) or transcriptomic (Holzscheck et al. 2021) data.

In our study, we derived biological age using the PhenoAge algorithm (Levine et al. 2018) from blood marker data. This was done by using UKBiobank, a large-scale biomedical database with in-depth genetic, lifestyle and health information from approximately 500,000 volunteers in the UK. In this work, we were able to extract longitudinal information by merging

participant's clinical histories data through GP records (Bycroft et al. 2018). If the predicted biological age (PhenoAge) was lower than chronological age we deemed these individuals healthy, and if higher, unhealthy. Measurements in time (average follow-up of 10 years) were transformed into linear trajectories and grouped into 4 similar clusters: healthy remaining healthy (HRH), healthy becoming unhealthy (HBU), unhealthy becoming healthy (UBH) and finally unhealthy remaining unhealthy (URU). Considering these different trajectories as representing participants engaging with the ageing process at different rates (Kubben and Misteli, 2017), or through different mechanism (C.-L. Kuo et al. 2021), we proceeded to determine if they could be characterised through the study of multimorbidity patterns and genetic makeup through GWAS. To the best of our knowledge, no previous research on both genetic background and longitudinal measurements has been done in the study of biological ageing metrics, with previous research focusing on either longitudinal assessment with no mechanistic input (P.-L. Kuo et al. 2022) or molecular associations but at a single point measurement (C.-L. Kuo et al. 2021).

Our analysis revealed that male sex, high BMI, comorbidities, smoking, and shorter parental lifespan dominated the unhealthy categories (URU). URU was also enriched for cardiometabolic diseases with diabetes, obesity, or heart failure being highly prevalent, In contrast, HRH had less prevalence of most diseases, was female dominated and had the best health outcomes overall. Finally, the genetic makeup of the different trajectories revealed a marked immune signature for HRH participants and a metabolic signature driven by lipid biosynthesis for URU. After enrichment analysis, strong associations to age-related genes were found for both HRH and URU, with pro-longevity associated variants (e.g rs429358, rs4420638, rs2075650) dominating HRH and SREBF1 and SREBF2 transcription factors postulated as key regulatory element driving the URU ageing phenotype. No significant genetic associations were found for UBH and HBU which we postulate to be more

highly influenced by lifestyle associated factors. An overview of the study methodology can be found in Figure 4-1.

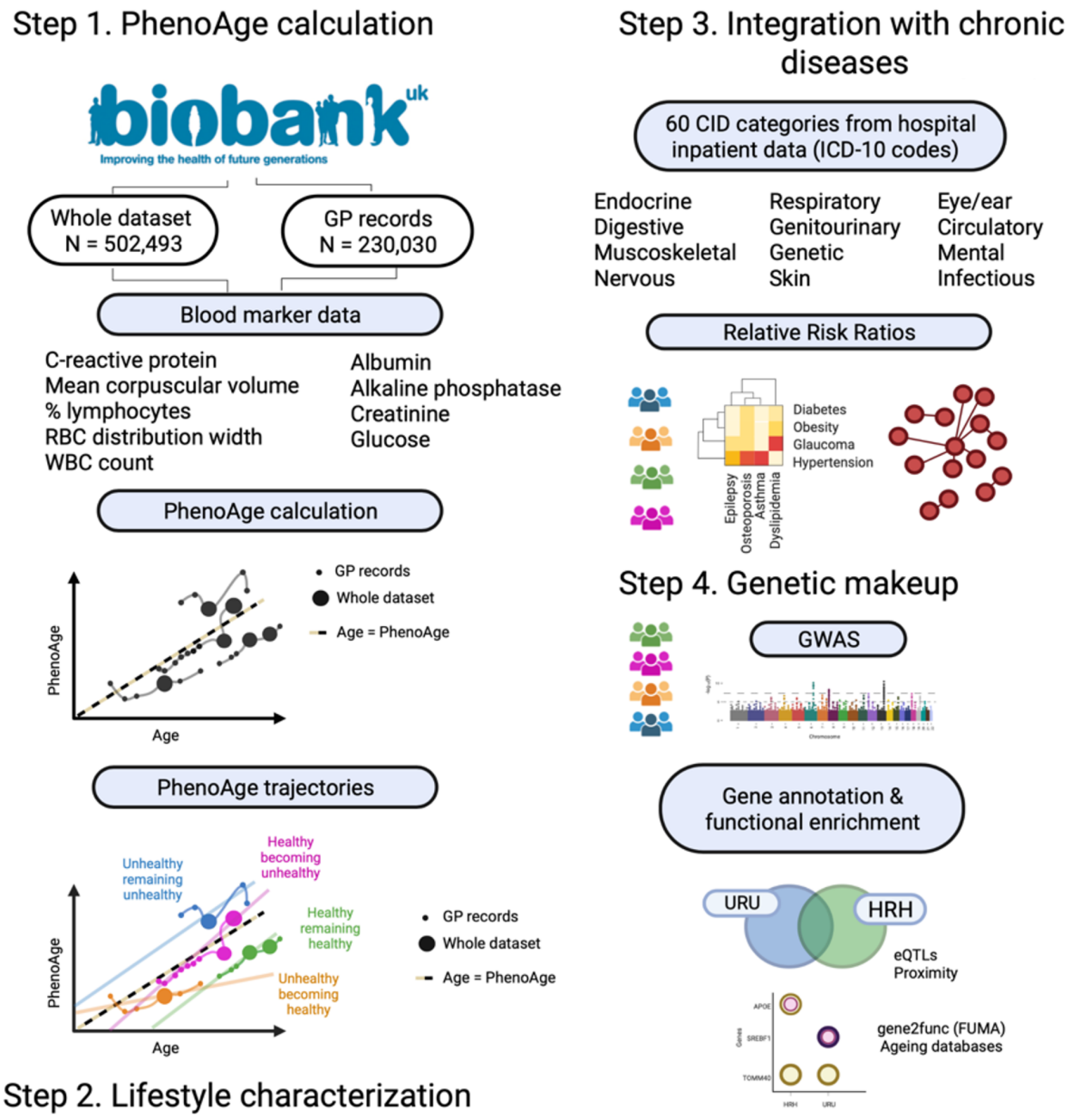


Figure 4-1 Study overview and methodology with key steps highlighted.

4.3 Methods

4.3.1 UK Biobank data

The UKBB is a prospective cohort study of approximately 502,000 volunteers of 40–69 years of age at recruitment. Participants were recruited in the UK between 2006 and 2010 and have been followed up through time. Participant data include health records that are periodically updated, self-reported survey information, linkage to death and cancer registries, collection of urine and blood biomarkers, imaging data, accelerometer data and various other phenotypic end points. (Bycroft et al. 2018). In our study we have used both disease diagnosis information and biomarker data. Biomarker data was collected through two different avenues, the own UKBB's assessments and through GP records supplied as part of each participant´s clinical history. More specifically, we have biomarker information of 502,493 participants at their initial visit (baseline measurement) with a few going on to follow up measurements. Regarding GP records, they were downloaded from UKBB repository on April 2021 and all read codes (Kuan Po Ai 2020) were mapped following UKBB 592 Resource (https://biobank.ndph.ox.ac.uk/ukb/refer.cgi?id=592). Read codes for the 9 clinical chemistry biomarkers from the PhenoAge linear model (albumin, creatinine, glucose, c-reactive protein, lymphocyte percent, mean cell volume, red cell distribution width, alkaline phosphatase and white blood cell count) were extracted (Levine et al. 2018) and are reported in Appendix 2 **.**Lifestyle factors such as BMI or smoking and exercise habits were also extracted in the same way. Regarding disease diagnosis, 60 chronic categories were produced following Calderón-Larrañaga´s exhaustive classification of the International Classification of Diseases, tenth revision (ICD10) codes (Appendix 3). In their research a disease or condition was considered to be chronic if "it had a prolonged duration and either (a) left residual disability or worsening quality of life or (b) required a long period of care, treatment, or rehabilitation" (Calderón-Larrañaga et al. 2017). Participant´s ICD-10 codes were obtained from the Hospital Episode Statistics (HES) recorded in category 2002 for main and secondary diagnosis, alongside date of diagnosis.

We also used genotype datasets from the UK Biobank (release version 2 for the directly genotyped variants and the imputed HLA allelotype datasets and release version 3 for the imputed genotype dataset) for all analyses in the study. We also restricted our analyses to unrelated individuals and followed quality control criteria provided by UK Biobank.

The UKBB has obtained Research Tissue Bank approval from its ethics committee that covers the majority of proposed uses of the resource with the Research Ethics Committee reference for UKBB being 16/NW/0274. All necessary patient/participant consent was obtained and appropriate institutional forms have been archived. We accessed data through application number 31224.

4.3.2 PhenoAge calculation

The equation to calculate PhenoAge was based on the parametrization of two Gompertz proportional hazard models, one fit using the 10 selected variables from a Cox proportional hazard elastic net model for mortality and the other fit using only chronological age. What it shows is that an individual with 30 chronological years, if predicted a PhenoAge of 45 years, has an average morbidity and/or mortality risk of someone who is 45 years old chronologically instead of 30 (Levine et al. 2018; Liu et al. 2018) The equation is:

$$PhenoAge = 141.50225 + \frac{\ln ln\,(0.00553 * \ln ln\,(1 - xb))}{0.090165}$$

$$xb = -19.907 - 0.0336 * albumin + 0.0095 * creatinine + 0.0195 * glucose + 0.0954 * \ln ln\,(C - reactive\ portein) - 0.0120 * lymphocyte\ percent + 0.0268 * mean\ cell\ volumne + 0.3356 * red\ blood\ cell\ distribution\ width + 0.00188 * alkaline\ phosphatase + 0.0554 * white\ blood\ cell\ count + 0.0804 * age$$

Equation 7

From all biomarkers obtained through UKBB assessment (first or second visit), we obtained a PhenoAge measurement given all these 9 markers are measured each time. But, in GP records, we do not always get a reading of all 9 markers in the same date and so a PhenoAge measurement is created every time one of the 9 markers has a new reading, with the missing values for that date keeping the closest record for it in time (moving window imputation shown in Appendix 4). Participants with missing blood marker data or just one recorded date entry are deleted.

To emulate a patients phenotypic trajectory in time, a linear regression model (phenoage = m * age + c) is fitted through all calculated PhenoAge measurements and their chronological age. This has been compared to the PhenoAgeAccell measure derived by Liu et al. 2018, as a means to extract an actionable parameter that allows for the participants healthspan classification. In our case, these linear trajectories are then grouped into 4 categories (Healthy remaining Healthy (HRH); Healthy becoming Unhealthy (HBU); Unhealthy becoming Healthy (UBH); Unhealthy remaining Unhealthy (URU)) depending on their relationship with the line chronological age = PhenoAge. More specifically, given a 2-dimensional equation of. a line ($y = mx + c$) depending on their intercept (c) and slope coefficient (m) combinations we have:

$$c > 0 \; and \; m \geq 1 : URU$$
$$c > 0 \; and \; m < 1 : UBH$$
$$c < 0 \; and \; m > 1 : HBU$$
$$c < 0 \; and \; m <= 1 \; and \; m > 0 : HRH$$

Equation 8

4.3.3 Disease studies: prevalence, co-occurrence calculations and correlation

After extracting disease information as determined by (Calderón-Larrañaga et al. 2017), shown in Appendix 3, disease prevalence was studied for the whole population as well as for each separate category using: existing cases/whole population*100 and including the BinomCI function and clopper-pearson person method for confidence interval calculation. We next determined whether there were distinct disease patterns or comorbidities (two or more diseases in the same individual), found in the different ageing trajectories by calculating disease co-occurrence (relative risk) in pairwise associations. Relative risk (RR) was determined following (Dönertaş et al. 2021), as an estimate of having disease A when already affected by disease B. It is calculated as a fraction between the number of patients diagnosed with both diseases and a random expectation based on disease prevalence. Mathematically it can be expressed as:

$$RR = \frac{\frac{N\ disease\ a\ and\ b}{N\ disease\ a}}{\frac{N\ without\ disease\ a\ but\ disease\ b}{N\ without\ disease\ a}}$$

Equation 9

To better understand subgroups between diseases, main ICD-10 code categories were mapped to each of the 60 categories as seen in Appendix 5

4.3.4 Genome Wide Association Studies (GWAS)

Next, to better understand the four different ageing categories and the associated disease prevalence combinations, we sought to integrate the genetic data available in UKBB. Genotype data from UKBB (GECh37 (hg19) coordinates) was extracted on the 12/01/21 and included 804,470 variants of 488,377 participants. After file pre-processing, genotype information can be found as merged files in PLINK (bed, bim, fam). Quality control (QC) was done using PLINK (Purcell et al. 2007) with the following filters (samples and variants of

missing call rates exceeding 10%, a Hardy-Weinberg equilibrium exact test p-value below 1^10-15, and a minor allele frequency of 0.01) 522,963 variants remained and 6 participants were removed (488,371 samples). REGENIE (Mbatchou et al. 2021), a computational efficient machine learning method that first reduces the genotype space into local blocks (of 1000 variants) using ridge regression and then performs association testing, was used to perform GWAS. Four parallel GWAS analysis were performed on participants belonging to each of the four defined categories. A covariate file used as an input for PLINK included the following fields: sex, age (maximum age of recorded disease), batch and first 10 PCs. Finally, the phenotype file included dummified information on which category each participant belonged to: URU, HRH, UBH, HBU.

4.3.5 Gene annotation

Considering the literature-standard *P* value $<5 \times 10^{-8}$ as significant (Panagiotou, Ioannidis, and Genome-Wide Significance Project, 2012), summary statistics of significant variants can be found in Appendix 6**.** Continuing the downstream analysis, we then proceeded to annotate them. Linkage disequilibrium (LD) was not taken into account to expand SNPs (Dong et al. 2021) or assess highly correlated SNPs (fine-mapping) for prioritization, with only proximity and colocalization with functional elements ((expression quantitative trait loci) cis-eQTL) used.

4.3.5.1 Proximity analysis

Based on the analysis performed in (Dönertaş et al. 2021), we used VariantAnnotation (Obenchain et al. 2014), TxDb.Hsapiens.UCSC.hg19.knownGene and GenomicRanges (Lawrence et al. 2013) packages in R, to map the genomic coordinates for each SNP to

genes. Specifically, if a gene is within the coding region, intron, 5' or 3' UTR, or 1kb down- or up-stream of the transcription start site, we annotated that SNP to the gene. As a result, we had 6713 SNP-gene associations for the 1680 significant variants found in the 4 GWAS analysis. We used the Ensembl biomaRt (Durinck et al. 2009) package in R to retrieve HGNC symbols, Ensembl Gene IDs and gene descriptions for the Entrez gene IDs obtained from TxDb.Hsapiens.UCSC.hg19.knownGene database.

4.3.5.2 eQTLs

Biological functional inference linking possible regulatory elements was studied using expression quantitative trait loci eQTL information for all tissues in GTEx v8 eQTL (Consortium 2015; Lonsdale et al. 2013; Gamazon et al. 2018) through the Phenoscanner (Kamat et al. 2019) database ((adjusted *P* value<0.05, FDR corrected)). eQTLs, where a variant influences mRNA levels, have been shown to be 5-40% of functional variants (Connally et al. 2021)) and integrating this intermediate regulatory information is key to connect overlapping gene modules associating disease mechanisms together (Hormozdiari et al. 2018). This final GTEx data release (v8) includes 54 primary human tissues, with 49 including at least 70 samples with both whole genome sequencing and tissue specific RNA-seq.

4.3.6 Functional enrichment

Given our major motivation was to use these identified gene associations to help infer any biological underpinning for the four ageing phenotypes, genes were then functionally enriched using the gene2func tool in FUMA (Watanabe et al. 2017) where annotated genes are tested for over-representation in publicly available gene sets (MsigDB v7.0) with

information from: BioCarta, KEGG, Pathway Interaction Database (PID), Reactome, WikiPathways and GO biological processes (Subramanian et al. 2005).

4.3.7 Association with ageing

Genes were also compared to literature-based ageing databases (Dönertaş et al. 2021). These databases are GenAge Human (genes associated with lifespan in humans or closely related species), human orthologs in GenAge Model Organism (genes modulating lifespan in model organisms) and CellAge (genes regulating cellular senescence) (Tacutu et al. 2018; Barardo et al. 2017). Longevity associated genes were explored too, based on work from (Zenin et al. 2019), with SNPs previously reported to be associated with lifespan, (extreme) longevity and disease-free survival included. We also incorporated the overlapping immune and age associated genes found as conserved orthologs in (Fabian et al. 2021) research.

4.4 Results

4.4.1 UKBB statistics and PhenoAge calculation.

To calculate PhenoAge from UKBB participants we accessed the data for nine blood markers: albumin; alkaline phosphatase; creatinine; glucose; C reactive protein; mean corpuscular volume; percentage lymphocytes; red blood cell distribution width; white blood cell count (Levine et al., 2018). In order to obtain as many longitudinal readings of PhenoAge as possible, we exploited two different resources, GP records and the UKBB baseline and follow-up visit measurements (Bycroft et al. 2018). In total, UKBB holds information from 502,493 participants with 61,061 of them having at least a second visit data, rising to 230,030 counting linkage to GP record data. In total 494,679 participants in UKBB and 206,218 participants with GP records had at least a data entry on one of the markers selected with more details described in Table 4-1. To unify both datasets, missing values per

entry date where imputed through a moving window where missing blood marker values were updated from the closest recorded value. Extra filtering steps were taken as described in methods.

Blood Markers	Maximum per participant	Mean +/- sd per participant	Participants with at least one value
Albumin	250	7,04 ± 8,15	188995
Alkaline Phosphatase	254	6,87 ± 8,06	192615
Creatinine	288	8,32 ± 9,15	199542
Glucose	83	3,75 ± 3,59	160441
Mean corpuscular volume (MCV)	354	6,44 ± 8,37	191536
Percentage lymphocytes	217	5,08 ± 7,18	26748
Plasma C-reactive protein level	110	2,1 ± 3,08	19103
Red blood cell dist width	353	5,51 ± 7,21	113970
White blood count	258	6,67 ± 9,67	53699

Blood Markers	Total number of readings
Albumin	1330964
Alkaline Phosphatase	1322920
Creatinine	1661000
Glucose	601562
Mean corpuscular volume (MCV)	1234337
Percentage lymphocytes	135971
Plasma C-reactive protein level	40029
Red blood cell dist width	628220
White blood count	358153

Number of readings	Number of participants
0-5	15022
5-10	26016
10-15	23213
15-20	20097
20-30	32158
30-40	24160
40-50	18123
50-60	13494
60-80	16713
80-100	8041
100-200	7745
200-300	789
300-500	448
500-1000	182
1000-1532	17

Table 4-1. GP records data on the 9 markers extracted to calculate PhenoAge
Top) Number of participants with at least one measurement on each of the blood markers including maximum, mean and standard deviation (sd) of measurements reported per

participant. Middle) Total number of readings of each blood marker in GP records Bottom) Binned number of participants with the number of GP records data of any of the blood markers studied

In the end, the PhenoAge value was calculated using the equation in Levine et al., 2018 (Equation 7) for 2,261,614 date entries of 199,228 participants. From these, a mean of 11.34 (maximum of 356) data points in time per patient, covering an average of 9.6 years (maximum 60.15 years) was used for the analysis. As the objective is to extract actionable ageing patterns, these PhenoAge values per participant were defined into trajectories fitting a simple linear model (Methods). An example of the different trajectories for PhenoAge scores against chronological age for 5 participants is shown in Figure 4-2. According to their intercept and slope coefficient, 4 different categories were defined (Equation 8): Healthy remaining Healthy (HRH; n= 20,395); Healthy becoming Unhealthy (HBU: n = 75,122); Unhealthy becoming Healthy (UBH: n = 94,130); Unhealthy remaining Unhealthy (URU; n = 9569) (Figure 4-2 B). 12 participants with no fitted line were excluded.

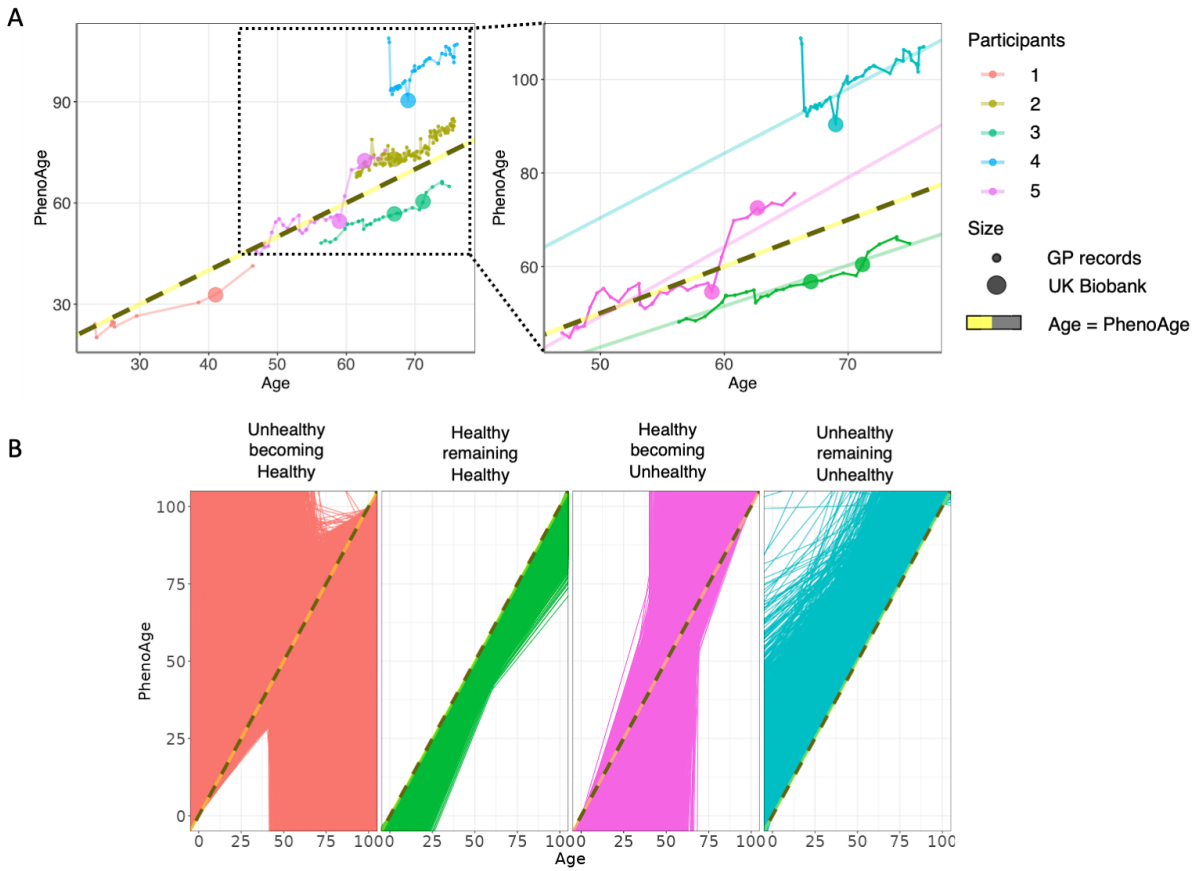


Figure 4-2. PhenoAge measurements, trajectories and categories

A) Data preparation. PhenoAge measurements of 5 participants.Data points calculated from original UKBB data are shown in large circles and those coming from GP records, in smaller ones. Zoomed in (right panel), we show the linear regression line of the PhenoAge trajectories of three participants. B) Trajectories were then grouped into four categories based on Equation 8

4.4.2 Characterisation of ageing trajectory categories using lifestyle factors and health outcomes.

We next assessed key outcomes such as mortality (Data Field (DF): 40000), together with lifestyle indicators (e.g townsend deprivation index (DF: 189), overall health rating (DF: 2178), BMI (DF: 21001), and smoking status (DF: 20116), physical activity (Metabolic Equivalent Task (MET) data based on IPAQ (International Physical Activity Questionnaire (DF: 22040 and 22032), or other potential healthspan modifiers such as parental longevity (DF (1807,3526, 2946, 1845)), (Cassidy et al. 2016). Demographic descriptors such as age, sex (DF: 22001) or comorbidities were also included with results shown in Table 4-2.

Comparing the HRH with URU trajectories the former are female dominated, have a lower number of comorbidities (2.8 vs 5.6), better overall health rating, lower BMI (26 vs 30), more never-smokers (60.4% vs 44.9%) and lower deprivation score, all factors we would expect to influence the rate of biological ageing. This was confirmed by lower mortality in the healthy group (2.6% vs 11.8%). Interestingly both the mother's and father's lifespan were also higher in those with a lower PhenoAge, i.e. on a healthy trajectory. Comparing those with an altered trajectory across our study, those who changed from a low PhenoAge to a higher PhenoAge (healthy becoming unhealthy, HBU) with those who changed from unhealthy to a healthy trajectory, we can see higher number of comorbidities, more deprivation, higher BMI and slightly shorter parental lifespan, resulting in higher mortality (8.1% vs 4.4%) in the HBU group.

Lifestyle factors		Healthy remaining healthy	Unhealthy becoming healthy	Healthy becoming unhealthy	Unhealthy remaining unhealthy
Total N (%)		20395 (10.2)	94130 (47.3)	75122 (37.7)	9569 (4.8)
Mortality (%)		537 (2.6)	4163 (4.4)	6096 (8.1)	1127 (11.8)
Age		66.4 (8.1)	66.1 (8.3)	66.2 (8.4)	67.4 (8.3)
Comorbidities		2.8 (2.7)	3.4 (3.3)	4.0 (3.9)	5.6 (4.6)
IPAQ activity group (MET)* (%)	low	2636 (12.9)	14289 (15.2)	11939 (15.9)	1802 (18.8)
	moderate	6989 (34.3)	30797 (32.7)	24356 (32.4)	2911 (30.4)
	high	6975 (34.2)	30962 (32.9)	24175 (32.2)	2677 (28.0)
	(Missing)	3795 (18.6)	18082 (19.2)	14652 (19.5)	2179 (22.8)
Minutes of activity/week (MET)*		2732.7 (2664.6)	2673.8 (2714.5)	2651.4 (2748.0)	2503.0 (2789.6)
Last Recorded Age Father		71.2 (16.1)	70.4 (16.7)	70.0 (16.9)	68.8 (18.0)
Last Recorded Age Mother		75.7 (13.8)	74.8 (14.8)	74.3 (15.2)	73.4 (16.8)
Sex (%)	Female	13497 (66.2)	51751 (55.0)	37890 (50.4)	3853 (40.3)
	Male	6721 (33.0)	41627 (44.2)	36672 (48.8)	5625 (58.8)
	(Missing)	177 (0.9)	752 (0.8)	560 (0.7)	91 (1.0)
Smoking (%)	Previous	6879 (33.7)	32550 (34.6)	26441 (35.2)	3573 (37.3)
	Current	1123 (5.5)	9421 (10.0)	8144 (10.8)	1628 (17.0)
	Never	12322 (60.4)	51776 (55.0)	40196 (53.5)	4296 (44.9)
	PNA	63 (0.3)	320 (0.3)	287 (0.4)	62 (0.6)
	(Missing)	8 (0.0)	63 (0.1)	54 (0.1)	10 (0.1)
Townsend deprivation index		-1.8 (2.8)	-1.4 (3.0)	-1.3 (3.1)	-0.8 (3.3)
Overall health rating (%)	Excellent	4148 (20.3)	14658 (15.6)	11454 (15.2)	716 (7.5)
	Good	12641 (62.0)	55263 (58.7)	42840 (57.0)	4416 (46.1)
	Fair	3125 (15.3)	19839 (21.1)	16487 (21.9)	3219 (33.6)
	Poor	404 (2.0)	3908 (4.2)	3903 (5.2)	1132 (11.8)
	DNK	57 (0.3)	348 (0.4)	342 (0.5)	62 (0.6)
	PNA	12 (0.1)	51 (0.1)	41 (0.1)	14 (0.1)
	(Missing)	8 (0.0)	63 (0.1)	55 (0.1)	10 (0.1)
BMI		26.0 (3.9)	27.5 (4.7)	27.8 (4.8)	30.0 (5.9)

Table 4-2. Lifestyle and health outcomes data for the different ageing categories.
Continous values are presented as mean (standard deviation). Discrete values as categories with number (& of total category).*Physical activity quantified using MET (Metabolic Equivalent Task) scores based on IPAQ (International Physical Activity Questionnaire) guidelines (Cassidy et al. 2016) Townsend deprivation index: Based on the preceding national census output areas. Each participant is assigned a score corresponding to the output area in which their postcode is located.PNA: Prefer not to answer. DNK: Do not know

4.4.3 Characterisation of ageing trajectory categories using disease patterns

To further define our four ageing categories, disease information was obtained from UKBB from hospital inpatient records. Date of diagnosis was obtained, with diagnosis classified through ICD10 codes. We focused on chronic conditions most frequently occurring in multimorbidity, as these are most likely to represent downstream impacts of the ageing process, following the categorization defined by Calderon-Larrañga and colleagues (Calderón-Larrañaga et al. 2017). In their categorisation, from 19,000 ICD-10 codes, 918 were selected as defining chronic conditions and subsequently mapped onto 60 chronic disease categories. Only participants with ICD-10 codes belonging to one of these categories were selected (n = 146,845), leaving the total of participants in each category as: HRH (n= 13,838); HBU (n = 56,354); UBH (n = 68,450); URU (n = 8,203). Disease category prevalence information is summarized in Table 4-3. The biggest differences in prevalence between disease categories (2.5 times) is highlighted, with obesity and diabetes showing the highest differences with baseline values (10.85% versus 2.64% and 26.58%) or (8.25% versus 4.06% and 14.99%) respectively. Highlighted diseases: anaemia, diabetes, obesity, COPD, heart failure, chronic kidney disease (CKD), cardiovascular diseases, sleep disorders, other respiratory disease, peripheral vascular disease and bradycardias, can be mainly grouped as cardiometabolic disorders, a combination often seen associated to males with higher severity and morbidity associations in multimorbidity clustering studies (Violan et al. 2014; Prados-Torres et al. 2014; Vetrano et al. 2020).

Chronic disease categories	All		Healthy remaining healthy		Healthy becoming Unhealthy		Healthy becoming unhealthy		Unhealthy remaining unhealthy	
	N 146,845	Prev (%)	N = 13, 383	Prev (%)	N = 56,354	Prev (%)	N = 68,450	Prev (%)	N = 8,203	Prev (%)
Hypertension	57979	38.66	4240	29.800	26257	37.538	23023	40.031	4459	53.81
Colitis and related diseases	39956	26.64	3553	24.972	17805	25.455	16076	27.952	2522	30.43
Esophagus, stomach and duodenum diseases	36901	24.60	3118	21.915	16760	23.961	14630	25.438	2393	28.88
Osteoarthritis and other degenerative joint diseases	35393	23.60	3045	21.401	16271	23.262	13779	23.958	2298	27.73
Solid neoplasms	34630	23.09	3137	22.048	15255	21.809	14137	24.581	2101	25.35
Other musculoskeletal and joint diseases	29639	19.76	2565	18.028	13636	19.494	11546	20.075	1892	22.83
Dyslipidemia	28929	19.29	2136	15.013	13030	18.628	11565	20.108	2198	26.52
Other genitourinary diseases	27003	18.00	2626	18.457	12290	17.570	10611	18.450	1476	17.81
Ischemic heart disease	22956	15.31	1473	10.353	9812	14.028	9652	16.782	2019	24.36
Cataract and other lens diseases	22120	14.75	2059	14.471	10101	14.441	8429	14.656	1531	18.47
Inflammatory arthropathies	18965	12.65	1326	9.320	8074	11.543	7937	13.800	1628	19.65
Asthma	18053	12.04	1308	9.193	8192	11.712	7280	12.658	1273	15.36
Anemia	**17197**	**11.47**	**994**	**6.986**	**7058**	**10.090**	**7535**	**13.101**	**1610**	**19.43**
Diabetes	**16279**	**10.85**	**376**	**2.643**	**6084**	**8.698**	**7616**	**13.242**	**2203**	**26.58**
Dorsopathies	14453	9.64	1207	8.483	6348	9.075	5930	10.311	968	11.68
Atrial fibrillation	13568	9.05	857	6.023	5408	7.731	6118	10.638	1185	14.30
Peripheral neuropathy	13102	8.74	1102	7.745	5984	8.555	5096	8.861	920	11.10
Thyroid diseases	12690	8.46	1163	8.174	5820	8.320	4883	8.490	824	9.94
Obesity	**12369**	**8.25**	**578**	**4.062**	**5042**	**7.208**	**5507**	**9.575**	**1242**	**14.99**
Depression and mood diseases	12123	8.08	863	6.066	5299	7.576	5063	8.803	898	10.84
Other digestive diseases	12108	8.07	958	6.733	5301	7.578	4974	8.648	875	10.56
Other eye diseases	10814	7.21	959	6.740	4841	6.921	4210	7.320	804	9.70
Chronic pancreas, biliary tract and gallbladder diseases	10644	7.10	830	5.834	4732	6.765	4301	7.478	781	9.42
Prostate diseases	10205	6.80	678	4.765	4397	6.286	4321	7.513	809	9.76
Copd, emphysema, chronic bronchitis	**9649**	**6.43**	**424**	**2.980**	**3913**	**5.594**	**4279**	**7.440**	**1033**	**12.47**
Cerebrovascular disease	8936	5.96	602	4.231	3783	5.408	3743	6.508	808	9.75
Heart failure	**8759**	**5.84**	**426**	**2.994**	**3174**	**4.538**	**4233**	**7.360**	**926**	**11.17**
Neurotic, stress-related and somatoform diseases	8565	5.71	720	5.060	3799	5.431	3495	6.077	551	6.65
Venous and lymphatic diseases	8032	5.36	835	5.869	3667	5.242	3097	5.385	433	5.23
Ear, nose, throat diseases	7822	5.22	654	4.597	3550	5.075	3170	5.512	448	5.41
Chronic kidney diseases	**7710**	**5.14**	**230**	**1.617**	**2495**	**3.567**	**3949**	**6.866**	**1036**	**12.50**
Other cardiovascular diseases	**7491**	**4.99**	**443**	**3.114**	**3024**	**4.323**	**3319**	**5.771**	**705**	**8.51**

Cardiac valve diseases	7168	4.78	473	3.324	2874	4.109	3240	5.634	581	7.01
Osteoporosis	6785	4.52	728	5.117	3056	4.369	2595	4.512	406	4.90
Other neurological diseases	6088	4.06	433	3.043	2512	3.591	2662	4.629	481	5.80
Other metabolic diseases	5866	3.91	351	2.467	2349	3.358	2618	4.552	548	6.61
Autoimmune diseases	5015	3.34	335	2.355	2098	2.999	2155	3.747	427	5.15
Deafness, hearing impairment	4749	3.17	376	2.643	2033	2.906	1970	3.425	370	4.46
Sleep disorders	**4364**	**2.91**	**224**	**1.574**	**1827**	**2.612**	**1874**	**3.258**	**439**	**5.30**
Other respiratory diseases	**4163**	**2.78**	**174**	**1.223**	**1627**	**2.326**	**1949**	**3.389**	**413**	**4.98**
Glaucoma	4021	2.68	388	2.727	1844	2.636	1519	2.641	270	3.26
Migraine and facial pain syndromes	3457	2.31	328	2.305	1593	2.277	1373	2.387	163	1.97
Peripheral vascular disease	**3446**	**2.30**	**185**	**1.300**	**1373**	**1.963**	**1524**	**2.650**	**364**	**4.39**
Blood and blood forming organ diseases	3419	2.28	224	1.574	1320	1.887	1551	2.697	324	3.91
Bradycardias and conduction diseases	**3216**	**2.14**	**195**	**1.371**	**1207**	**1.726**	**1523**	**2.648**	**291**	**3.51**
Inflammatory bowel diseases	2992	1.99	184	1.293	1265	1.808	1234	2.146	309	3.73
Other psychiatric and behavioral diseases	2925	1.95	137	963	1191	1.703	1297	2.255	300	3.62
Dementia	2853	1.90	220	1.546	1117	1.597	1265	2.200	251	3.03
Hematological neoplasms	2791	1.86	172	1.209	1029	1.471	1362	2.368	228	2.75
Chronic ulcer of the skin	2630	1.75	78	548	956	1.367	1262	2.194	334	4.03
Allergy	2456	1.64	218	1.532	1164	1.664	958	1.666	116	1.40
Epilepsy	2430	1.62	178	1.251	1095	1.565	978	1.700	179	2.16
Chronic liver diseases	1911	1.27	64	0.450	704	1.006	909	1.581	234	2.82
Parkinson and parkinsonism	1368	0.91	120	0.843	576	0.823	566	0.984	106	1.28
Other skin diseases	1356	0.90	118	0.829	606	0.866	520	0.904	112	1.35
Blindness, visual impairment	986	0.66	67	0.471	395	0.565	427	0.742	97	1.17
Multiple sclerosis	846	0.56	82	0.576	396	0.566	329	0.572	39	0.47
Schizophrenia and delusional diseases	534	0.36	31	0.218	219	0.313	238	0.414	46	0.56
Chronic infectious diseases	380	0.25	18	0.127	156	0.223	168	0.292	38	0.46
Chromosomal abnormalities	71	0.05	3	0.21	26	0.37	37	0.64	5	0.06

Table 4-3. Disease prevalence of 60 chronic disease categories by ageing trajectory groups. Comparison of chronic condition prevalence between the whole cohort of participants with at least one diagnosis (n = 146845) and each ageing phenotype. In bold we have those diagnosis with the highest (X2.5) change in prevalence between whole cohort and subgroup.

Regarding number of diseases, more information can be seen in Figure 4-3, where participants had a minimum of 1 disease category and a maximum of 33 , with the average being 3 and the upper quartile 6.

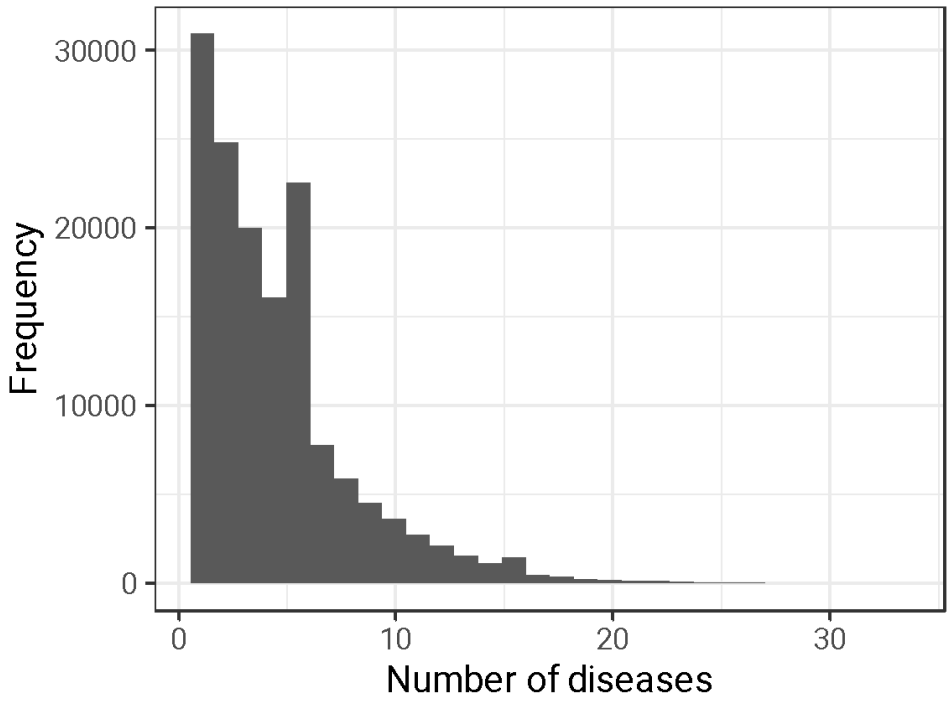


Figure 4-3. Frequency of the number of disease categories per participant.
Only those participants with at least one category diagnosis were selected (minimum = 1, maximum = 33)

We next determined whether there were distinct disease patterns or comorbidities (two or more diseases in the same individual), found in the different ageing phenotypes by calculating disease co-occurrence in pairwise associations Figure 4-4

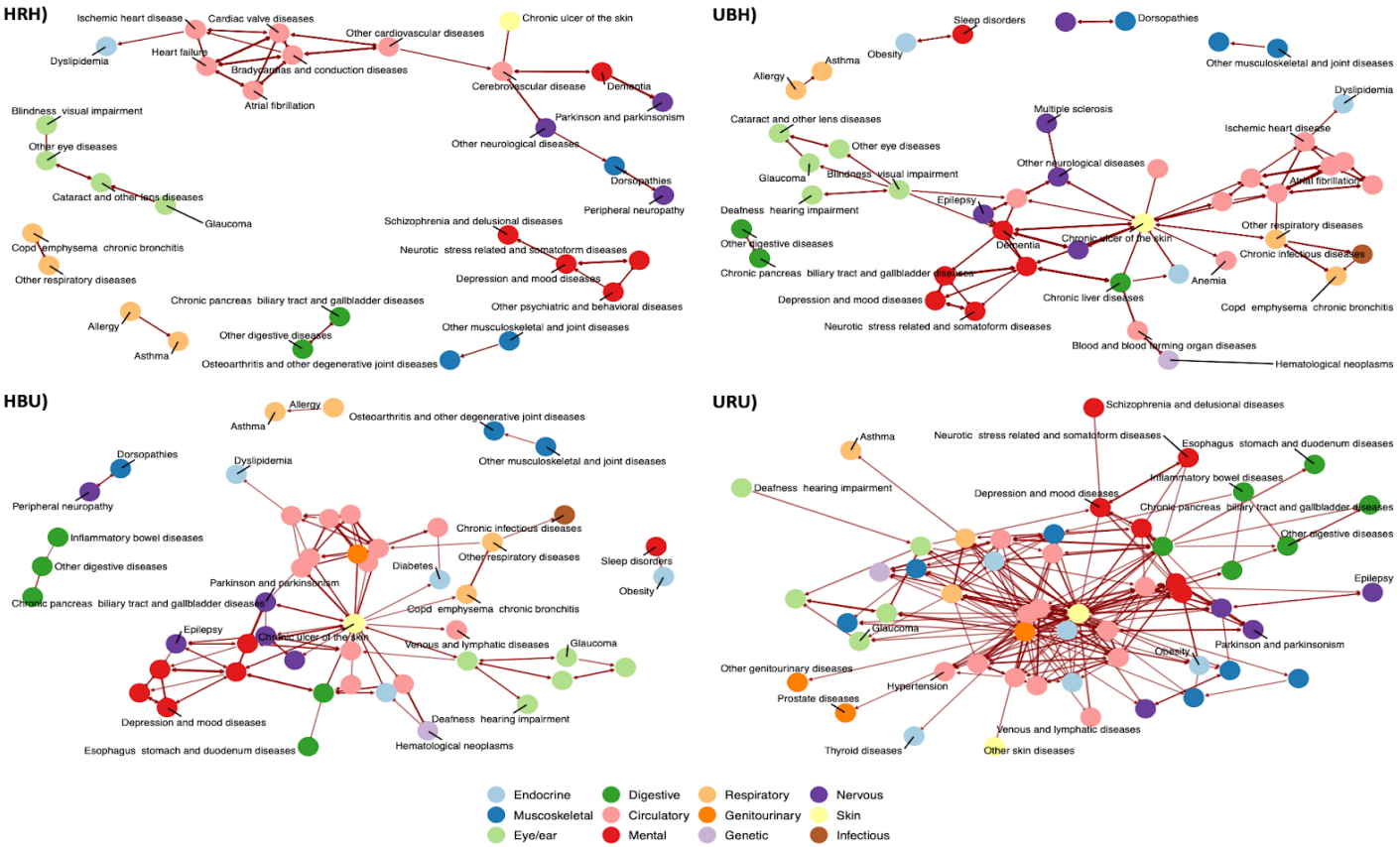


Figure 4-4. Relative Risk between disease categories plot for the four ageing phenotypes. Disease pair associations with an absolute relative risk higher than 2 and confidence lower than 0.2 were plotted. All results can be seen in https://github.com/InFlamUOB/Thesis_Chap4. Graphs were created using the package igraph and function gggnet2 and legends defined through Appendix 5

Relative risk (RR) is a measure of comorbidity strength used in previous studies of disease networks (Fotouhi et al. 2018; Dönertaş et al. 2021) where RR is equal to the ratio of the number of times that an ordered pair of diseases occur in the empirical data to the expected number of times it would occur in a random network. Looking into disease pair risk differences between categories (confidence < 0.2), the highest RR between conditions was found to be in URU participants between “Parkinson’s and parkinsonism” and “Dementia”: 19.72 (14.38 – 24.38)) with the other phenotypes displaying strong relationships between nervous and mental health categories. Similar findings have been previously reported (Ben Hassen et al. 2022), with neurological disorders clustering strongly together in multimorbidity analysis, with a strong age dependency. All results can be found in our repository at https://github.com/InFlamUOB/Thesis_Chap4.

Also, a higher number of significant disease pairs were found in URU participants (37 nodes), with a higher number of associations between ICD10 code categories (Figure 4-4). Diseases categories with the most number of associations were heart failure, chronic ulcer of the skin and COPD, emphysema and chronic bronchitis. Also, differences in RR between categories in diseases with a well-known spectrum of associated comorbidities could be assessed (D. Tomic, Shaw, and Magliano 2022; Pearson-Stuttard et al. 2022), with diabetes associated with higher relative risk to more severe conditions (i.e ischaemic heart disease, blindness and chronic kidney disease) than HRH.

4.4.4 Genomic characterization of ageing trajectory categories.

Genetic makeup of our ageing clusters was assessed through GWAS (Figure 4-5). Considering associations with the literature-standard P value $<5 \times 10^{-8}$ as significant

(Panagiotou, Ioannidis, and Genome-Wide Significance Project, 2012), summary statistics of significant variants can be found in Appendix 6.

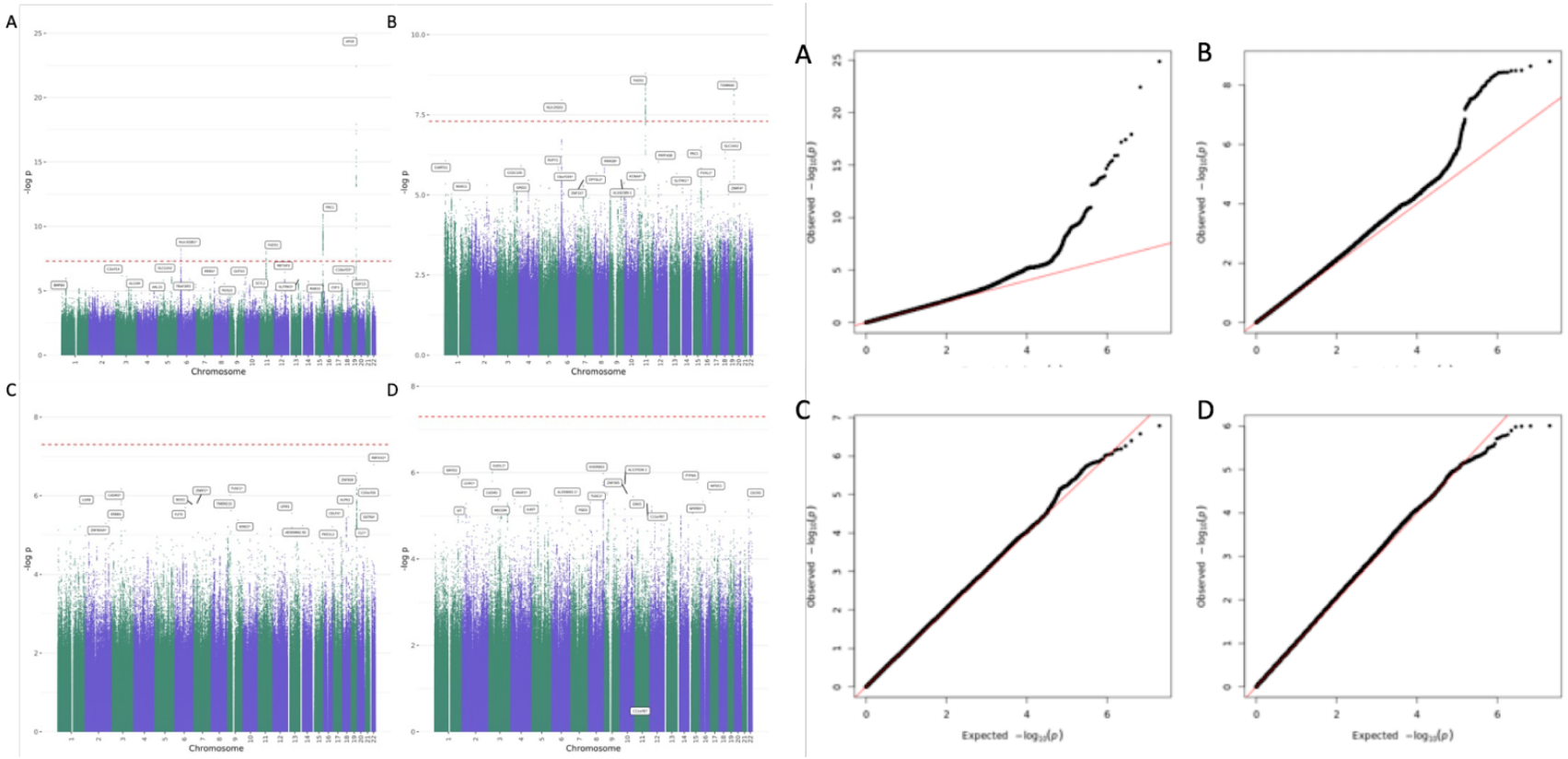


Figure 4-5. GWAS results for the four ageing categories.
Left. Manhattan plots showing 20 randomly annotated intergenic variants of above log(p)>5 value. Annotation was performed using the Ensembl Variant Effect Predictor (VEP) database. Right. QQ plots showing the observed p-value for each SNP A) HRH B) URU, C)UBH, D) HBU.

In total, as seen in Figure 4-6 and in Appendix 6**,** between all four categories, significant variant associations were found only in the URU and HRH categories. 108 (95 with “viable” rsID) were unique to the HRH participants with annotated genes corresponding to APOE, NECTIN2, APOC1, TOMM40, PRC1, APOC1P1, RCCD1, VPS33B, TBC1D22B and HLA-DQB1. The lead SNP for the HRH category was rs429358 (P = 1.36 x 10^25), a missense APOE e4 determinant, with an effect size (β = -0.18) in agreement with previous pro-longevity/healthy ageing phenotypes (Timmers et al. 2019). APOE e4 allele is traditionally associated with adverse health outcomes, including an increased risk of Alzheimer's disease, cardiovascular disease, and reduced life expectancy. The e2 allele (rs7412), said to confer protection, was not found as significant association in any category (C.-L. Kuo et al. 2021) . 12 associations were found for the URU set (rs174541 (FADS2), rs174576 (FADS2), rs4246215 (FEN1), rs174544, rs174548, rs174549, rs174555, rs174556, rs174557, rs174560, rs174561, rs751229834, all annotated to FADS1) and 8 association were shared between both (rs12972156, rs12972970, rs34342646 associated to NECTIN2, rs28456 (FADS2), rs2075650, rs34404554, rs71352238, rs11556505, associated with TOMM40). All gene annotations were performed using Phenoscanner (Kamat et al. 2019), which uses Ensembl Variant Effect Predictor tool (VEP) (McLaren et al. 2016) and BEDOPS (Neph et al. 2012) .In those 8 SNPs shared between categories, similar effect sizes were seen but with opposite directions (negative in HRH and positive in URU). Upon checking GWAS Catalog (Sollis et al. 2023) and the Phenotype-Genotype Integrator Database (Ramos et al. 2014), traits followed previously reported longevity studies, with significant associations found for coronary artery disease (rs4420638), Alzheimer’s disease (rs2075650, rs769449, rs157580, rs6857, rs4420638, rs56131196, rs429358), blood pressure (rs4420638) and cholesterol related traits (Harold et al. 2009; Seshadri et al. 2010; Fortney et

al. 2015; Dönertaş et al. 2021) . A full list of all GWAS Catalog results can be seen in the GitHub repository (https://github.com/InFlamUOB/Thesis_Chap4)

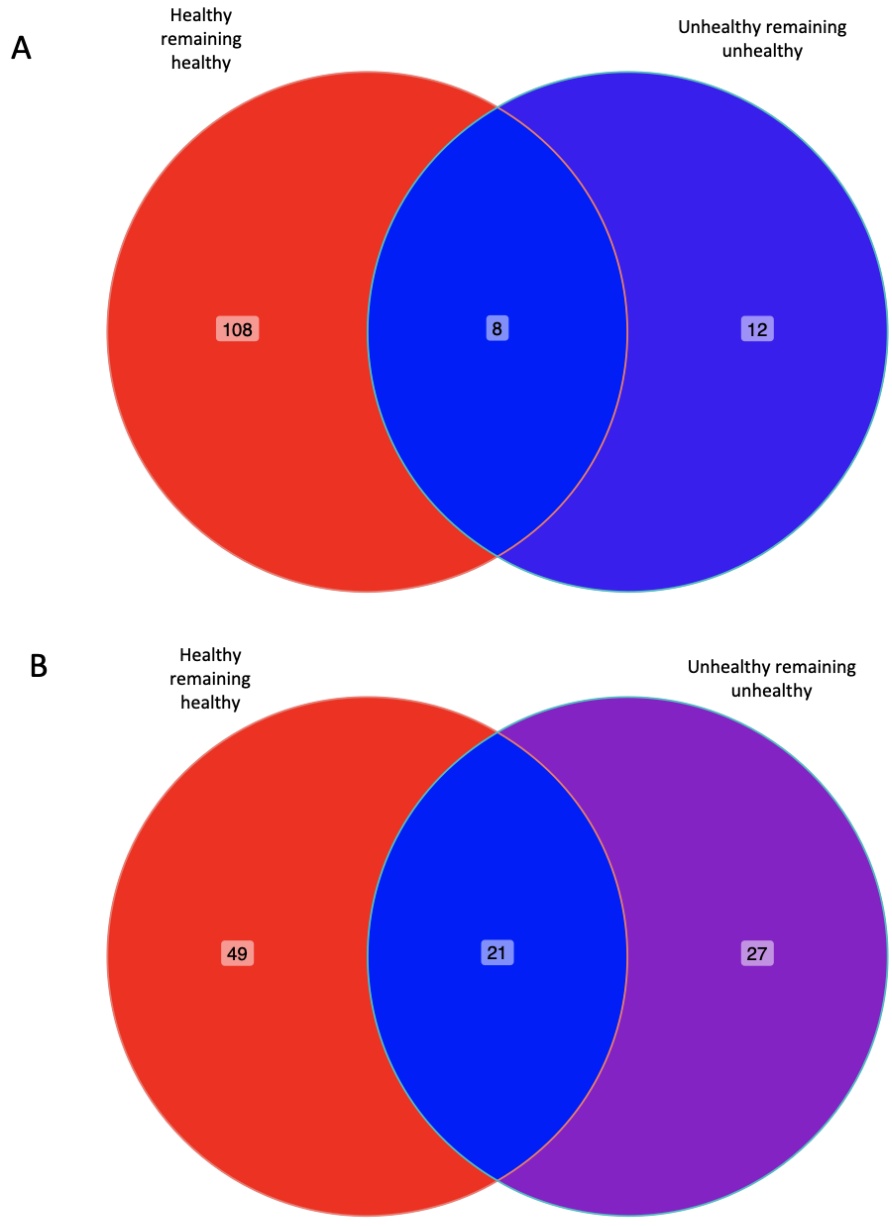


Figure 4-6. Venn diagrams showing significant genetic associations in URU and HRH categories.
A) Significant variants extracted from GWAS. B) Annotated genes

4.4.5 Gene annotation

As described in Methods, linkage disequilibrium (LD) was not taken into account to expand SNPs (Dong et al. 2021) or assess highly correlated SNPs (fine-mapping) for prioritisation, with only proximity and colocalisation with functional elements eQTL integrated. In total 97 HGCN genes were annotated, with 13 annotated based on both proximity and eQTL data (APOE, APOC1, TOMM40, PVRL2, PRC1-AS1, PRC1, RCCD1, VPS33B, HLA-DQB1, FADS2, FADS1, MIR1908, FEN1) and the rest (84) based on eQTL information alone. The most annotated gene was PRC1, with 78 SNPs found as significant in HRH categories annotated to it. All annotations can be found in https://github.com/InFlamUOB/Thesis_Chap4.

4.4.6 Functional enrichment

Given our major motivation was to use these identified gene associations to help infer any biological underpinning for the four ageing phenotypes, genes were then functionally enriched using the gene2func tool in FUMA (Watanabe et al. 2017), where annotated genes are tested for over-representation in publicly available gene sets (MsigDB v7.0)(Subramanian et al. 2005). Significant mappings on both HRH and URU annotated genes revealed canonical pathways and gene ontology biological processes enriched for immune related and metabolic pathways respectively (Figure 4-7). Interestingly, regulatory target gene sets were only enriched in URU annotated genes. All genes per pathway and databases available are summarised in https://github.com/InFlamUOB/Thesis_Chap4, including analysis with and without the overlapping 21 genes.

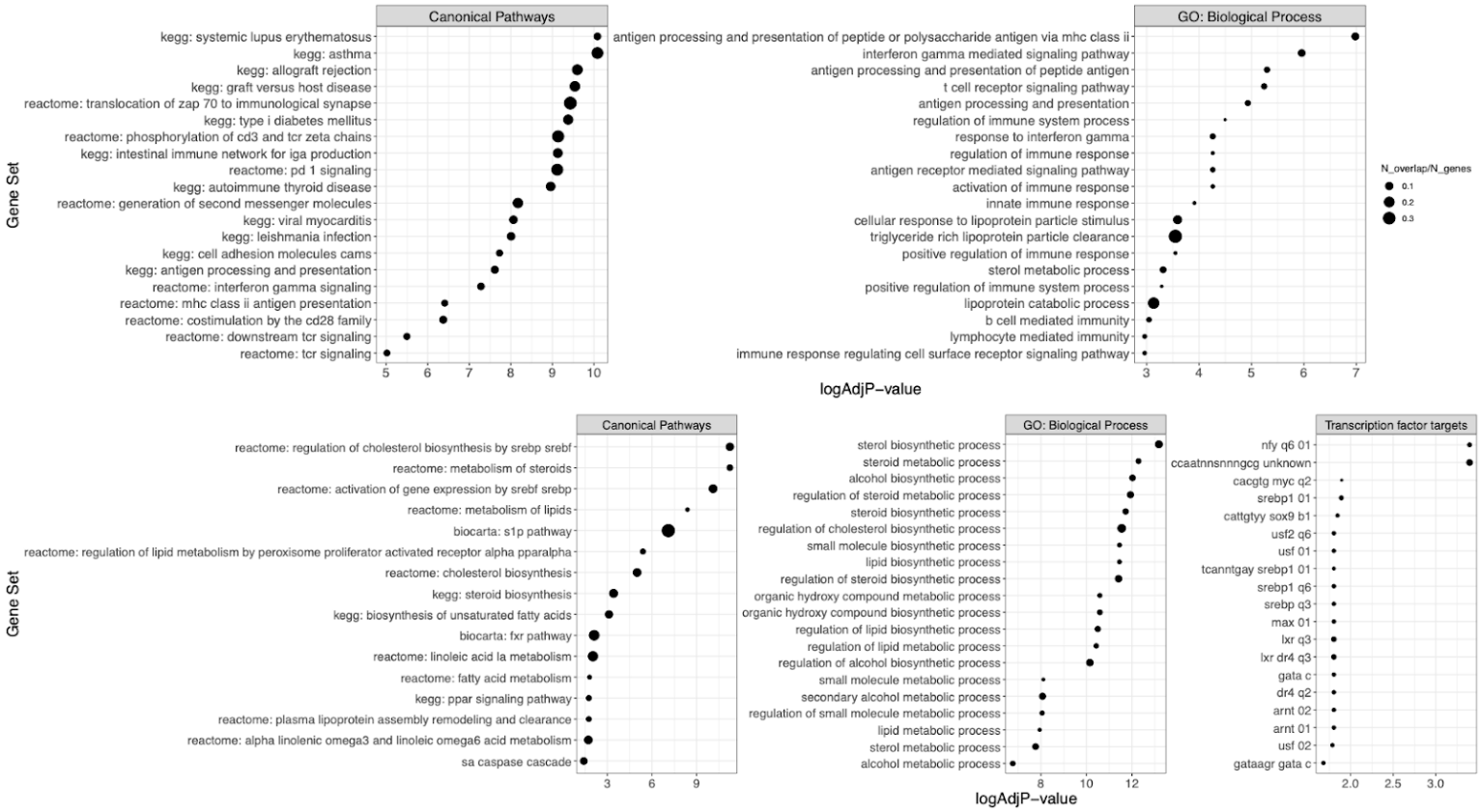


Figure 4-7. FUMA gene enrichment results

(Above) Top 20 results for the 49 genes unique to HRH signature and 21 genes belonging to both HRH and URU. (Below) Top 20 results for the 27 genes unique to URU signature and 21 genes belonging to both HRH and URU. Results obtained using FUMA's gene2func tool.

The HRH gene signature is enriched with immune biological processes, with the most significant association being positional gene sets (adjusted p value of 2.2 x 10^17), with C4A, C4B, CYP21A2, RNF5, AGER, NOTCH4, HLA-DRB9, HLA-DRB5, HLA-DRB6, HLA-DRB1, HLA-DQA1, HLA-DQB1, HLA-DQB1-AS1, HLA-DQA2, HLA-DQB2, HLA-DOB, HLA-DPA1, HLA-DPB2 genes all associated with chromosome 6p 21. This position belongs to the Human Leukocyte Antigen (HLA) critically related to immunity and inflammation and functionally associated to more than 120 major diseases (Borchers et al. 2014).

In contrast, the URU signature is dominated by metabolic and lipid synthesis enrichment, with the main dominating module being sterol biosynthetic process with an adjusted p value of (6.664x10-14) and 10 genes (DHCR24, SCD, SREBF1, ACACA, ABCG1, SREBF2, HMGCS1, INSIG1, FDFT1, SQLE) involved. Of special relevance is the fact that transcription factor targets were found for the URU subgroup and not the HRH, with recent studies highlighting a few conserved transcription factors (TFs), such as the above mentioned SREBF1 and SREBF2, as key controllers of the molecular footprint of ageing (Bou Sleiman et al. 2020).

4.4.7 Ageing databases

Given our focus on ageing trajectories, all annotated genes were then compared to literature-based ageing databases (Figure 4-8). These databases are GenAge Human (genes associated with lifespan in humans or closely related species), human orthologs in GenAge Model Organism (genes modulating lifespan in model organisms), CellAge (genes regulating cellular senescence) (Barardo et al. 2017; Tacutu et al. 2018), Aging Bank (Y. Gao et al. 2022) and longevity related genes extracted from different recent studies (Timmers et al. 2019; Zenin et al. 2019). We also incorporated the overlapping immune and age

associated genes found as conserved orthologs (Fabian et al. 2021). All information can be found in https://github.com/InFlamUOB/Thesis_Chap4.

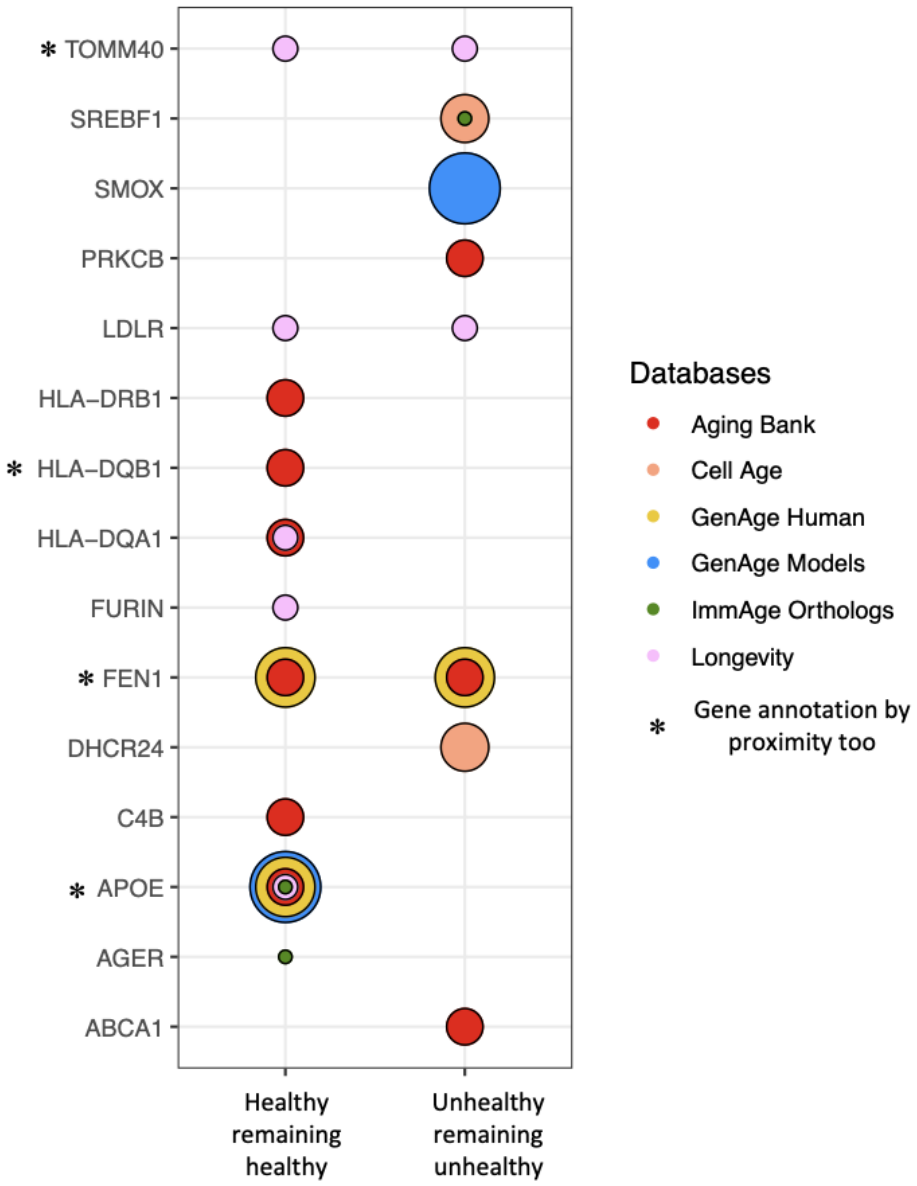


Figure 4-8. Association of annotated genes with key ageing databases.
Presence of ageing-related genes determined by 5 ageing databases and 1 study in our ageing categories. Genes annotated through proximity as well as eQTL are signalled through a *.

APOE, is one of the only consistent genome-wide significant associations for longevity and as such, present in nearly all of the databases included. Through our annotation mechanisms, rs6857, rs157580, rs429358 and rs769449 were the only variants associated with APOE, with them all found significant associations only with HRH cohort and them all displaying a negative effect. Regardless, TOMM40 which is located close to gene APOE at chromosome 19q13.32 (Joshi et al. 2016), did have representation in both categories. Also, as key highlights, shared genes between categories were found to be associated to genetic risk with opposite effects. The TOMM40 gene previously mentioned, was annotated by 18 SNPs through eQTL data and by 10 through proximity analysis. From these, 7 genetic loci were shared between HRH and URU categories (rs71352238, rs11556505, rs12972156, rs12972970, rs2075650, rs34342646, rs34404554) and in all different effects were reported with HRH consistently reporting negative effect sizes (betas) and URU positive. For LDLR, the low-density lipoprotein receptor, 11 SNPs were found associated through eQTLs and one (rs28456 coding for FADS2) was shared between categories, and again reported opposite effects (Beta = -0.0782 for HRH; Beta = +0.1079 for URU). From Zenin et al. 2019 database, 3 out of 40 significant SNPs identified in their study (rs2075650 (Fortney et al. 2015), rs4420638 (Deelen et al. 2014) and rs429358 (Joshi et al. 2016)) all associated to the the APOE/TOMM40 locus were also found in our analysis albeit none where the final 12 significant loci identified with health span in their study (Appendix 7). All annotated genes and summary stats can be found in https://github.com/InFlamUOB/Thesis_Chap4.

4.5 Conclusion and discussion

In this study, we have made use of UKBB data to determine an individual's ageing trajectory through calculation of the PhenoAge metric (Levine et al. 2018), a well-established biological age proxy. Longitudinal information available through GP clinical records, with repeated

measurements were used to help achieve approximations of an individual's biological age over time (Alpert et al. 2019). Acknowledging the high inter-individual variability seen in deeply phenotyped longitudinal healthy cohorts (Kaczorowski et al. 2017; Alpert et al. 2019; Ahadi et al. 2020; Tebani et al. 2020) and the challenge of studying ageing genetics due to low heritability and limited samples, participants with similar trajectory behavior were grouped in to closely related phenotypes to capture elements of the ageing process (Kubben and Misteli, 2017). Ageing trajectories were identified by acknowledging that higher predicted PhenoAge scores over real chronologic age represented unhealthy signatures and lower ones, healthy. In these way, four trajectories were built: Healthy remaining Healthy (HRH; n= 20,395); Healthy becoming Unhealthy (HBU: n = 75,122); Unhealthy becoming Healthy (UBH: n = 94,130); Unhealthy remaining Unhealthy (URU; n = 9569). Considering these different trajectories as examples of participants ageing at different rates, we proceeded to understand if these four groupings could be further characterized by potentially different biological mechanisms.

In agreement with previous studies, a predominant cardiometabolic disease pattern was seen associated with our unhealthy trajectory groups, with diseases such as diabetes, obesity, or heart failure being highly prevalent in URU participants, associated too with male sex and worse health outcomes and lifestyle practices (high BMI, comorbidities, smoking, and shorter parental lifespan) In turn, HRH displayed the opposite signature and had less prevalence of most diseases with the only disease seen at a higher prevalence than baseline being osteoporosis, a disease highly associated to the female sex, predominant in the category. The intermediate, transitioning categories HBU and UBH displayed intermediate signatures with HBU consistently displaying worse health outcomes, higher disease prevalence and worse lifestyle practices.

Enrichment analysis through FUMA of our annotated GWAS results for each ageing category revealed marked metabolic and immunological signatures in URU and HRH respectively. The metabolic signature uncovered in URU is characterized by reactive biolipid molecules synthesis, which amplify the inflammatory response through enhancing the inflammatory cytokine cascade, upregulating the innate response to tissue damage or pathogens, including increasing inflammatory cytokine expression (Rea et al. 2018). Of special relevance too, are the sterol regulatory element-binding protein (SREBP) transcription factors, SREBF1 and SREBF2, which we hypothesize might act as a unifying link between the enriched cardiometabolic disease patterns, metabolic signature and ageing genetic risk seen in URU. In recent studies, SREBP TFs have been claimed as one of the key putative ageing modulators underlying ageing footprints across tissues and species (Bou Sleiman et al. 2020). Responsible for regulating cellular lipogenesis and lipid homeostasis, as well as promoting inflammation and mediating metabolic reprogramming (Fowler et al. 2022; Kusnadi et al. 2019) they have been closely associated with type II diabetes, cancer, innate immunity, neuroprotection, and autophagy (Shao and Espenshade 2012). Also, other TFs pathways were identified such as autophagy related GATA TF or liver X receptor (LXR) (Kang et al. 2015; Zhu et al. 2021). In order to further ascertain possible underlying molecular patterns, a more exhaustive analysis of regulatory regions associated to our GWAS associations would be needed (Dong et al. 2021), with current leading theories on disease development and complexity focusing on the study of non-coding regions of the genome (Boyle, Li, and Pritchard 2017; Barbeira et al. 2021). Even though eQTL mappings were used, harnessing other resources such as ENCODE (ENCODE Project Consortium 2012) or Roadmap Epigenomics projects (Kundaje et al. 2015) would be useful to identify extra mechanistic insights, with maps of histone modifications, transcription-factor binding, chromatin accessibility, three-dimensional genome structure and other annotations acknowledging different cell types and tissues (Tansey, Cameron, and Hill, 2018). With the

integration of such resources, elucidation of the TFs modules as key integrating agents of complex ageing and inflammatory responses would be possible, with promising guides such as LXR. LXR plays a dual role in metabolism and inflammation (Dixon et al. 2021), acting as a regulator of the inflammatory response through changes in lipid metabolism with the activation of ABCA1, a sterol transporter altering membrane cholesterol homeostasis able to inhibit NF-KB (Ito et al. 2015). LXR also acts as a glucose sensor regulating glucose homeostasis and insulin sensitivity (Gerin et al. 2005). As such these genes have been highly associated to diseases such as Type 2 diabetes, obesity or other cardiometabolic diseases, highly prevalent in the URU category. Also TF GATA, associated to genes TPM3, MYRF, FTH1 and PVRL2, is highly associated with the hallmarks of ageing: DNA repair and impaired autophagy (Kang et al., 2015; Zhu et al., 2021).

With regards to the HRH signature, genetic associations point towards an immune-enriched signature (albeit unspecific) and most relevantly, a number of longevity-associated SNPs. Examples such as rs2075650, rs4420638, rs429358 with strong negative effect sizes denoting a protective phenotype matching the literature (Zenin et al. 2019). Such is the case of the SNP with the biggest effect size from our analysis, APOE e4 variant rs429358, implicated in age-related diseases and traits such as late-onset Alzheimer's disease or cardiovascular disease and depleted in centenarians. This same variant was also found as key genetic driver of PhenoAgeAccel in a recent analysis exploring UKBB's data (C.-L. Kuo et al. 2021). Regardless , in their study, opposite effect direction to expected (and to ours) was seen. They concluded that cofounding through the e3e3 genotype and plasma level cholesterol might be the reason, with inflammation acting differently in subtypes. Interestingly this effect seems to have been mitigated in our study by splitting it into subtypes, getting rid of this genotype in one of the other categories (C.-L. Kuo et al. 2021). In our work the second

APOE isoform and e2 allele (rs7412), was not found as significant, but warrants future inspection.

Worth mentioning too is the fact that no significant associations were found in the intermediate ageing phenotypes HBU and UBH, with maybe a higher sample size required to properly uncover signals, given these trajectories might be grouping more heterogenous individuals with stronger environmental and lifestyle effects. In previous work using deep and longitudinal phenotyping (Ahadi et al. 2020), 4 enriched ageing patterns or “ageotypes” were derived from key molecular pathways that changed over time in a given individual. Two of these four “ageotypes” matched our metabolic and immune signatures, with the other two being defined by liver and kidney dysfunction pathways. Liver and kidney dysregulations have also been found in recent studies associating multimorbidity and metabolic studies of healthspan, with impaired glucose homeostasis, low-grade inflammation, and lipoprotein metabolism (Pietzner et al. 2021) as other key markers too. Further work, such as a closer study of multimorbidity clusters and network modules in our diseases (Dong et al. 2021), would help understand if this kidney and liver dysregulation is seen in HBU and UBH too. Also, associations with disease sequences in time would be of special interest too, with chains of disease causation separated in time allowing us to further profile the development of each ageing cluster (Jensen et al. 2014; Giannoula et al. 2018; 2021; Kuan et al. 2023). Also, and as studied in (Ahadi et al. 2020), further research onto the associations between lifestyle changes (i.e BMI variations, diet restrictions, smoking cessation or specific medications such as metformin) to transitions from unhealthy to healthy trajectories could pave the way for better understanding of how improvements in health can be targeted using selective interventions or lifestyle changes.

Also, many different biological age metrics exist as reported before. PhenoAge was used due to its well-established use in research and the simple serum biomarkers required, routinely assessed in clinical practice and so more readily available for longitudinal analysis. Since its derivation in 2018 (Levine et al. 2018), many studies have employed serum proteins as instruments to predict health risk and disease (Emilsson et al. 2022; 2018; Sinnott-Armstrong, Tanigawa, et al. 2021; Sinnott-Armstrong, Naqvi, et al. 2021), with Li and co-workers recently assessing the genetic underpinnings and health span association of most of the markers involved in PhenoAge (X. Li et al. 2021). From their work, glucose control was hypothesised to be a putative causal mechanism in healthspan maintenance, with glycemic biomarkers and those related to HDL metabolism showing the greatest genetic effects. As expected, variants significantly associated to our more extreme categories (HRH and URU), were highly associated through GWAS Catalog mappings (https://github.com/InFlamUOB/Thesis_Chap4.) to cholesterol metabolism and most PhenoAge biomarkers. For the future, a closer look onto how our results relate to each of the 9 markers used to build our score would be warranted with work such as (C.-L. Kuo et al. 2021) disentangling its effect in single point measurements. Although, in the end, with each of these markers being massively polygenic and complex by themselves (e.g 266 independent loci identified for CRP-regulation (Said et al. 2022)), the power of our analysis probably emerges from their combination, where shared underlying molecular pathways are helping uncover novel and overlapping genetic variants and associations (Ligthart et al. 2016).

Also, investigating other markers and studying other biological age metrics would be of interest, with the previous mentioned work (C.-L. Kuo et al. 2021) concluding, after comparing the genetic predisposition of both PehnoAge and BioAge (another biological age metric), that the two measures capture different ageing domains with maybe more than 1

metric required to generate a complete understanding of the ageing signature. Probably these metrics should encompass as many omics layers as possible, with recent research focusing on DNA methylation and epigenetic clocks or metabolomic studies with haematological metabolism, described as a key player in ageing phenotyping (Pietzner et al. 2021). Also, a comprehensive 76 protein model was found to be highly correlated (r = 0.94) with chronological age (Tanaka et al. 2018). In the end, the genetic basis of the endpoint phenotype, in this case different health span/ ageing trajectories is expected to reflect the genetics of all the intermediate endophenotypes, here represented by the choice of molecular markers (Sinnott-Armstrong, Naqvi, et al. 2021).

In terms of study methodology, there are limitations. We acknowledge different analysis criteria could have been performed for tackling the imputation of missing blood marker data, or different disease classification categories implemented, with for example a broader range of age-related diseases considered (Kim et al. 2021). Our hypothesis was based on uncovering mechanisms driving ageing trajectories and so we chose the 60 chronic diseases most associated with multimorbidity (Calderón-Larrañaga et al. 2017), but it is true that this grouping of ICD-10 codes makes comparisons with the current literature more complex (Dong et al. 2021). Also, the use of hospital data from UKBB and not self-reported health data or the integration of GP records, can produce a biased disease description of the population, with really healthy patient markers not captured, given GP visits are normally a consequence of ill-health. Regardless the use of primary care data codes in UKBB is still a novel tool (Hanlon et al. 2022) and we are benefiting from the long follow-up time (25 years) and an extensive coverage of diseases provided by the UKBB linked hospital inpatient data.

Another limitation is that no fine-mapping was performed to assess if genomic variance is truly independent and not merely due to linkage disequilibrium, with all SNPs studied

downstream and their functional enrichment prioritized (Uffelmann et al. 2021). Also, UKBB participants are recognised to be less socioeconomically deprived, have fewer lifestyle risk factors, and have a lower prevalence of long term conditions (LTCs) than the UK population in general (Hanlon et al. 2022), making multimorbidity studies challenging (Kuan et al. 2023) and not fully representative of worldwide populations.

5 Chapter 5: Discussion and thesis conclusions

This thesis investigated three applications of computational biology focusing on three key areas of research:

- Discovery: Machine learning as a bioinformatic tool (Chapter 2)
- Translation: Clinical translation through the study of prognostic models (Chapter 3)
- Complexity: Data integration and the study of complex systems (Chapter 4)

For all pieces of work, the objective was twofold, to create and improve on an analysis pipeline and to generate insight into the specific biological problems studied, either as case-studies (Chapter 2) or as main drivers of the work (Chapter 3 and 4). As it is common in such big pieces of work, the projects and aims of the thesis evolved and adapted in time, with the journey starting through the development of data analysis frameworks for the study of biological data. With the pandemic, our work got repurposed and we got to explore a novel and exciting avenue: translational medicine. The analysis then grew naturally larger, through the exploration of big datasets such as UKBiobank and the questioning of higher order concepts such as ageing and health to disease transitions.

5.1 Overview of major findings

In Chapter 2, I recount the journey of developing a data-driven, generalizable, robust and low-bias machine learning tool that generates easily interpretable outputs for actionable biomarker discovery. This tool has as objective to both aid in translational research, helping guide the selection of explanatory features of -omics data, as well as generate insights into underlying patterns and help disentangle the ever-increasing amount of data available. Work is divided into three main sections. First, we developed a robust machine learning workflow based on bootstrapping of regularization algorithms for feature selection which was validated in three different datasets. Based on our approach, we were for example able to reduce the

number of potential causative genes in a transcriptomics experiment to under 100 (from an input of over 22,000 genes) whilst the high confidence selections were reduced to less than 15. Then, this same pipeline was adapted for the study of the novel “Golden Hour” dataset, presenting unique data on the ultra-early immune response to trauma. Through our work, we assessed the predictive power of early immunological data to detect MultiOrgan Dysfunction development, with measurements at three different time points. A final multivariable model identified three potential biomarkers relating neutrophil function. These three features were: decrease in CD62L expression after fMLF treatment, CD63 expression and monocyte percentage. With the inclusion of current clinical scoring systems such as NISS and P14, the final AUC model performance achieved was 0.92, improving the current clinical scoring systems PS14 by 8% and NISS by 13% respectively. Finally, this ML pipeline evolved to incorporate other areas of interest such as explainable AI methods for more comprehensive feature selection methods as well as included more algorithms to complement analysis. In order to make it useful for the community, it was deployed as a complementary tool to allow a quick and simple analysis through a webapp (at https://inflam.shinyapps.io/MachLearnBiomarker2/).

In Chapter 3 we repurposed our machine learning pipeline to implement clinical prediction models as part of a collaboration with the CovidSurg Collaborative (‘CovidSurg – Globalsurg’), a NIHR Global Surgery Research Unit built out of the need to derive evidence-based guidelines for surgical practice. This was done during the pandemic, as part of their research exploring how to restart surgery in a safely manner through the study of patients with SARS-CoV-2 infection who had surgery at different hospitals and countries. For thes datasets with 126,000 patients from more than 1158 from 119 countries was used. As part of the work, two different clinical prediction models were built for clinical decision-support:

- Prediction of postoperative mortality risk in patients with perioperative SARS-CoV-2 infection (CovidSurg Mortality Score) (COVIDSurg Collaborative 2021). Model deployed here: https://inflam.shinyapps.io/RiskAppLukeAugust2/
- Prediction of postoperative pulmonary complications in adults undergoing elective surgery (GlobalSurg-Pulmonary Score). Model deployed here: : www.surgicalrisk.app

Both models were deployed following similar methodological structures, with a strong focus on variable importance assessments and feature selection. After testing different algorithms and variable combinations, both models were implemented as logistic regression for the classification of patients at high risk of either mortality or postoperative pulmonary complications. Logistic regression provided the best compromise between performance and interpretability. Also, a thorough performance assessment through discrimination and calibration metrics was done in both, although the validation schemes were adapted to the resources and data available. More specifically regarding results, the COVIDSurg Mortality Score consists of a logistic regression model of the following categorical variables: age (in deciles), American Society of Anaesthesiologists (ASA) grade, revised cardiac risk index (RCRI) score and preoperative respiratory support. Regarding performance, it is a well calibrated model, that shows high discrimination for predicting mortality in patients with perioperative SARS-CoV-2 with AUROCs in the derivation (0.728 , 95% CI 0.714 – 0.7428) and validation (0.804, 95% CI 0.774 – 0.833) sets consistent with published non-surgical COVID risk scores. The predicted score is globally applicable through a webapp (https://inflam.shinyapps.io/RiskAppLukeAugust2/) with variables that are readily available across all resource settings. Continuing with our collaboration, a second model with updated data was designed targeting pulmonary events, which are the major driver of death after elective surgery and increased in frequency and severity during the SARS-CoV-2 pandemic. For the GlobalSurg-Pulmonary Score, ten predictor variables were selected: age, sex,

American Society of Anaesthesiologists (ASA) grade, operation grade, pre-existing respiratory disease, RCRI, anaesthesia type, indication, interaction between country income and background community SARS-CoV-2 infection and area of surgery. A logistic regression model was built that demonstrated good discrimination and calibration upon internal validation (AUROC 0.7729, 95%CI 0.7509-0.7948; Brier score: 0.020, calibration-in-the-large: 0.072, slope: 0.930). The model was then externally validated through two different datasets with an acceptable performance in CovidSurg-Cancer (AUROC 0.746, 95% CI: 0.733-0.760; Brier score: 0.036, CITL: 0.324, slope: 1.002), but with some miscalibration in RECON data (AUROC: 0.690, 95% CI 0.663-0.719; Brier score: 0.045, CITL: 0.711, slope: 0.875). Miscalibration was assessed through transportability frameworks and the model deployed to an online web application: https://surgicalrisk.app with the hope to inform surgical consent, resource allocation, and hospital level prioritisation. Both studies were designed and reported according to best practice guidelines and were at a low risk of bias. As key takeaway too, such a direct collaboration with the surgical team has been enriching, with a clear idea of the useful final product helping guide the study methodology. Also, during deployment, a real-time feedback on building easy, interpretable design interfaces with actionable outputs was essential too.

In Chapter 4, we have made use of UKBiobank data to determine an individual's ageing trajectory through longitudinal measurements of the PhenoAge score (Levine et al. 2018), a well-established biological age proxy. Four main aging trajectories were found based on its relationship with chronological age and were defined as: Healthy remaining Healthy (HRH; n= 20,395); Healthy becoming Unhealthy (HBU: n = 75,122); Unhealthy becoming Healthy (UBH: n = 94,130); Unhealthy remaining Unhealthy (URU; n = 9569). Considering these different trajectories as examples of participants ageing at different rates, we proceeded to understand if these four groups could be further characterized by potentially different

biological mechanisms. Looking at key lifestyle and health outcomes, chronic diseases patterns and genetic makeup, we discovered that: HRH are female dominated, have a lower number of comorbidities, better overall health rating, lower BMI, more never-smokers and lower deprivation score, all factors we would expect to influence the rate of biological ageing. This was confirmed by lower mortality in the healthy group. Interestingly both the mother's and father's lifespan was higher too. Regarding URU, it was dominated by the male sex, high BMI, comorbidities, smoking, and shorter parental lifespan. To better analyse chronic disease patterns in our ageing phenotypes, we defined 60 chronic disease categories from available ICD-10 codes following (Calderón-Larrañaga et al. 2017)'s work. Cardiometabolic diseases were the most prevalent URU, with dominating diseases such as diabetes, obesity, or heart failure. In turn, HRH had less prevalence of most disease. Also, upon pairwise analysis of diseases assessing relative risk, a higher number of significant disease pairs were found in URU participants (37 nodes), with parkinsons and dementia being the most highly associated. The intermediate, transitioning categories HBU and UBH displayed intermediate signatures with HBU consistently displaying worse health outcomes, higher disease prevalence and worse lifestyle practices. Finally, GWAS was performed revealing the genetic makeup of the different trajectories. After significant variants were annotated through both proximity and eQTLs, enrichment analysis revealed a marked immune signature for HRH participants and a metabolic signature driven by lipid biosynthesis for URU. Also, conserved transcription factors SREBF1 and SREBF2, key controllers of the molecular footprint of ageing (Bou Sleiman et al. 2020), postulated as key regulatory element driving the URU ageing phenotype Finally, focusing on ageing databases, significant associations to age-related genes were found for both HRH and URU, with pro-longevity associated variants (e.g rs429358, rs4420638, rs2075650) dominating HRH, with opposite effect directions in the shared significant variants seen (rs2075650). No significant genetic associations were found for UBH and HBU which we postulate to be more highly influenced by lifestyle associated

factors.To the best of our knowledge, no previous research on both genetic background and longitudinal measurements has been done in the study of biological ageing metrics, with previous research focusing on either longitudinal assessment with no mechanistic input or molecular associations but at a single point measurement.

5.2 Contributions to the literature

- Developed and deployed a publicly available tool that includes bootstrapping of regularized models and permutation analysis for biomarker discovery (https://inflam.shinyapps.io/MachLearnBiomarker2/).
- Evaluated the predictive power of ultra-early immunological data in determining adverse outcomes in trauma patients, identifying three potential biomarkers (reduced CD62L expression after fMLF treatment, CD63 expression, and monocyte percentage) and offering a model with superior performance compared to existing clinical scoring systems.
- Launched a model for predicting postoperative mortality risk in patients with perioperative SARS-CoV-2 infection (CovidSurg Mortality Score): https://inflam.shinyapps.io/RiskAppLukeAugust2/.
- Introduced a model for assessing postoperative pulmonary complications in adults undergoing elective surgery (GlobalSurg-Pulmonary Score): www.surgicalrisk.app.
- Pioneered the integration of genetic makeup and longitudinal measurements in the study of biological ageing metrics, identifying four ageing phenotypes characterized by a spectrum of health outcomes and chronic disease prevalence.
- Uncovered longevity-associated genetic determinants with opposite directional effects in both the healthiest and unhealthiest signatures and shared all associations.

- Postulated transcription factor genes, SREBF1 and SREBF2—key regulators of the molecular footprint of ageing—as possible driving elements of our unhealthy ageing phenotypes.

5.3 Limitations

Chapter 2. An evolution of the machine learning pipeline can be traced throughout the chapter, with new iterations improving on previous weaknesses. For instance, as a common limitation for the whole machine learning pipeline in all its developments, no interactions between features were assessed as seen in other comparable pieces of work (Ali 2020). Finally, the web app was deployed, and although still in development given its lack of parallelization in the public server and need for proper software engineering frameworks, with debugging and testing strategies, the GitHub source code is a fully operating, parallelized code. Regardless, we do envision that there might be problems when dealing with different operating systems and software dependencies, with proper containerization/environments of our work through platforms such as Docker, designed to help developers build, share, and run modern application as better options for the future. Regarding the work on the classification of multiorgan dysfunction patients, internal validation was performed using bootstrapping and permutation analysis but the analysis only counted with 61 trauma patients, with external validation required to properly ascertain our models validity. Regardless, it is true that finding data comparable to our experimental design is difficult. The Golden Hour study, to the best of our knowledge, represents one of the few studies in the world with data on the ultra-early response to trauma (< 1h) with only (Cabrera et al. 2017) performing similar work, albeit with information on the first 2h. Also, to generate a more complete analysis on the complexities underlying trauma, different types of data would be required. To date, I have been involved in follow up work with steroid data and we have also applied our pipeline to the transcriptomic information from (Cabrera et al. 2017) attempting

integration given similarities in phenotypes between our patients. As a result genes better seen to be able to classify between MODS vs non-MODS have been found to be enriched for neutrophil degranulation pathways (Bravo-Merodio, Acharjee, et al. 2019), in close association with our neutrophil related identified immune features. Regardless, this is far from a proper integration, with deep phenotyping providing a more insightful biological understanding of the early trauma response. Moreover, in our study, women were underrepresented (16% of patients being female).

Chapter 3. Given similar methodological approaches were taken for the two prognostic models, limitations are shared, with them being mainly related to the field of ML itself. These could be summarized through potential dangers of ML algorithms such as: dataset shift or unintended discriminatory bias. Data shift is especially relevant in our case, with both models devised for an acute problem, the COVID19 pandemic and as such the model will require strong updates and revisions. Also, although we have explored the whole space of model building and deployment, both prognostic scores are still pieces of research with no further clinical translation and regulatory assessments taken. Also, clinical predictor variables were coded following standardized medical criteria and so collected categorically, a simplification that may have reduced statistical power and may explain a degree of miscalibration in RECON data in the GlobalSurg-Pulmonary score work. It is worth mentioning too that more models and pre-processing steps could have been implemented, including using methods for managing class imbalance (i.e., over/under sampling or Synthetic Minority Over- Sampling Technique (SMOTE)). However, these are criticised in clinical datasets due to risk of miscalibration (strong overestimation of probability to belong to the minority class), and we opted not to adopt these here.

Chapter 4 In terms of study methodology, there are limitations. We acknowledge different analysis criteria could have been performed for tackling the imputation of missing blood marker data, or different disease classification categories implemented, with for example a broader range of age-related diseases considered (Kim et al. 2021). Also, because of the use of hospital data from UKBB and not self-reported health data, this can produce a biased disease description of the population. Regardless the use of primary care data codes in UKBB is still a novel tool (Hanlon et al. 2022) and we are benefiting from the long follow-up time (25 years) and an extensive coverage of diseases provided by the UKBB linked hospital inpatient data. Another limitation is that no fine-mapping was performed to assess if genomic variance is truly independent and not merely due to linkage disequilibrium, with all SNPs studied downstream and their functional enrichment prioritized (Uffelmann et al. 2021). Also, UKBB participants are recognised to be less socioeconomically deprived, have fewer lifestyle risk factors, and have a lower prevalence of long term conditions (LTCs) than the UK population in general (Hanlon et al. 2022), making multimorbidity studies challenging (Kuan et al. 2023) and not fully representative of worldwide populations.

5.4 Planned future work

There are several extensions to this work which may be done; some are planned for validation and others to continue themes of these projects.

5.4.1 Chapter 2

As well as testing our model in external datasets, several interesting questions stemmed out from the project. For instance, we hope to find a marker or aggregation of markers able to substitute the complex anatomical information reflected in the clinical scoring systems NISS and PS14, and so yield an objective and real time predictive parameter. cfDNA and HMGB1

have been proposed as plausible substitutes (M. J. Cohen et al. 2009) but could not be tested in our dataset at the moment given too much missing data.

5.4.2 Chapter 3

With regards to the COVID19 work, it would be of interest to pursue the whole clinical workflow pipeline, with the journey required to study how to successfully implement prognostic models into clinical frameworks. (Royen et al. 2022). For the moment, we are still closely collaborating with the NIHR group and monitoring the quality of our deployed models and their applicability.

5.4.3 Chapter 4

Tackling big questions in biology such as ageing, health to disease transitions, multimorbidity and chronic inflammation has opened many avenues of research and improvement that we are currently pursuing and plan to continue working on for the future.

5.4.3.1 PheWAS

To gain more insight into the variants found to be significantly associated to our ageing trajectories, we will use PheWAS through a pipeline of analysis already implemented in Gkoutos lab. Currently, associations of key variants extracted from our work were performed through GWAS Catalog (Chaptrer 4), where for instance rs2075650, highly significant in both URU and HRH trajectories, was found to be associated to longevity as well as to LDL cholesterol and Alzheimer disease. Through PheWAS, which stands for Phenome Wide Association Studies, we will be able to uncover any pleiotropic effects of our variants (predicted to be present in 60% of GWAS loci (Watanabe et al. 2019)) by examining the impact of genetic variants across a broad range of phenotypes using univariate regression models in our own UKBB dataset (Bush, Oetjens, and Crawford 2016). This amounts to flipping the direction of inference in GWAS, and going from exposure (genetic variant) to

outcome. Recent examples of PheWAS in multimorbidity studies have looked at the interplay between cardiovascular and neurological disorders (X. Zhang et al. 2022), where they found 5 distinct genomic loci with significant evidence of pleiotropy, with most displaying antagonistic effects.

5.4.3.2 Disease sequences

Taking advantage of the dates of diagnosis information, we are currently working on extending our disease research to understand more complex disease patterns, trying to determine if there are any distinct sequences of disease development in time found in the different categories. Similar work has been performed using Dynamic Time Warping (DTW) (Giannoula et al. 2018; 2021), an unsupervised clustering algorithm which measures similarities between two sequences that may vary in time or speed, with few other large-scale disease trajectory studies reported so far (Jensen et al. 2014). Given the large number of combinations available, most research focuses on targeted disease explorations (e.g diabetes (Haug et al. 2021), mortality (Paik and Kim 2021) or sequence analysis with just three diseases (CVD, Diabetes and Cancer) (Cezard, Sullivan, and Keenan 2022).

5.4.3.3 Lifestyle and transitions

Certain social, environmental and lifestyle factors such as dietary habits have been shown to be highly associated with the development of chronic diseases here studied (Furman et al. 2019). Therefore a closer look onto lifestyle factors longitudinally by extracting information from GP records on smoking, diet, BMI, or exercise could aid to better characterise our identified ageing trajectories. This will be of special interest for our transitioning phenotypes HBU and UBH, given their lack of genetic drivers might mean higher environmental influence driving their signature with even attempts to pinpoint the transition points as possibilities. Similar studies have been done, where (Ahadi et al. 2020) looked at associations in lifestyle

changes (i.e BMI variations, diet restrictions, smoking cessation or specific medications such as metformin). This approach could potentially facilitate a deeper understanding of how targeted interventions or lifestyle modifications can contribute to enhancing health outcomes.

5.4.3.4 Knowledge graphs

In this last package of work, we aim to study three key pieces of work, multi-omic integration, regulatory elements and network medicine. Our aim is to study complex biological systems by integrating all layers of biological information, from gene to trait/disease, across various scales, including protein-protein interactions, gene expression, epigenetic state, and signalling activity. Networks provide a natural framework for systematically investigating the various relationships between biological constituents within and between scales. Of particular interest to us is regulatory mechanisms, as cellular phenotypes depend on the influences of multiscale networks of molecular interactions that regulate gene expression, protein abundance, and signalling activity (J. C. Chen et al. 2014). However, obtaining such knowledge is challenging as it is often fragmented across publications, non-standardized repositories, and evolving ontologies. To overcome this challenge, we use knowledge graphs such as Bioteque (Fernández-Torras et al. 2022) and PrimeKG (Chandak, Huang, and Zitnik 2022), which provide a unified framework for biomedical data integration. Bioteque integrates 450,000 biological entities and 30 million relationships between them from over 150 data sources, while PrimeKG integrates 20 high-quality resources to describe 17,080 diseases with 4,050,249 relationships representing ten major biological scales. Our work aims to develop methods for integrating data describing these various levels of activity, ultimately enabling a better understanding of how the genome generates cellular phenotypes.

5.5 Conclusions

This thesis aims to investigate the applications of computational biology in biomedical research, focusing on three core areas: discovery science, clinical translation, and complex systems. In the realm of discovery science, we explored the use of machine learning in bioinformatics by developing a workflow based on bootstrapped regularization algorithms for dimension reduction and biomarker identification. We initially validated this approach using publicly available -omics data and then applied it to the novel "Golden Hour" study, assessing the predictive potential of immune information during the ultra-early (<1h) response to trauma. A model was constructed to identify patients at risk of developing Multi-Organ Dysfunction, revealing three immune markers related to neutrophil dysregulation as candidates for future validation studies. This workflow was subsequently converted into a functional pipeline and made publicly accessible through an open-source web application.Next, we adapted our machine learning research to develop clinical prognostic models and examine the translational potential of our work during the COVID-19 pandemic. Two prognostic models were constructed, one for evaluating postoperative mortality risk in patients with perioperative SARS-CoV-2 infection and another for assessing postoperative pulmonary complications in adults undergoing elective surgery. In collaboration with the NIHR Global Surgery Research Unit, these models were deployed online as decision support tools.Lastly, we focused on studying complex biological systems by leveraging the UK Biobank data as a rich source of deeply phenotyped information for analyzing ageing trajectories and health-to-disease transitions. By calculating biological age (PhenoAge) over time for nearly 400,000 participants, we identified four distinct ageing trajectories ranging from healthy to unhealthy ageing. Characterization of these ageing phenotypes revealed that factors such as male sex, high BMI, comorbidities, smoking, and shorter parental lifespan were more prevalent in unhealthier categories. We also mapped the disease information for

60 chronic disease categories, finding a predominant cardiometabolic disease signature associated with unhealthy trajectory groups and a healthier trajectory with lower disease prevalence overall.In an effort to uncover the biological underpinnings of these trajectories, we conducted a genome-wide association study (GWAS). The results revealed an immune signature for healthier participants and a metabolic signature driven by lipid biosynthesis and cardiometabolic diseases for unhealthier individuals. Enrichment analysis indicated strong associations with age-related genes in both groups, albeit in opposite directions. Pro-longevity associated variants (e.g., rs429358, rs4420638, rs2075650) were dominant in healthier ageing trajectories, while the SREBF transcription factor was postulated as a key regulatory element driving the unhealthier ageing phenotype.

Appendix

Appendix 1: Web app report of key results

Report downloaded from MachLearnWebApp run as found in (InFlamUOB/MachLearnWebApp).

Machine Learning for Biomarker Discovery

Laura Bravo Merodio

March 2023

Introduction

This is a report that summarizes the main results from the app, including plots and tables. It follows the steps from the analysis.

Data exploration

Your data looks like this:

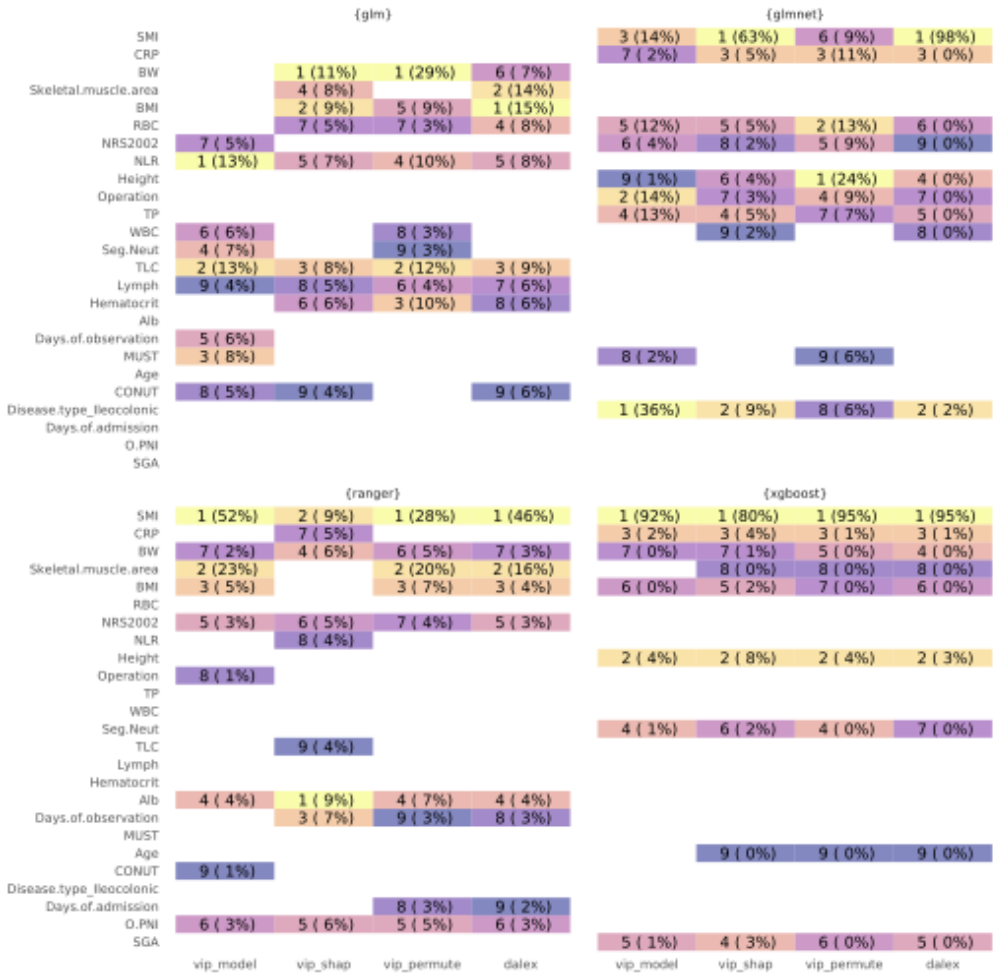
Relationship of Variables and Outcome



The analysis is going to be a ** Classification ** analysis

Feature Selection

Top10- Variable Importance Ranking: classification



The chosen features are:

Selected Features

x
SMI
CRP
BW
Skeletal.muscle.area

Given this ranking table

Selected Features

Variable Name	RankMean	NormMean
SMI	7.312500	0.2925000
CRP	8.875000	0.3550000
BW	9.062500	0.3625000
Skeletal.muscle.area	9.066667	0.3626667
BMI	9.625000	0.3850000
RBC	10.066667	0.4026667
NRS2002	11.133333	0.4453333
NLR	11.400000	0.4560000
Height	12.375000	0.4950000
Operation	12.533333	0.5013333
TP	12.800000	0.5120000
WBC	13.333333	0.5333333
Seg.Neut	13.937500	0.5575000
TLC	14.600000	0.5840000
Lymph	14.733333	0.5893333
Hematocrit	15.000000	0.6000000
Alb	15.933333	0.6373333
Days.of.observation	16.000000	0.6400000
MUST	16.200000	0.6480000
Age	16.333333	0.6533333

Model building

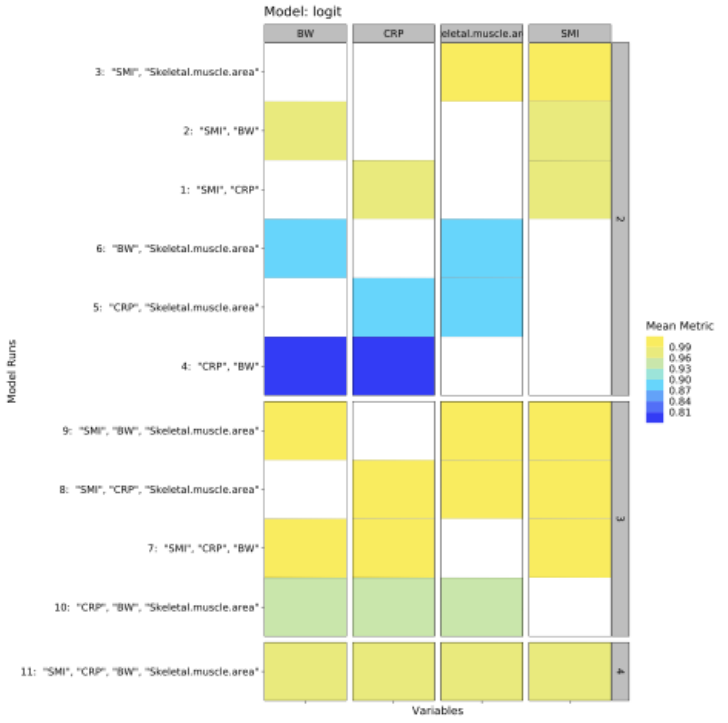
Then model building with the selected parameters: Models: and Bootstraps: 2 of the following models

Variables

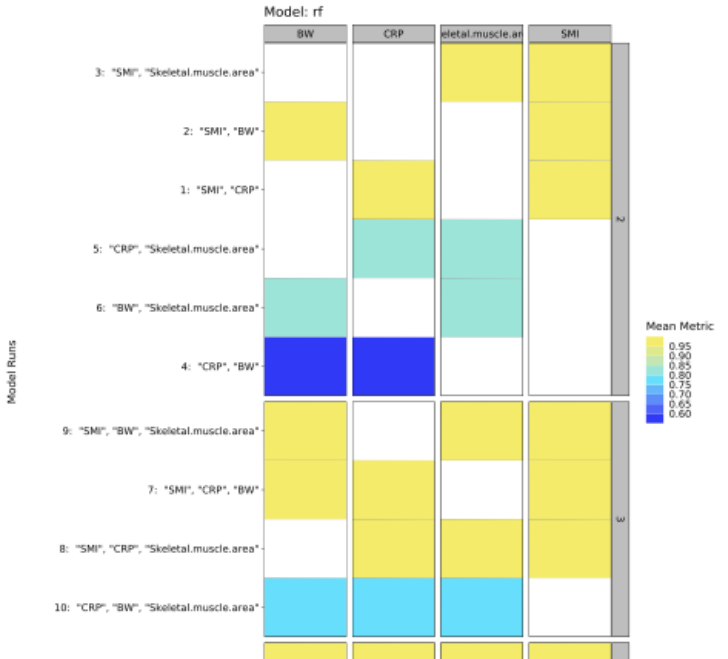
Names	NumVars	Run
"SMI", "CRP"	2	1
"SMI", "BW"	2	2
"SMI", "Skeletal.muscle.area"	2	3
"CRP", "BW"	2	4
"CRP", "Skeletal.muscle.area"	2	5
"BW", "Skeletal.muscle.area"	2	6
"SMI", "CRP", "BW"	3	7
"SMI", "CRP", "Skeletal.muscle.area"	3	8
"SMI", "BW", "Skeletal.muscle.area"	3	9
"CRP", "BW", "Skeletal.muscle.area"	3	10
"SMI", "CRP", "BW", "Skeletal.muscle.area"	4	11

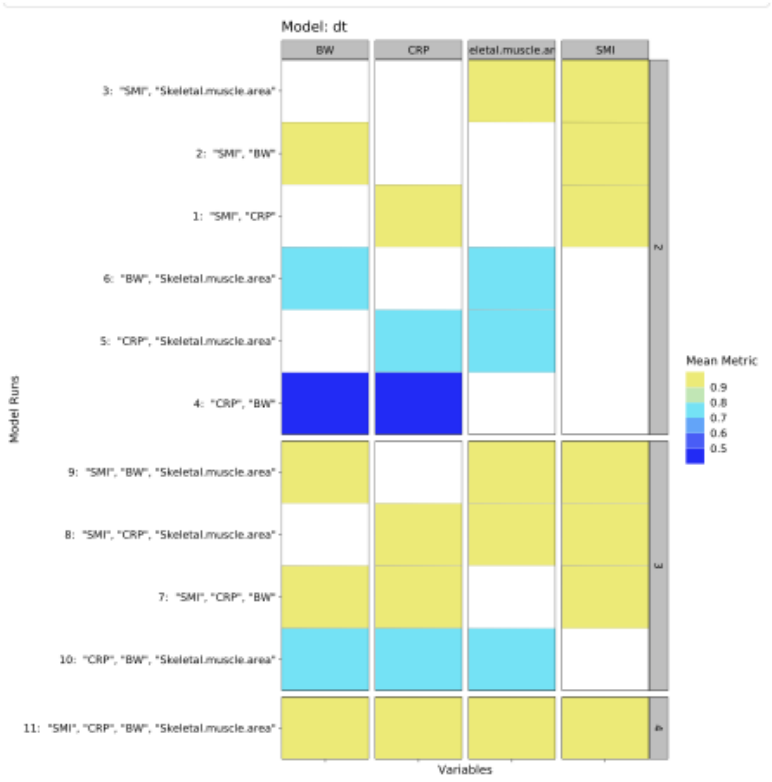
Results

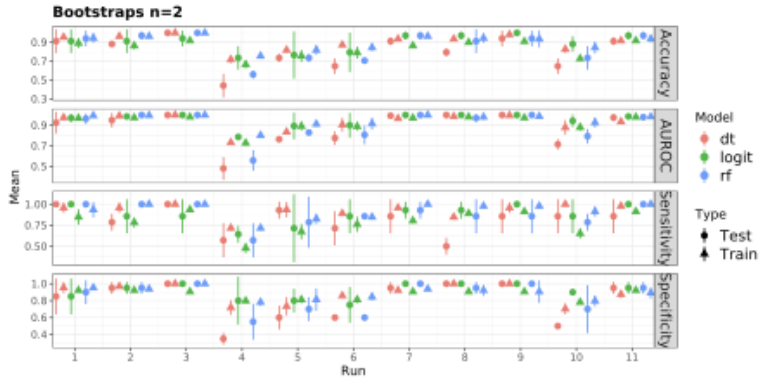
```
## [[1]]
```




```
##
## [[2]]
```







With top selected models being:

Selected Features

Model	Run	Mean	sd	Metrics	Names	NumVars
dt	3	1.000	0.000	AUROC	"SMI", "Skeletal.muscle.area"	2
dt	8	1.000	0.000	AUROC	"SMI", "CRP", "Skeletal.muscle.area"	3
dt	9	1.000	0.000	AUROC	"SMI", "BW", "Skeletal.muscle.area"	3
logit	3	1.000	0.000	AUROC	"SMI", "Skeletal.muscle.area"	2
logit	7	1.000	0.000	AUROC	"SMI", "CRP", "BW"	3
logit	8	1.000	0.000	AUROC	"SMI", "CRP", "Skeletal.muscle.area"	3
logit	9	1.000	0.000	AUROC	"SMI", "BW", "Skeletal.muscle.area"	3
rf	2	1.000	0.000	AUROC	"SMI", "BW"	2
rf	3	1.000	0.000	AUROC	"SMI", "Skeletal.muscle.area"	2
rf	7	1.000	0.000	AUROC	"SMI", "CRP", "BW"	3
rf	9	1.000	0.000	AUROC	"SMI", "BW", "Skeletal.muscle.area"	3
dt	7	0.993	0.010	AUROC	"SMI", "CRP", "BW"	3
logit	2	0.986	0.020	AUROC	"SMI", "BW"	2
logit	11	0.986	0.020	AUROC	"SMI", "CRP", "BW", "Skeletal.muscle.area"	4
rf	11	0.979	0.030	AUROC	"SMI", "CRP", "BW", "Skeletal.muscle.area"	4
dt	11	0.975	0.035	AUROC	"SMI", "CRP", "BW", "Skeletal.muscle.area"	4
logit	1	0.971	0.040	AUROC	"SMI", "CRP"	2
rf	8	0.971	0.040	AUROC	"SMI", "CRP", "Skeletal.muscle.area"	3
rf	1	0.964	0.051	AUROC	"SMI", "CRP"	2
dt	2	0.950	0.071	AUROC	"SMI", "BW"	2
logit	10	0.943	0.061	AUROC	"CRP", "BW", "Skeletal.muscle.area"	3
dt	1	0.925	0.106	AUROC	"SMI", "CRP"	2
logit	6	0.900	0.121	AUROC	"BW", "Skeletal.muscle.area"	2
logit	5	0.893	0.131	AUROC	"CRP", "Skeletal.muscle.area"	2
rf	5	0.829	0.000	AUROC	"CRP", "Skeletal.muscle.area"	2
rf	6	0.807	0.091	AUROC	"BW", "Skeletal.muscle.area"	2
rf	10	0.793	0.071	AUROC	"CRP", "BW", "Skeletal.muscle.area"	3
logit	4	0.786	0.020	AUROC	"CRP", "BW"	2
dt	6	0.775	0.066	AUROC	"BW", "Skeletal.muscle.area"	2
dt	5	0.764	0.020	AUROC	"CRP", "Skeletal.muscle.area"	2
dt	10	0.714	0.051	AUROC	"CRP", "BW", "Skeletal.muscle.area"	3
rf	4	0.557	0.101	AUROC	"CRP", "BW"	2
dt	4	0.479	0.111	AUROC	"CRP", "BW"	2

Appendix 2: Read codes

Data provider	Read code	Value	Term description	Marker
2	2	44TJ.	Blood glucose level	Glucose
4	2	44TJ.	Blood glucose level	Glucose
1	2	44TJ.	Blood glucose level	Glucose
3	3	X772z	Blood glucose level	Glucose
1	2	44U8.	Blood glucose normal	Glucose
2	2	44U8.	Blood glucose normal	Glucose
4	2	44U8.	Blood glucose normal	Glucose
1	2	44U..	Blood glucose result	Glucose
2	2	44U..	Blood glucose result	Glucose
4	2	44U..	Blood glucose result	Glucose
3	3	X772z	BS - Blood glucose level	Glucose
2	2	42H3.	Leucocytosis -high white count	White blood count
1	2	42H3.	Leucocytosis -high white count	White blood count
4	2	42H3.	Leucocytosis -high white count	White blood count
2	2	42H2.	Leucopenia - low white count	White blood count
1	2	42H2.	Leucopenia - low white count	White blood count
4	2	42H2.	Leucopenia - low white count	White blood count
1	2	42A4.	MCV - low	Mean corpuscular volume (MCV)
4	2	42A4.	MCV - low	Mean corpuscular volume (MCV)
2	2	42A4.	MCV - low	Mean corpuscular volume (MCV)
3	3	42A..	MCV - Mean cell volume	Mean corpuscular volume (MCV)
1	2	42A1.	MCV - normal	Mean corpuscular volume (MCV)
4	2	42A1.	MCV - normal	Mean corpuscular volume (MCV)
4	2	42A3.	MCV - raised	Mean corpuscular volume (MCV)
1	2	42A3.	MCV - raised	Mean corpuscular volume (MCV)
2	2	42A3.	MCV - raised	Mean corpuscular volume (MCV)
2	2	42A..	Mean corpuscular volume (MCV)	Mean corpuscular volume (MCV)
1	2	42A..	Mean corpuscular volume (MCV)	Mean corpuscular volume (MCV)
4	2	42A..	Mean corpuscular volume (MCV)	Mean corpuscular volume (MCV)
3	3	XE2mQ	Percentage lymphocyte count	Percentage lymphocytes
1	2	42b1.	Percentage lymphocytes	Percentage lymphocytes
2	2	42b1.	Percentage lymphocytes	Percentage lymphocytes
4	2	42b1.	Percentage lymphocytes	Percentage lymphocytes
3	3	XE2mQ	Percentage lymphocytes	Percentage lymphocytes
1	2	44MI.	Plasma albumin level	Albumin
4	2	44MI.	Plasma albumin level	Albumin
3	3	XaIRc	Plasma albumin level	Albumin
1	2	44CU.	Plasma alkaline phosphatase level	Alkaline Phosphatase

2	2	44CU.	Plasma alkaline phosphatase level	Alkaline Phosphatase
4	2	44CU.	Plasma alkaline phosphatase level	Alkaline Phosphatase
3	3	XaIRj	Plasma alkaline phosphatase level	Alkaline Phosphatase
3	3	XE2dy	Plasma C-reactive protein level	Plasma C-reactive protein level
2	2	44JF.	Plasma creatinine level	Creatinine
4	2	44JF.	Plasma creatinine level	Creatinine
1	2	44JF.	Plasma creatinine level	Creatinine
3	3	XaETQ	Plasma creatinine level	Creatinine
4	2	44TA.	Plasma glucose	Glucose
2	2	44TA.	Plasma glucose	Glucose
1	2	44TA.	Plasma glucose	Glucose
3	3	XM0ly	Plasma glucose	Glucose
1	2	44U..	Plasma glucose level	Glucose
1	2	44g..	Plasma glucose level	Glucose
2	2	44g..	Plasma glucose level	Glucose
4	2	44g..	Plasma glucose level	Glucose
2	2	44U..	Plasma glucose level	Glucose
4	2	44U..	Plasma glucose level	Glucose
3	3	XM0ly	Plasma glucose level	Glucose
2	2	44g0.	Plasma random glucose level	Glucose
4	2	44g0.	Plasma random glucose level	Glucose
1	2	44g0.	Plasma random glucose level	Glucose
3	3	XE2mp	Random blood glucose	Glucose
3	3	XE2mp	Random blood glucose level	Glucose
3	3	XE2mp	RBG - Random blood glucose	Glucose
3	3	XE2mO	RDW - Red blood cell distribution width	Red blood cell dist width
2	2	42Z7.	Red blood cell distribution width	Red blood cell dist width
1	2	42Z7.	Red blood cell distribution width	Red blood cell dist width
4	2	42Z7.	Red blood cell distribution width	Red blood cell dist width
3	3	XE2mO	Red blood cell distribution width	Red blood cell dist width
3	3	42C..	Red blood cell size	Red blood cell dist width
3	3	XE2eA	SA - Serum albumin	Albumin
2	2	44M4.	Serum albumin	Albumin
1	2	44M4.	Serum albumin	Albumin
4	2	44M4.	Serum albumin	Albumin
3	3	XE2eA	Serum albumin	Albumin
3	3	XE2eA	Serum albumin level	Albumin
2	2	44M41	Serum albumin low	Albumin
4	2	44M41	Serum albumin low	Albumin
2	2	44M40	Serum albumin normal	Albumin
4	2	44M40	Serum albumin normal	Albumin
1	2	44M40	Serum albumin normal	Albumin
2	2	44F1.	Serum alk. phos. normal	Alkaline Phosphatase

1	2	44F1.	Serum alk. phos. normal	Alkaline Phosphatase
4	2	44F1.	Serum alk. phos. normal	Alkaline Phosphatase
2	2	44F2.	Serum alk. phos. raised	Alkaline Phosphatase
1	2	44F2.	Serum alk. phos. raised	Alkaline Phosphatase
4	2	44F2.	Serum alk. phos. raised	Alkaline Phosphatase
2	2	44F..	Serum alkaline phosphatase	Alkaline Phosphatase
1	2	44F..	Serum alkaline phosphatase	Alkaline Phosphatase
4	2	44F..	Serum alkaline phosphatase	Alkaline Phosphatase
3	3	XE2px	Serum alkaline phosphatase	Alkaline Phosphatase
3	3	XE2px	Serum alkaline phosphatase level	Alkaline Phosphatase
2	2	44J3.	Serum creatinine	Creatinine
1	2	44J3.	Serum creatinine	Creatinine
4	2	44J3.	Serum creatinine	Creatinine
3	3	XE2q5	Serum creatinine	Creatinine
1	2	44J30	Serum creatinine abnormal	Creatinine
2	2	44J30	Serum creatinine abnormal	Creatinine
4	2	44J30	Serum creatinine abnormal	Creatinine
3	3	XE2q5	Serum creatinine level	Creatinine
2	2	44J31	Serum creatinine low	Creatinine
4	2	44J31	Serum creatinine low	Creatinine
2	2	44J32	Serum creatinine normal	Creatinine
4	2	44J32	Serum creatinine normal	Creatinine
1	2	44J32	Serum creatinine normal	Creatinine
2	2	44J33	Serum creatinine raised	Creatinine
4	2	44J33	Serum creatinine raised	Creatinine
1	2	44J33	Serum creatinine raised	Creatinine
1	2	44f..	Serum glucose level	Glucose
4	2	44f..	Serum glucose level	Glucose
2	2	44f..	Serum glucose level	Glucose
3	3	44f..	Serum glucose level	Glucose
2	2	44f0.	Serum random glucose level	Glucose
1	2	44f0.	Serum random glucose level	Glucose
4	2	44f0.	Serum random glucose level	Glucose
4	2	44F3.	Total alkaline phosphatase	Alkaline Phosphatase
1	2	44F3.	Total alkaline phosphatase	Alkaline Phosphatase
2	2	44F3.	Total alkaline phosphatase	Alkaline Phosphatase
2	2	42H..	White blood count	White blood count
1	2	42H..	White blood count	White blood count
4	2	42H..	White blood count	White blood count

Appendix 3: ICD-10 categorisation

Category	ICD-10 code
ALLERGY	'J301','J302','J303','J304','J450','K522','L20','L200','L208','L209','L23','L230','L231','L232','L233','L234','L235','L236','L237','L238','L239','L500','Z516'
ANEMIA	'D50','D500','D501','D508','D509','D51','D510','D511','D512','D513','D518','D519','D52','D520','D521','D528','D529','D53','D530','D531','D532','D538','D539','D55','D550','D551','D552','D553','D558','D559','D56','D560','D561','D562','D564','D568','D569','D57','D570','D571','D572','D573','D578','D58','D580','D581','D582','D588','D589','D59','D591','D594','D595','D598','D599','D60','D600','D608','D609','D61','D610','D613','D618','D619','D63','D630','D631','D638','D64','D640','D641','D643','D644','D648','D649'
ASTHMA	'J45','J450','J451','J458','J459'
ATRIAL FIBRILLATION	'I48','I480','I481','I482','I483','I484','I489'
AUTOIMMUNE DISEASES	'I731','L10','L100','L101','L102','L103','L104','L108','L109','L12','L120','L121','L122','L123','L128','L129','L40','L400','L401','L402','L403','L404','L405','L408','L409','L41','L410','L411','L412','L413','L414','L415','L418','L419','L93','L930','L931','L932','L94','L940','L941','L942','L943','L944','L945','L946','L948','L949','L95','L950','L951','L958','L959','M30','M300','M301','M302','M303','M308','M31','M310','M311','M312','M313','M314','M315','M316','M317','M318','M319','M32','M321','M328','M329','M3290','M33','M330','M331','M332','M339','M34','M340','M341','M348','M349','M35','M350','M351','M352','M353','M354','M355','M356','M3599','M36','M364','M368'
BLINDNESS, VISUAL IMPAIRMENT	'H54','H540','H541','H542','H544','H545','H546','H547','H549','Z442','Z970'
BLOOD AND BLOOD FORMING ORGAN DISEASES	'D66','D67','D68','D680','D681','D682','D685','D686','D688','D689','D69','D690','D691','D692','D693','D694','D696','D698','D699','D71','D720','D730','D731','D732','D74','D740','D749','D750','D761','D763','D77','D80','D800','D801','D802','D803','D804','D805','D806','D808','D809','D81','D810','D811','D812','D813','D814','D815','D816','D817','D818','D819','D82','D820','D821','D822','D823','D824','D828','D829','D83','D830','D831','D832','D838','D839','D84','D840','D841','D848','D849','D86','D860','D861','D862','D863','D868','D869','D89','D890','D892','D898','D899'
BRADYCARDIAS AND CONDUCTION DISEASES	'I441','I442','I443','I453','I455','Z950'
CARDIAC VALVE DISEASES	'I05','I050','I051','I052','I058','I059','I06','I060','I061','I062','I068','I069','I07','I070','I071','I072','I078','I079','I08','I080','I081','I082','I083','I088','I089','I091','I098','I34','I340','I341','I342','I348','I349','I35','I350','I351','I352','I358','I359','I36','I360','I361','I362','I368','I369','I37','I370','I371','I372','I378','I379','I38','I390','I391','I392','I393','I394','Q22','Q220','Q221','Q222','Q223','Q224','Q225','Q226','Q228','Q229','Q23','Q230','Q231','Q232','Q233','Q234','Q238','Q239','Z952','Z953','Z954'
CATARACT AND OTHER LENS DISEASES	'H25','H250','H251','H252','H258','H259','H26','H260','H261','H262','H263','H264','H268','H269','H27','H270','H271','H278','H279','H28','H280','H281','H282','H288','Q12','Q120','Q121','Q122','Q123','Q124','Q128','Q129','Z961'
CEREBROVASCULAR DISEASE	'G45','G450','G451','G452','G453','G454','G458','G459','G46','G460','G461','G462','G463','G464','G465','G466','G467','G468','I60','I600','I601','I602','I603','I604','I605','I606','I607','I608','I609','I61','I610','I611','I612','I613','I614','I615','I616','I618','I619','I62','I620','I621','I629','I63','I630','I631','I632','I633','I634','I635','I636','I638','I639','I64','I67','I670','I671','I672','I673','I674','I675','I676','I677','I678','I679','I69','I690','I691','I692','I693','I694','I698'
CHROMOSOMAL ABNORMALITIES	'Q90','Q900','Q901','Q902','Q909','Q91','Q910','Q911','Q912','Q913','Q914','Q915','Q916','Q917','Q92','Q920','Q921','Q922','Q923','Q924','Q925','Q926','Q927','Q928','Q929','Q93','Q930','Q931','Q932','Q933','Q934','Q935','Q936','Q937','Q938','Q939','Q95','Q950','Q951','Q952','Q953','Q954','Q955','Q958','Q959','Q96','Q960','Q961','Q962','Q963','Q964','Q968','Q969','Q97','Q970','Q971','Q972','Q973','Q978','Q979','Q98','Q980','Q981','Q982','Q983','Q984','Q985','Q986','Q987','Q988','Q989','Q99','Q990','Q991','Q992','Q998','Q999'
CHRONIC INFECTIOUS DISEASES	'A15','A150','A151','A152','A153','A154','A155','A156','A157','A158','A159','A16','A160','A161','A162','A163','A164','A165','A167','A168','A169','A17','A170','A171','A178','A179','A18','A180','A181','A182','A183','A184','A185','A186','A187','A188','A19','A190','A191','A192','A198','A199','A30','A300','A301','A302','A303','A304','A305','A308','A309','A31','A310','A311','A318','A319','A50','A500','A501','A502','A503','A504','A505','A506','A507','A509','A52','A520','A521','A522','A523','A527','A528','A529','A53','A530','A539','A65','A66','A660','A661','A662','A663','A664','A665','A666','A667','A668','A669','A67','A670','A671','A672','A673','A679','A692','A81','A810','A811','A812','A818','A819','B20','B200','B201','B202','B203','B204','B205','B206','B207','B208','B209','B21','B210','B211','B212','B213','B217','B218','B219','B22','B220','B221','B222','B227','B23','B230','B231','B232','B238','B24','B381','B391','B401','B572','B573','B574'

	,'B575','B65','B650','B651','B652','B653','B658','B659','B92','B94','B940','B941','B942','B948','B949','J65','M863','M864','M865','M866'
CHRONIC KIDNEY DISEASES	'I120','I130','I131','I132','I139','N01','N010','N011','N012','N013','N014','N015','N016','N017','N018','N019','N03','N030','N031','N032','N033','N034','N035','N036','N037','N038','N039','N04','N040','N041','N042','N043','N044','N045','N046','N047','N048','N049','N05','N050','N051','N052','N053','N054','N055','N056','N057','N058','N059','N07','N070','N071','N072','N073','N074','N075','N076','N077','N078','N079','N08','N080','N081','N082','N083','N084','N085','N088','N11','N110','N111','N118','N119','N183','N184','N185','N189','Q60','Q600','Q601','Q602','Q603','Q604','Q605','Q606','Q611','Q612','Q613','Q614','Q615','Q618','Q619','Z905','Z940'
CHRONIC LIVER DISEASES	'B18','B180','B181','B182','B188','B189','K70','K702','K703','K704','K709','K713','K714','K715','K717','K721','K73','K730','K731','K732','K738','K739','K74','K740','K741','K742','K743','K744','K745','K746','K753','K754','K758','K761','K766','K767','K778','Q446','Z944'
CHRONIC PANCREAS, BILIARY TRACT AND GALLBLADDER DISEASES	'K800','K801','K802','K808','K811','K86','K860','K861','K868','Q440','Q441','Q442','Q443','Q444','Q445','Q450'
CHRONIC ULCER OF THE SKIN	'I830','I832','L89','L890','L891','L892','L893','L899','L97','L984'
COLITIS AND RELATED DISEASES	'K520','K528','K551','K552','K572','K573','K574','K575','K578','K579','K58','K580','K589','K590','K592','K62','K622','K623','K624','K627','K628','K629','K634','K64','K640','K641','K642','K643','K644','K648','K649'
COPD, EMPHYSEMA, CHRONIC BRONCHITIS	'J41','J410','J411','J418','J42','J43','J430','J431','J432','J438','J439','J44','J440','J441','J448','J449','J47'
DEAFNESS, HEARING IMPAIRMENT	'H80','H800','H801','H802','H808','H809','H90','H900','H901','H902','H903','H904','H905','H906','H907','H908','H911','H913','H919','Q16','Q160','Q161','Q162','Q163','Q164','Q165','Q169','Z453','Z461','Z962','Z974'
DEMENTIA	'F00','F000','F001','F002','F009','F01','F010','F011','F012','F013','F018','F019','F02','F020','F021','F022','F023','F024','F028','F03','F051','G30','G300','G301','G308','G309','G31','G310','G311','G312','G318','G319'
DEPRESSION AND MOOD DISEASES	'F30','F300','F301','F302','F308','F309','F31','F310','F311','F312','F313','F314','F315','F316','F317','F318','F319','F32','F320','F321','F322','F323','F328','F329','F33','F330','F331','F332','F333','F334','F338','F339','F34','F340','F341','F348','F349','F38','F380','F381','F388','F39','F412'
DIABETES	'E10','E100','E101','E102','E103','E104','E105','E106','E107','E108','E109','E11','E110','E111','E112','E113','E114','E115','E116','E117','E118','E119','E13','E130','E131','E132','E133','E134','E135','E136','E137','E138','E139','E14','E140','E141','E142','E143','E144','E145','E146','E147','E148','E149','E891'
DORSOPATHIES	'M40','M400','M4000','M4001','M4002','M4003','M4004','M4005','M4006','M4007','M4008','M4009','M401','M4010','M4011','M4012','M4013','M4014','M4015','M4016','M4017','M4018','M4019','M402','M4020','M4021','M4022','M4023','M4024','M4025','M4026','M4027','M4028','M4029','M403','M4030','M4031','M4032','M4033','M4034','M4035','M4036','M4037','M4038','M4039','M404','M4040','M4041','M4042','M4043','M4044','M4045','M4046','M4047','M4048','M4049','M405','M4050','M4051','M4052','M4053','M4054','M4055','M4056','M4057','M4058','M4059','M41','M410','M4100','M4101','M4102','M4103','M4104','M4105','M4106','M4107','M4108','M4109','M411','M4110','M4111','M4112','M4113','M4114','M4115','M4116','M4117','M4118','M4119','M412','M4120','M4121','M4122','M4123','M4124','M4125','M4126','M4127','M4128','M4129','M413','M4130','M4131','M4132','M4133','M4134','M4135','M4136','M4137','M4138','M4139','M414','M4140','M4141','M4142','M4143','M4144','M4145','M4146','M4147','M4148','M4149','M415','M4150','M4151','M4152','M4153','M4154','M4155','M4156','M4157','M4158','M4159','M418','M4180','M4181','M4182','M4183','M4184','M4185','M4186','M4187','M4188','M4189','M419','M4190','M4191','M4192','M4193','M4194','M4195','M4196','M4197','M4198','M4199','M42','M420','M4200','M4201','M4202','M4203','M4204','M4205','M4206','M4207','M4208','M4209','M421','M4210','M4211','M4212','M4213','M4214','M4215','M4216','M4217','M4218','M4219','M429','M4290','M4291','M4292','M4293','M4294','M4295','M4296','M4297','M4298','M4299','M43','M430','M4300','M4301','M4302','M4303','M4304','M4305','M4306','M4307','M4308','M4309','M431','M4310','M4311','M4312','M4313','M4314','M4315','M4316','M4317','M4318','M4319','M432','M4320','M4321','M4322','M4323','M4324','M4325','M4326','M4327','M4328','M4329','M433','M4330','M4331','M4332','M4333','M4334','M4335','M4336','M4337','M4338','M4339','M434','M4340','M4341','M4342','M4343','M4344','M4345','M4346','M4347','M4348','M4349','M435','M4350','M4351','M4352','M4353','M4354','M4355','M4356','M4357','M4358','M4359','M436','M4360','M4361','M4362','M4363','M4364','M4365','M4366','M4367','M4368','M4369','M438','M4380','M4381','M4382','M4383','M4384','M4385','M4386','M4387','M4388','M4389','M439','M4390','M4391','M4392','M4393','M4394','M4395','M4396','M4397','M4398','M4399','M47','M470','M4700','M4701','M4702','M4703','M4704','M4705','M4706','M4707','M4708','M4709','M471','M4710','M4711','M4712','M4713','M4714','M4715','M4716','M4717','M4718','M4719','M472','M4720','M4721','M4722','M4723','M4724','M4725','M4726','

	M4727','M4728','M4729','M478','M4780','M4781','M4782','M4783','M4784','M4785','M4786','M4787','M4788','M4789','M479','M4790','M4791','M4792','M4793','M4794','M4795','M4796','M4797','M4798','M4799','M48','M480','M4800','M4801','M4802','M4803','M4804','M4805','M4806','M4807','M4808','M4809','M481','M4810','M4811','M4812','M4813','M4814','M4815','M4816','M4817','M4818','M4819','M482','M4820','M4821','M4822','M4823','M4824','M4825','M4826','M4827','M4828','M4829','M483','M4830','M4831','M4832','M4833','M4834','M4835','M4836','M4837','M4838','M4839','M484','M4840','M4841','M4842','M4843','M4844','M4845','M4846','M4847','M4848','M4849','M485','M4850','M4851','M4852','M4853','M4854','M4855','M4856','M4857','M4858','M4859','M488','M4880','M4881','M4882','M4883','M4884','M4885','M4886','M4887','M4888','M4889','M489','M4890','M4891','M4892','M4893','M4894','M4895','M4896','M4897','M4898','M4899','M49','M490','M4900','M4901','M4902','M4903','M4904','M4905','M4906','M4907','M4908','M4909','M491','M4910','M4911','M4912','M4913','M4914','M4915','M4916','M4917','M4918','M4919','M492','M4920','M4921','M4922','M4923','M4924','M4925','M4926','M4927','M4928','M4929','M493','M4930','M4931','M4932','M4933','M4934','M4935','M4936','M4937','M4938','M4939','M494','M4940','M4941','M4942','M4943','M4944','M4945','M4946','M4947','M4948','M4949','M495','M4950','M4951','M4952','M4953','M4954','M4955','M4956','M4957','M4958','M4959','M498','M4980','M4981','M4982','M4983','M4984','M4985','M4986','M4987','M4988','M4989','M50','M500','M501','M502','M503','M508','M509','M51','M510','M511','M5116','M5119','M512','M513','M5137','M514','M518','M519','M53','M530','M5300','M5301','M5302','M5303','M5304','M5305','M5306','M5307','M5308','M5309','M531','M5310','M5311','M5312','M5313','M5314','M5315','M5316','M5317','M5318','M5319','M532','M5320','M5321','M5322','M5323','M5324','M5325','M5326','M5327','M5328','M5329','M533','M5330','M5331','M5332','M5333','M5334','M5335','M5336','M5337','M5338','M5339','M538','M5380','M5381','M5382','M5383','M5384','M5385','M5386','M5387','M5388','M5389','M539','M5390','M5391','M5392','M5393','M5394','M5395','M5396','M5397','M5398','M5399','Q675','Q761','Q764'
DYSLIPIDEMIA	'E78','E780','E781','E782','E783','E784','E785','E786','E788','E789'
EAR, NOSE, THROAT DISEASES	'H604','H661','H662','H663','H701','H71','H731','H741','H810','H831','H832','H95','H950','H951','H958','H959','J300','J31','J310','J311','J312','J32','J320','J321','J322','J323','J324','J328','J329','J33','J330','J331','J338','J339','J341','J342','J343','J35','J350','J351','J352','J353','J358','J359','J37','J370','J371','J380','J386','K051','K053','K07','K070','K071','K072','K073','K074','K075','K076','K078','K079','K110','K117','Q30','Q300','Q301','Q302','Q303','Q308','Q309','Q31','Q310','Q311','Q312','Q313','Q314','Q315','Q318','Q319','Q32','Q320','Q321','Q322','Q323','Q324','Q35','Q350','Q351','Q352','Q353','Q354','Q355','Q356','Q357','Q358','Q359','Q36','Q360','Q361','Q369','Q37','Q370','Q371','Q372','Q373','Q374','Q375','Q378','Q379','Q38','Q380','Q381','Q382','Q383','Q384','Q385','Q386','Q387','Q388'
EPILEPSY	'G40','G400','G401','G402','G403','G404','G406','G407','G408','G409'
ESOPHAGUS, STOMACH AND DUODENUM DISEASES	'I85','I850','I859','I864','I982','I983','K21','K210','K219','K220','K222','K224','K225','K227','K230','K231','K254','K255','K256','K257','K264','K265','K266','K267','K274','K275','K276','K277','K284','K285','K286','K287','K293','K294','K295','K296','K297','K298','K299','K311','K312','K313','K314','K315','Q39','Q390','Q391','Q392','Q393','Q394','Q395','Q396','Q398','Q399','Q40','Q400','Q401','Q402','Q403','Q408','Q409','Z903'
GLAUCOMA	'H401','H402','H403','H404','H405','H406','H408','H409'
HEART FAILURE	'I110','I130','I132','I27','I270','I271','I272','I278','I279','I280','I42','I420','I421','I422','I423','I424','I425','I426','I427','I428','I429','I43','I430','I431','I432','I438','I50','I500','I501','I509','I515','I517','I528','Z941','Z943'
HEMATOLOGICAL NEOPLASMS	'C81','C810','C811','C812','C813','C814','C817','C819','C82','C820','C821','C822','C823','C824','C825','C826','C827','C829','C83','C830','C831','C832','C833','C834','C835','C836','C837','C838','C839','C84','C840','C841','C842','C843','C844','C845','C846','C847','C848','C849','C85','C850','C851','C852','C857','C859','C86','C860','C861','C862','C863','C864','C865','C866','C88','C880','C881','C882','C883','C884','C887','C889','C90','C900','C901','C902','C903','C91','C910','C911','C912','C913','C914','C915','C916','C917','C918','C919','C92','C920','C921','C922','C923','C924','C925','C926','C927','C928','C929','C93','C930','C931','C932','C933','C937','C939','C94','C940','C941','C942','C943','C944','C945','C946','C947','C95','C950','C951','C952','C957','C959','C96','C960','C961','C962','C963','C964','C965','C966','C967','C968','C969'
HYPERTENSION	'I10','I11','I110','I119','I12','I120','I129','I13','I130','I131','I132','I139','I15','I150','I151','I152','I158','I159'
INFLAMMATORY ARTHROPATHIES	'M023','M05','M050','M0500','M0501','M0502','M0503','M0504','M0505','M0506','M0507','M0508','M0509','M051','M0510','M0511','M0512','M0513','M0514','M0515','M0516','M0517','M0518','M0519','M052','M0520','M0521','M0522','M0523','M0524','M0525','M0526','M0527','M0528','M0529','M053','M0530','M0531','M0532','M0533','M0534','M0535','M0536','M0537','M0538','M0539','M058','M0580','M0581','M0582','M0583','M0584','M0585','M0586','M0587','M0588','M0589','M059','M0590','M0591','M0592','M0593','M0594','M0595','M0596','M0597','M0598','M0599','M06','M060','M060

	0','M0601','M0602','M0603','M0604','M0605','M0606','M0607','M0608','M0609','M061','M0610','M0611','M0612','M0613','M0614','M0615','M0616','M0617','M0618','M0619','M062','M0620','M0621','M0622','M0623','M0624','M0625','M0626','M0627','M0628','M0629','M063','M0630','M0631','M0632','M0633','M0634','M0635','M0636','M0637','M0638','M0639','M064','M0640','M0641','M0642','M0643','M0644','M0645','M0646','M0647','M0648','M0649','M068','M0680','M0681','M0682','M0683','M0684','M0685','M0686','M0687','M0688','M0689','M069','M0690','M0691','M0692','M0693','M0694','M0695','M0696','M0697','M0698','M0699','M07','M070','M0700','M0701','M0702','M0703','M0704','M0705','M0706','M0707','M0708','M0709','M071','M0710','M0711','M0712','M0713','M0714','M0715','M0716','M0717','M0718','M0719','M072','M0720','M0721','M0722','M0723','M0724','M0725','M0726','M0727','M0728','M0729','M073','M0730','M0731','M0732','M0733','M0734','M0735','M0736','M0737','M0738','M0739','M074','M0740','M0741','M0742','M0743','M0744','M0745','M0746','M0747','M0748','M0749','M075','M0750','M0751','M0752','M0753','M0754','M0755','M0756','M0757','M0758','M0759','M076','M0760','M0761','M0762','M0763','M0764','M0765','M0766','M0767','M0768','M0769','M08','M080','M0800','M0801','M0802','M0803','M0804','M0805','M0806','M0807','M0808','M0809','M081','M0810','M0811','M0812','M0813','M0814','M0815','M0816','M0817','M0818','M0819','M082','M0820','M0821','M0822','M0823','M0824','M0825','M0826','M0827','M0828','M0829','M083','M0830','M0831','M0832','M0833','M0834','M0835','M0836','M0837','M0838','M0839','M084','M0840','M0841','M0842','M0843','M0844','M0845','M0846','M0847','M0848','M0849','M088','M0880','M0881','M0882','M0883','M0884','M0885','M0886','M0887','M0888','M0889','M089','M0890','M0891','M0892','M0893','M0894','M0895','M0896','M0897','M0898','M0899','M09','M090','M0900','M0901','M0902','M0903','M0904','M0905','M0906','M0907','M0908','M0909','M091','M0910','M0911','M0912','M0913','M0914','M0915','M0916','M0917','M0918','M0919','M092','M0920','M0921','M0922','M0923','M0924','M0925','M0926','M0927','M0928','M0929','M098','M0980','M0981','M0982','M0983','M0984','M0985','M0986','M0987','M0988','M0989','M10','M100','M1000','M1001','M1002','M1003','M1004','M1005','M1006','M1007','M1008','M1009','M101','M1010','M1011','M1012','M1013','M1014','M1015','M1016','M1017','M1018','M1019','M102','M1020','M1021','M1022','M1023','M1024','M1025','M1026','M1027','M1028','M1029','M103','M1030','M1031','M1032','M1033','M1034','M1035','M1036','M1037','M1038','M1039','M104','M1040','M1041','M1042','M1043','M1044','M1045','M1046','M1047','M1048','M1049','M109','M1090','M1091','M1092','M1093','M1094','M1095','M1096','M1097','M1098','M1099','M11','M110','M1100','M1101','M1102','M1103','M1104','M1105','M1106','M1107','M1108','M1109','M111','M1110','M1111','M1112','M1113','M1114','M1115','M1116','M1117','M1118','M1119','M112','M1120','M1121','M1122','M1123','M1124','M1125','M1126','M1127','M1128','M1129','M118','M1180','M1181','M1182','M1183','M1184','M1185','M1186','M1187','M1188','M1189','M119','M1190','M1191','M1192','M1193','M1194','M1195','M1196','M1197','M1198','M1199','M12','M120','M1200','M1201','M1202','M1203','M1204','M1205','M1206','M1207','M1208','M1209','M121','M1210','M1211','M1212','M1213','M1214','M1215','M1216','M1217','M1218','M1219','M122','M1220','M1221','M1222','M1223','M1224','M1225','M1226','M1227','M1228','M1229','M123','M1230','M1231','M1232','M1233','M1234','M1235','M1236','M1237','M1238','M1239','M124','M1240','M1241','M1242','M1243','M1244','M1245','M1246','M1247','M1248','M1249','M125','M1250','M1251','M1252','M1253','M1254','M1255','M1256','M1257','M1258','M1259','M128','M1280','M1281','M1282','M1283','M1284','M1285','M1286','M1287','M1288','M1289','M13','M130','M1300','M1301','M1302','M1303','M1304','M1305','M1306','M1307','M1308','M1309','M131','M1310','M1311','M1312','M1313','M1314','M1315','M1316','M1317','M1318','M1319','M138','M1380','M1381','M1382','M1383','M1384','M1385','M1386','M1387','M1388','M1389','M139','M1390','M1391','M1392','M1393','M1394','M1395','M1396','M1397','M1398','M1399','M14','M140','M141','M142','M143','M144','M145','M146','M148','M45','M45X0','M45X1','M45X2','M45X3','M45X4','M45X5','M45X6','M45X7','M45X8','M45X9','M460','M461','M468','M469'
INFLAMMATORY BOWEL DISEASES	'K50','K500','K501','K508','K509','K51','K510','K511','K512','K513','K514','K515','K518','K519'
ISCHEMIC HEART DISEASE	'I20','I200','I201','I208','I209','I21','I210','I211','I212','I213','I214','I219','I21X','I22','I220','I221','I228','I229','I24','I240','I241','I248','I249','I25','I250','I251','I252','I253','I254','I255','I256','I258','I259','Z951','Z955'
MIGRAINE AND FACIAL PAIN SYNDROMES	'G43','G430','G431','G432','G433','G438','G439','G440','G441','G442','G443','G448','G50','G500','G501','G508','G509'
MULTIPLE SCLEROSIS	'G35'
NEUROTIC, STRESS-RELATED AND SOMATOFORM DISEASES	'F40','F400','F401','F402','F408','F409','F41','F410','F411','F412','F413','F418','F419','F42','F420','F421','F422','F428','F429','F43','F431','F438','F439','F44','F440','F441','F442','F443','F444','F445','F446','F447','F448','F449','F45','F450','F451','F452','F453','F454','F458','F459','F48','F480','F481','F488','F489'
OBESITY	'E66','E660','E661','E662','E668','E669'

OSTEOARTHRITIS AND OTHER DEGENERATIVE JOINT DISEASES	'M15','M150','M1500','M151','M152','M153','M154','M158','M159','M1599','M16','M160','M161','M162','M163','M164','M165','M166','M167','M169','M17','M170','M171','M172','M173','M174','M175','M179','M18','M180','M181','M182','M183','M184','M185','M189','M19','M190','M1900','M1901','M1902','M1903','M1904','M1905','M1906','M1907','M1908','M1909','M191','M1910','M1911','M1912','M1913','M1914','M1915','M1916','M1917','M1918','M1919','M192','M1920','M1921','M1922','M1923','M1924','M1925','M1926','M1927','M1928','M1929','M198','M1980','M1981','M1982','M1983','M1984','M1985','M1986','M1987','M1988','M1989','M199','M1990','M1991','M1992','M1993','M1994','M1995','M1996','M1997','M1998','M1999','M362','M363'
OSTEOPOROSIS	'M80','M800','M8000','M8001','M8002','M8003','M8004','M8005','M8006','M8007','M8008','M8009','M801','M8010','M8011','M8012','M8013','M8014','M8015','M8016','M8017','M8018','M8019','M802','M8020','M8021','M8022','M8023','M8024','M8025','M8026','M8027','M8028','M8029','M803','M8030','M8031','M8032','M8033','M8034','M8035','M8036','M8037','M8038','M8039','M804','M8040','M8041','M8042','M8043','M8044','M8045','M8046','M8047','M8048','M8049','M805','M8050','M8051','M8052','M8053','M8054','M8055','M8056','M8057','M8058','M8059','M808','M8080','M8081','M8082','M8083','M8084','M8085','M8086','M8087','M8088','M8089','M809','M8090','M8091','M8092','M8093','M8094','M8095','M8096','M8097','M8098','M8099','M81','M810','M8100','M8101','M8102','M8103','M8104','M8105','M8106','M8107','M8108','M8109','M811','M8110','M8111','M8112','M8113','M8114','M8115','M8116','M8117','M8118','M8119','M812','M8120','M8121','M8122','M8123','M8124','M8125','M8126','M8127','M8128','M8129','M813','M8130','M8131','M8132','M8133','M8134','M8135','M8136','M8137','M8138','M8139','M814','M8140','M8141','M8142','M8143','M8144','M8145','M8146','M8147','M8148','M8149','M815','M8150','M8151','M8152','M8153','M8154','M8155','M8156','M8157','M8158','M8159','M816','M8160','M8161','M8162','M8163','M8164','M8165','M8166','M8167','M8168','M8169','M818','M8180','M8181','M8182','M8183','M8184','M8185','M8186','M8187','M8188','M8189','M819','M8190','M8191','M8192','M8193','M8194','M8195','M8196','M8197','M8198','M8199','M82','M820','M8200','M8201','M8202','M8203','M8204','M8205','M8206','M8207','M8208','M8209','M821','M8210','M8211','M8212','M8213','M8214','M8215','M8216','M8217','M8218','M8219','M823','M828','M8280','M8281','M8282','M8283','M8284','M8285','M8286','M8287','M8288','M8289'
OTHER CARDIOVASCULAR DISEASES	'I09','I090','I092','I099','I281','I310','I311','I456','I495','I498','I70','I700','I7000','I7001','I701','I7010','I7011','I7020','I7021','I708','I7080','I7081','I709','I7090','I7091','I71','I710','I711','I712','I713','I714','I715','I716','I718','I719','I72','I720','I721','I722','I723','I724','I725','I726','I728','I729','I790','I791','I950','I951','I958','Q20','Q200','Q201','Q202','Q203','Q204','Q205','Q206','Q208','Q209','Q21','Q210','Q211','Q212','Q213','Q214','Q218','Q219','Q24','Q240','Q241','Q242','Q243','Q244','Q245','Q246','Q248','Q249','Q25','Q250','Q251','Q252','Q253','Q254','Q255','Q256','Q257','Q258','Q259','Q26','Q260','Q261','Q262','Q263','Q264','Q265','Q266','Q268','Q269','Q27','Q270','Q271','Q272','Q273','Q274','Q278','Q279','Q28','Q280','Q281','Q282','Q283','Q288','Q289','Z958','Z959'
OTHER DIGESTIVE DISEASES	'K660','K900','K901','K902','K911','K93','K930','K931','K938','Q41','Q410','Q411','Q412','Q418','Q419','Q42','Q420','Q421','Q422','Q423','Q428','Q429','Q43','Q430','Q431','Q432','Q433','Q434','Q435','Q436','Q437','Q438','Q439','R15','Z904','Z980'
OTHER EYE DISEASES	'H022','H023','H024','H025','H04','H040','H041','H042','H044','H045','H046','H048','H049','H05','H051','H052','H053','H054','H055','H058','H059','H104','H17','H170','H171','H178','H179','H184','H185','H186','H187','H188','H189','H193','H198','H201','H21','H210','H211','H212','H213','H214','H215','H218','H219','H310','H311','H312','H318','H319','H33','H330','H331','H332','H333','H334','H335','H352','H353','H354','H355','H357','H358','H359','H36','H360','H368','H47','H472','H473','H474','H475','H476','H477','H48','H480','H488','H49','H490','H491','H492','H493','H494','H498','H499','H51','H510','H511','H512','H518','H519','Q10','Q100','Q101','Q102','Q103','Q104','Q105','Q106','Q107','Q11','Q110','Q111','Q112','Q113','Q13','Q130','Q131','Q132','Q133','Q134','Q135','Q138','Q139','Q14','Q140','Q141','Q142','Q143','Q148','Q149','Q15','Q150','Q158','Q159','Z947'
OTHER GENITOURINARY DISEASES	'B901','N200','N202','N209','N210','N218','N219','N22','N220','N228','N301','N302','N303','N304','N31','N310','N311','N312','N318','N319','N320','N323','N328','N329','N33','N330','N338','N35','N350','N351','N358','N359','N393','N394','N480','N484','N489','N701','N711','N731','N734','N736','N761','N763','N81','N810','N811','N812','N813','N814','N815','N816','N818','N819','N88','N880','N881','N882','N883','N884','N888','N889','N895','N905','N952','Q54','Q540','Q541','Q542','Q543','Q544','Q548','Q549','Q620','Q621','Q622','Q623','Q624','Q627','Q628','Q638','Q639','Q640','Q641','Q643','Q644','Q645','Q646','Q647','Q648','Q649','Z906','Z907','Z960'
OTHER METABOLIC DISEASES	'E20','E200','E201','E208','E209','E21','E210','E211','E212','E213','E214','E215','E22','E220','E221','E222','E228','E229','E23','E230','E232','E233','E236','E237','E24','E240','E241','E243','E248','E249','E25','E250','E258','E259','E26','E260','E261','E268','E269','E27','E270','E271','E272','E274','E275','E278','E279','E28','E280','E281','E282','E283','E288','E289','E29','E290','E291','E298','E299','E31','E310','E311','E318','E319','E34','E340','E341','E342','E345','E348','E349','E35','E351','E358','E40','E41','E

	42','E43','E44','E440','E45','E46','E64','E640','E641','E642','E643','E648','E649','E70','E700','E701','E702','E703','E708','E709','E71','E710','E711','E712','E713','E72','E720','E721','E722','E723','E724','E725','E728','E729','E74','E740','E741','E742','E743','E744','E748','E749','E75','E750','E751','E752','E753','E754','E755','E756','E76','E760','E761','E762','E763','E768','E769','E77','E770','E771','E778','E779','E79','E791','E798','E799','E80','E800','E801','E802','E803','E805','E806','E807','E83','E830','E831','E832','E833','E834','E835','E838','E839','E84','E840','E841','E848','E849','E85','E850','E851','E852','E853','E854','E858','E859','E88','E880','E881','E882','E888','E889','E89','E891','E893','E894','E895','E896','E898','E899','K903','K904','K908','K909','K912','M83','M830','M8300','M8301','M8302','M8303','M8304','M8305','M8306','M8307','M8308','M8309','M831','M8310','M8311','M8312','M8313','M8314','M8315','M8316','M8317','M8318','M8319','M832','M8320','M8321','M8322','M8323','M8324','M8325','M8326','M8327','M8328','M8329','M833','M8330','M8331','M8332','M8333','M8334','M8335','M8336','M8337','M8338','M8339','M834','M8340','M8341','M8342','M8343','M8344','M8345','M8346','M8347','M8348','M8349','M835','M8350','M8351','M8352','M8353','M8354','M8355','M8356','M8357','M8358','M8359','M838','M8380','M8381','M8382','M8383','M8384','M8385','M8386','M8387','M8388','M8389','M839','M8390','M8391','M8392','M8393','M8394','M8395','M8396','M8397','M8398','M8399','M88','M880','M8800','M8801','M8802','M8803','M8804','M8805','M8806','M8807','M8808','M8809','M888','M8880','M8881','M8882','M8883','M8884','M8885','M8886','M8887','M8888','M8889','M889','M8890','M8891','M8892','M8893','M8894','M8895','M8896','M8897','M8898','M8899','N25','N250','N251','N258','N259'
OTHER MUSCULOSKELETAL AND JOINT DISEASES	'B902','M212','M213','M214','M215','M216','M217','M218','M219','M22','M220','M221','M2210','M222','M223','M224','M228','M229','M23','M230','M2300','M2301','M2302','M2303','M2304','M2305','M2306','M2307','M2309','M231','M2310','M2311','M2312','M2313','M2314','M2315','M2316','M2317','M2319','M232','M2320','M2321','M2322','M2323','M2324','M2325','M2326','M2327','M2329','M233','M2330','M2331','M2332','M2333','M2334','M2335','M2336','M2337','M2339','M234','M2340','M2341','M2342','M2343','M2344','M2345','M2346','M2347','M2349','M235','M2350','M2351','M2352','M2353','M2354','M2355','M2356','M2357','M2359','M236','M2360','M2361','M2362','M2363','M2364','M2365','M2366','M2367','M2369','M238','M2380','M2381','M2382','M2383','M2384','M2385','M2386','M2387','M2389','M239','M2390','M2391','M2392','M2393','M2394','M2395','M2396','M2397','M2399','M24','M240','M2400','M2401','M2402','M2403','M2404','M2405','M2406','M2407','M2408','M2409','M241','M2410','M2411','M2412','M2413','M2414','M2415','M2416','M2417','M2418','M2419','M242','M2420','M2421','M2422','M2423','M2424','M2425','M2426','M2427','M2428','M2429','M243','M2430','M2431','M2432','M2433','M2434','M2435','M2436','M2437','M2438','M2439','M244','M2440','M2441','M2442','M2443','M2444','M2445','M2446','M2447','M2448','M2449','M245','M2450','M2451','M2452','M2453','M2454','M2455','M2456','M2457','M2458','M2459','M246','M2460','M2461','M2462','M2463','M2464','M2465','M2466','M2467','M2468','M2469','M247','M2470','M2471','M2472','M2473','M2474','M2475','M2476','M2477','M2478','M2479','M248','M2480','M2481','M2482','M2483','M2484','M2485','M2486','M2487','M2488','M2489','M249','M2490','M2491','M2492','M2493','M2494','M2495','M2496','M2497','M2498','M2499','M252','M253','M357','M61','M610','M6100','M6101','M6102','M6103','M6104','M6105','M6106','M6107','M6108','M6109','M611','M6110','M6111','M6112','M6113','M6114','M6115','M6116','M6117','M6118','M6119','M612','M6120','M6121','M6122','M6123','M6124','M6125','M6126','M6127','M6128','M6129','M613','M6130','M6131','M6132','M6133','M6134','M6135','M6136','M6137','M6138','M6139','M614','M6140','M6141','M6142','M6143','M6144','M6145','M6146','M6147','M6148','M6149','M615','M6150','M6151','M6152','M6153','M6154','M6155','M6156','M6157','M6158','M6159','M619','M6190','M6191','M6192','M6193','M6194','M6195','M6196','M6197','M6198','M6199','M652','M653','M654','M700','M720','M722','M724','M750','M751','M753','M754','M797','M841','M89','M890','M8900','M8901','M8902','M8903','M8904','M8905','M8906','M8907','M8908','M8909','M891','M8910','M8911','M8912','M8913','M8914','M8915','M8916','M8917','M8918','M8919','M892','M8920','M8921','M8922','M8923','M8924','M8925','M8926','M8927','M8928','M8929','M893','M8930','M8931','M8932','M8933','M8934','M8935','M8936','M8937','M8938','M8939','M894','M8940','M8941','M8942','M8943','M8944','M8945','M8946','M8947','M8948','M8949','M895','M8950','M8951','M8952','M8953','M8954','M8955','M8956','M8957','M8958','M8959','M896','M8960','M8961','M8962','M8963','M8964','M8965','M8966','M8967','M8968','M8969','M898','M8980','M8981','M8982','M8983','M8984','M8985','M8986','M8987','M8988','M8989','M899','M8990','M8991','M8992','M8993','M8994','M8995','M8996','M8997','M8998','M8999','M91','M910','M9100','M9101','M9102','M9103','M9104','M9105','M9106','M9107','M9108','M9109','M911','M9110','M9111','M9112','M9113','M9114','M9115','M9116','M9117','M9118','M9119','M912','M9120','M9121','M9122','M9123','M9124','M9125','M9126','M9127','M9128','M9129','M913','M9130','M9131','M9132','M9133','M9134','M9135','M9136','M9137','M9138','M9139','M918','M9180','M9181','M9182','M9183','M9184','M9185','M9186','M9187','M9188','M9189','M919','M9190','M9191','M9192','M9193','M9194','M9195','M9196','M9197','M9198','M9199','M93','M930','M931','M932','M9326','

	M938','M9386','M939','M94','M940','M9400','M9401','M9402','M9403','M9404','M9405','M9406','M9407','M9408','M9409','M941','M9410','M9411','M9412','M9413','M9414','M9415','M9416','M9417','M9418','M9419','M942','M9420','M9421','M9422','M9423','M9424','M9425','M9426','M9427','M9428','M9429','M943','M9430','M9431','M9432','M9433','M9434','M9435','M9436','M9437','M9438','M9439','M948','M9480','M9481','M9482','M9483','M9484','M9485','M9486','M9487','M9488','M9489','M949','M9490','M9491','M9492','M9493','M9494','M9495','M9496','M9497','M9498','M9499','M96','M960','M961','M962','M963','M964','M965','M966','M9663','M9665','M968','M969','M99','M990','M9900','M9901','M9902','M9903','M9904','M9905','M9906','M9907','M9908','M9909','M991','M9910','M9911','M9912','M9913','M9914','M9915','M9916','M9917','M9918','M9919','M992','M9920','M9921','M9922','M9923','M9924','M9925','M9926','M9927','M9928','M9929','M993','M9930','M9931','M9932','M9933','M9934','M9935','M9936','M9937','M9938','M9939','M994','M9940','M9941','M9942','M9943','M9944','M9945','M9946','M9947','M9948','M9949','M995','M9950','M9951','M9952','M9953','M9954','M9955','M9956','M9957','M9958','M9959','M996','M9960','M9961','M9962','M9963','M9964','M9965','M9966','M9967','M9968','M9969','M997','M9970','M9971','M9972','M9973','M9974','M9975','M9976','M9977','M9978','M9979','M998','M9980','M9981','M9982','M9983','M9984','M9985','M9986','M9987','M9988','M9989','M999','M9990','M9991','M9992','M9993','M9994','M9995','M9996','M9997','M9998','M9999','Q65','Q650','Q651','Q652','Q653','Q654','Q655','Q656','Q658','Q659','Q66','Q660','Q661','Q662','Q663','Q664','Q665','Q666','Q667','Q668','Q669','Q68','Q680','Q681','Q682','Q683','Q684','Q685','Q688','Q71','Q710','Q711','Q712','Q713','Q714','Q715','Q716','Q718','Q719','Q72','Q720','Q721','Q722','Q723','Q724','Q725','Q726','Q727','Q728','Q729','Q73','Q730','Q731','Q738','Q74','Q740','Q741','Q742','Q743','Q748','Q749','Q77','Q770','Q771','Q772','Q773','Q774','Q775','Q776','Q777','Q778','Q779','Q78','Q780','Q781','Q782','Q783','Q784','Q785','Q786','Q788','Q789','Q796','Q798','Q87','Q870','Q871','Q872','Q873','Q874','Q875','Q878','S382','S48','S480','S481','S489','S58','S580','S581','S589','S68','S680','S681','S682','S683','S684','S688','S689','S78','S780','S781','S789','S88','S880','S881','S889','S98','S980','S981','S982','S983','S984','T05','T050','T051','T052','T053','T054','T055','T056','T058','T059','T096','T116','T136','T147','T90','T900','T901','T902','T903','T904','T905','T908','T909','T91','T910','T911','T912','T913','T914','T915','T918','T919','T92','T920','T921','T922','T923','T924','T925','T926','T928','T929','T93','T930','T931','T932','T933','T934','T935','T936','T938','T939','T94','T940','T941','T95','T950','T951','T952','T953','T954','T958','T959','T96','T97','T98','T980','T981','T982','T983','Z440','Z441','Z891','Z892','Z893','Z894','Z895','Z896','Z897','Z898','Z899','Z946','Z966','Z971'
OTHER NEUROLOGICAL DISEASES	'B900','D482','G041','G09','G10','G11','G110','G111','G112','G113','G114','G118','G119','G12','G120','G121','G122','G128','G129','G13','G132','G138','G24','G240','G241','G242','G243','G244','G245','G248','G249','G25','G250','G252','G253','G255','G258','G259','G26','G32','G320','G328','G37','G370','G371','G372','G373','G374','G375','G378','G379','G51','G511','G512','G513','G514','G518','G519','G52','G520','G521','G522','G523','G527','G528','G529','G53','G530','G531','G532','G533','G538','G70','G700','G701','G702','G708','G709','G71','G710','G711','G712','G713','G718','G719','G723','G724','G728','G729','G73','G730','G731','G735','G736','G737','G80','G800','G801','G802','G803','G804','G808','G809','G81','G810','G811','G819','G82','G820','G821','G822','G823','G824','G825','G83','G830','G831','G832','G833','G834','G835','G839','G90','G900','G901','G902','G903','G904','G908','G909','G91','G910','G911','G912','G913','G918','G919','G938','G939','G95','G950','G951','G952','G958','G959','G99','G990','G991','G992','G998','M471','Q00','Q000','Q001','Q002','Q01','Q010','Q011','Q012','Q018','Q019','Q02','Q03','Q030','Q031','Q038','Q039','Q04','Q040','Q041','Q042','Q043','Q044','Q045','Q046','Q048','Q049','Q05','Q050','Q051','Q052','Q053','Q054','Q055','Q056','Q057','Q058','Q059','Q06','Q060','Q061','Q062','Q063','Q064','Q068','Q069','Q07','Q070','Q078','Q079','Q760'
OTHER PSYCHIATRIC AND BEHAVIORAL DISEASES	'F04','F06','F060','F061','F062','F063','F064','F065','F066','F067','F068','F069','F07','F070','F071','F072','F078','F079','F09','F102','F106','F107','F112','F116','F117','F122','F126','F127','F132','F136','F137','F142','F146','F147','F152','F156','F157','F162','F166','F167','F172','F176','F177','F182','F186','F187','F192','F196','F197','F50','F500','F501','F502','F503','F504','F505','F508','F509','F52','F520','F521','F522','F523','F524','F525','F526','F527','F528','F529','F60','F600','F601','F602','F603','F604','F605','F606','F607','F608','F609','F61','F62','F620','F621','F628','F629','F63','F630','F631','F632','F633','F638','F639','F68','F680','F681','F688','F70','F700','F701','F708','F709','F71','F710','F711','F718','F719','F72','F720','F721','F728','F729','F73','F730','F731','F738','F739','F78','F780','F781','F788','F789','F79','F790','F791','F798','F799','F80','F800','F801','F802','F803','F808','F809','F81','F810','F811','F812','F813','F818','F819','F82','F83','F84','F840','F841','F842','F843','F844','F845','F848','F849','F88','F89','F95','F950','F951','F952','F958','F959','F99'
OTHER RESPIRATORY DISEASES	'B909','E662','J60','J61','J62','J620','J628','J63','J630','J631','J632','J633','J634','J635','J638','J64','J65','J66','J660','J661','J662','J668','J67','J670','J671','J672','J673','J674','J675','J676','J677','J678','J679','J684','J701','J703','J704','J84','J840','J841','J848','J849','J92','J920','J929','J941','J953','J955','J961','J98','J980','J982','J983','J984

	','J985','J986','J988','J989','Q33','Q330','Q331','Q332','Q333','Q334','Q335','Q336','Q338','Q339','Q34','Q340','Q341','Q348','Q349','Z902','Z942','Z943','Z963'
OTHER SKIN DISEASES	'L13','L130','L131','L138','L139','L28','L280','L281','L282','L301','L43','L430','L431','L433','L438','L439','L508','L581','L85','L850','L851','L852','L853','L858','L859','Q80','Q800','Q801','Q802','Q803','Q804','Q808','Q809','Q81','Q810','Q811','Q812','Q818','Q819','Q821','Q822','Q829'
PARKINSON AND PARKINSONISM	'G20','G21','G211','G212','G213','G214','G218','G219','G22','G23','G230','G231','G232','G233','G238','G239'
PERIPHERAL NEUROPATHY	'B91','G14','G54','G540','G541','G542','G543','G544','G545','G546','G547','G548','G549','G55','G550','G551','G552','G553','G558','G56','G560','G561','G562','G563','G564','G568','G569','G57','G570','G571','G572','G573','G574','G575','G576','G578','G579','G58','G580','G587','G588','G589','G59','G590','G598','G60','G600','G601','G602','G603','G608','G609','G628','G629','G63','G630','G632','G633','G634','G635','G636','G638','M472','M531','M541'
PERIPHERAL VASCULAR DISEASE	'I702','I73','I730','I739','I792','I798'
PROSTATE DISEASES	'N40','N411','N418'
SCHIZOPHRENIA AND DELUSIONAL DISEASES	'F20','F200','F201','F202','F203','F204','F205','F206','F208','F209','F22','F220','F228','F229','F24','F25','F250','F251','F252','F258','F259','F28'
SLEEP DISORDERS	'F510','F511','F512','F513','G47','G470','G471','G472','G473','G474','G478','G479'
SOLID NEOPLASMS	'C','D00','D000','D001','D002','D01','D010','D011','D012','D013','D014','D015','D017','D019','D02','D020','D021','D022','D023','D024','D03','D030','D031','D032','D033','D034','D035','D036','D037','D038','D039','D04','D040','D041','D042','D043','D044','D045','D046','D047','D048','D049','D05','D050','D051','D057','D059','D06','D060','D061','D067','D069','D07','D070','D071','D072','D073','D074','D075','D076','D09','D090','D091','D092','D093','D097','D099','D320','D321','D329','D330','D331','D332','D333','D334','Q85','Q850','Q851','Q858','Q859'
THYROID DISEASES	'E00','E000','E001','E002','E009','E01','E010','E011','E012','E018','E02','E03','E030','E031','E032','E033','E034','E038','E039','E05','E050','E051','E052','E053','E054','E055','E058','E059','E062','E063','E065','E07','E070','E071','E078','E079','E350','E890'
VENOUS AND LYMPHATIC DISEASES	'I780','I83','I830','I831','I832','I839','I87','I870','I871','I872','I878','I879','I89','I890','I891','I898','I899','I972','Q820'

Appendix 4: Moving window imputation

Mean corpuscular volume (MCV)	White blood count	Alkaline Phosphatase	Creatinine	Albumin	Red blood cell dist width	Glucose	Percentage lymphocytes	Plasma C-reactive protein level
93.00	6.50	60.4	66.8	48.21	13.20	4.784	39.40	1.72
96.02	8.16	96.8	68.6	45.47	13.34	4.900	22.37	2.58
NA	NA	32.0	84.0	40.00	NA	4.600	NA	NA
89.60	NA	NA	NA	NA	NA	NA	NA	NA
NA	NA	NA	83.0	NA	NA	5.400	NA	NA
NA	NA	57.0	88.0	41.00	NA	5.000	NA	NA
NA	NA	NA	83.0	NA	NA	4.900	NA	NA
86.00	NA	51.0	67.0	41.00	13.70	4.400	NA	NA
87.50	NA	NA	85.0	NA	13.70	NA	NA	NA
NA	NA	58.0	76.0	46.00	NA	NA	NA	NA
86.40	NA	54.0	78.0	42.00	13.70	6.000	NA	NA
NA	NA	51.0	68.0	44.00	NA	NA	NA	NA
86.20	NA	NA	61.0	NA	NA	NA	NA	NA
87.00	NA	46.0	NA	39.00	NA	NA	NA	NA
93.95	3.50	65.7	68.6	NA	20.90	NA	66.40	2.21
NA	NA	NA	66.0	NA	NA	NA	NA	NA
NA	NA	41.0	67.0	36.00	NA	4.800	NA	NA
NA	NA	NA	70.0	NA	NA	NA	NA	NA
84.70	NA	32.0	NA	37.00	14.10	7.800	NA	NA
NA	NA	NA	68.0	NA	NA	NA	NA	NA

→

Mean corpuscular volume (MCV)	White blood count	Alkaline Phosphatase	Creatinine	Albumin	Red blood cell dist width	Glucose	Percentage lymphocytes	Plasma C-reactive protein level
93.00	6.50	60.4	66.8	48.21	13.20	4.784	39.40	1.72
96.02	8.16	96.8	68.6	45.47	13.34	4.900	22.37	2.58
89.60	3.50	32.0	84.0	40.00	13.70	4.600	66.40	2.21
89.60	3.50	57.0	83.0	41.00	13.70	5.400	66.40	2.21
86.00	3.50	57.0	83.0	41.00	13.70	5.400	66.40	2.21
86.00	3.50	57.0	88.0	41.00	13.70	5.000	66.40	2.21
86.00	3.50	51.0	83.0	41.00	13.70	4.900	66.40	2.21
86.00	3.50	51.0	67.0	41.00	13.70	4.400	66.40	2.21
87.50	3.50	58.0	85.0	46.00	13.70	6.000	66.40	2.21
86.40	3.50	58.0	76.0	46.00	13.70	6.000	66.40	2.21
86.40	3.50	54.0	78.0	42.00	13.70	6.000	66.40	2.21
86.20	3.50	51.0	68.0	44.00	20.90	4.800	66.40	2.21
86.20	3.50	46.0	61.0	39.00	20.90	4.800	66.40	2.21
87.00	3.50	46.0	68.6	39.00	20.90	4.800	66.40	2.21
93.95	3.50	65.7	68.6	36.00	20.90	4.800	66.40	2.21
84.70	3.50	41.0	66.0	36.00	14.10	4.800	66.40	2.21
84.70	3.50	41.0	67.0	36.00	14.10	4.800	66.40	2.21
84.70	3.50	32.0	70.0	37.00	14.10	7.800	66.40	2.21
84.70	3.50	32.0	68.0	37.00	14.10	7.800	66.40	2.21
85.90	3.50	50.0	68.0	42.00	14.20	5.300	66.40	2.21

Appendix 5: Disease legend

Labels	Categories
Endocrine	DIABETES
Muscoskeletal	OSTEOARTHRITIS AND OTHER DEGENERATIVE JOINT DISEASES
Eye/ear	CATARACT AND OTHER LENS DISEASES
Muscoskeletal	OTHER MUSCULOSKELETAL AND JOINT DISEASES
Digestive	COLITIS AND RELATED DISEASES
Circulatory	HYPERTENSION
Endocrine	DYSLIPIDEMIA
Mental	DEPRESSION AND MOOD DISEASES
Muscoskeletal	DORSOPATHIES
Endocrine	THYROID DISEASES
Respiratory	ALLERGY
Respiratory	ASTHMA
Eye/ear	EAR, NOSE, THROAT DISEASES
Muscoskeletal	INFLAMMATORY ARTHROPATHIES
Eye/ear	OTHER EYE DISEASES
Digestive	ESOPHAGUS, STOMACH AND DUODENUM DISEASES
Digestive	CHRONIC PANCREAS, BILIARY TRACT AND GALLBLADDER DISEASES
Endocrine	OBESITY
Digestive	OTHER DIGESTIVE DISEASES
Circulatory	ANEMIA
Endocrine	OTHER METABOLIC DISEASES
Circulatory	CEREBROVASCULAR DISEASE
Circulatory	BRADYCARDIAS AND CONDUCTION DISEASES
Circulatory	CARDIAC VALVE DISEASES
Circulatory	ISCHEMIC HEART DISEASE
Eye/ear	DEAFNESS, HEARING IMPAIRMENT
Circulatory	OTHER CARDIOVASCULAR DISEASES
Circulatory	BLOOD AND BLOOD FORMING ORGAN DISEASES
Genitourinary	OTHER GENITOURINARY DISEASES
Respiratory	OTHER RESPIRATORY DISEASES
Genitourinary	PROSTATE DISEASES
Genetic	HEMATOLOGICAL NEOPLASMS
Nervous	OTHER NEUROLOGICAL DISEASES
Nervous	PERIPHERAL NEUROPATHY
Mental	SLEEP DISORDERS
Genitourinary	CHRONIC KIDNEY DISEASES
Genetic	SOLID NEOPLASMS
Circulatory	HEART FAILURE

Circulatory	ATRIAL FIBRILLATION
Muscoskeletal	OSTEOPOROSIS
Muscoskeletal	AUTOIMMUNE DISEASES
Mental	DEMENTIA
Nervous	PARKINSON AND PARKINSONISM
Mental	SCHIZOPHRENIA AND DELUSIONAL DISEASES
Circulatory	VENOUS AND LYMPHATIC DISEASES
Nervous	EPILEPSY
Circulatory	PERIPHERAL VASCULAR DISEASE
Respiratory	COPD, EMPHYSEMA, CHRONIC BRONCHITIS
Skin	CHRONIC ULCER OF THE SKIN
Nervous	MIGRAINE AND FACIAL PAIN SYNDROMES
Mental	OTHER PSYCHIATRIC AND BEHAVIORAL DISEASES
Skin	OTHER SKIN DISEASES
Digestive	CHRONIC LIVER DISEASES
Eye/ear	GLAUCOMA
Digestive	INFLAMMATORY BOWEL DISEASES
Infectious	CHRONIC INFECTIOUS DISEASES
Nervous	MULTIPLE SCLEROSIS
Eye/ear	BLINDNESS, VISUAL IMPAIRMENT
Genetic	CHROMOSOMAL ABNORMALITIES
Mental	NEUROTIC, STRESS-RELATED AND SOMATOFORM DISEASES

Appendix 6: Summary Statistics

Variable	CHROM	GENPOS	ID	ALLELE0	ALLELE1	A1FREQ	BETA	SE	CHISQ	LOG10 P	p
HRH	19	45411941	rs429358	C	T	0,845997	-0,17859	0,017078	109,351	24,8667	1,36E-25
HRH	19	45413234	19:45413234_GGT_G	G	GGT	0,865605	-0,19016	0,019185	98,2454	22,4322	3,7E-23
HRH	19	45410002	rs769449	A	G	0,876959	-0,16552	0,018778	77,698	17,9206	1,2E-18
HRH	19	45392254	rs6857	T	C	0,831118	-0,14417	0,016602	75,4092	17,4173	3,83E-18
HRH	19	45424351	rs814573	T	A	0,812201	-0,14151	0,01641	74,3663	17,1878	6,49E-18
HRH	19	45415935	rs7256200	T	G	0,873442	-0,15471	0,018676	68,6241	15,9239	1,19E-16
HRH	19	45415713	rs10414043	A	G	0,873551	-0,15464	0,018683	68,5098	15,8988	1,26E-16
HRH	19	45422160	rs12721051	G	C	0,81418	-0,13034	0,016001	66,3572	15,4246	3,76E-16
HRH	19	45422846	rs56131196	A	G	0,812601	-0,12908	0,01594	65,5782	15,253	5,58E-16
HRH	19	45422946	rs4420638	G	A	0,81236	-0,12784	0,015932	64,3787	14,9886	1,03E-15
HRH	19	45424514	rs157592	C	A	0,815105	-0,1306	0,016468	62,8938	14,6612	2,18E-15
HRH	19	45395714	rs157581	C	T	0,780554	-0,11728	0,015173	59,7484	13,9674	1,08E-14
HRH	19	45390333	rs283815	G	A	0,781434	-0,11731	0,015238	59,2721	13,8623	1,37E-14
HRH	19	45396665	rs59007384	T	G	0,792507	-0,11896	0,015468	59,1431	13,8338	1,47E-14
HRH	19	45388500	rs283811	G	A	0,780305	-0,11756	0,015325	58,8465	13,7684	1,7E-14
HRH	19	45396219	rs157582	T	C	0,782265	-0,11647	0,015222	58,5412	13,701	1,99E-14
HRH	19	45394969	rs184017	G	T	0,781748	-0,11594	0,015209	58,1168	13,6073	2,47E-14
HRH	19	45386467	rs142042446	GTAA	G	0,853542	-0,1352	0,017854	57,3406	13,4359	3,67E-14
HRH	19	45396144	rs11556505	T	C	0,855113	-0,13296	0,017667	56,6416	13,2815	5,23E-14
HRH	19	45388130	rs34342646	A	G	0,850675	-0,13211	0,017579	56,4742	13,2446	5,69E-14
HRH	19	45394336	rs71352238	C	T	0,853519	-0,1321	0,017612	56,2604	13,1974	6,35E-14
HRH	19	45387459	rs12972156	G	C	0,853791	-0,13297	0,017732	56,23	13,1906	6,45E-14

HRH	19	45395619	rs2075650	G	A	0,854836	-0,13232	0,017647	56,2176	13,1879	6,49E-14
HRH	19	45395909	rs34404554	G	C	0,855338	-0,13234	0,017679	56,0359	13,1478	7,12E-14
HRH	19	45387596	rs12972970	A	G	0,853694	-0,13242	0,017723	55,8249	13,1012	7,92E-14
HRH	19	45427125	rs111789331	A	T	0,844426	-0,11948	0,017583	46,1755	10,9661	1,08E-11
HRH	15	91530158	rs11853073	C	T	0,752627	0,102528	0,015119	45,9849	10,9238	1,19E-11
HRH	15	91523036	rs6496742	T	C	0,753881	0,102413	0,015112	45,9272	10,911	1,23E-11
HRH	15	91530820	rs8028856	G	A	0,754539	0,102297	0,015115	45,8039	10,8837	1,31E-11
HRH	15	91528794	15:91528794_CT_C	C	CT	0,753614	0,102652	0,015172	45,7757	10,8775	1,33E-11
HRH	15	91522070	rs8032722	C	T	0,752505	0,102041	0,015118	45,5557	10,8287	1,48E-11
HRH	15	91519660	rs8036430	C	A	0,752177	0,101604	0,015131	45,0899	10,7254	1,88E-11
HRH	15	91520308	rs12595052	T	C	0,753883	0,101663	0,01518	44,8511	10,6724	2,13E-11
HRH	19	45428234	rs66626994	A	G	0,843722	-0,11791	0,017655	44,601	10,617	2,42E-11
HRH	15	91529924	15:91529924_TTTTATTTA_T	T	TTTTATTTA	0,754782	0,100619	0,015113	44,3235	10,5554	2,78E-11
HRH	19	45421254	rs12721046	A	G	0,845669	-0,11453	0,017312	43,7605	10,4305	3,71E-11
HRH	15	91512067	rs2290203	A	G	0,793884	0,105949	0,016059	43,5279	10,3789	4,18E-11
HRH	15	91507202	rs7171333	C	G	0,62247	0,087703	0,013466	42,4174	10,1324	7,37E-11
HRH	15	91511260	rs12910825	G	A	0,623476	0,087066	0,013385	42,314	10,1094	7,77E-11
HRH	15	91525197	rs2301826	T	C	0,79141	0,104589	0,016081	42,2984	10,1059	7,84E-11
HRH	15	91508803	rs4932182	A	C	0,632147	0,086842	0,013474	41,5404	9,93757	1,15E-10
HRH	15	91533621	15:91533621_ATT_A	A	ATT	0,797626	0,104089	0,016188	41,3447	9,89409	1,28E-10
HRH	19	45395266	rs157580	A	G	0,393501	-0,08385	0,013061	41,2179	9,86593	1,36E-10
HRH	15	91506452	rs2034085	C	T	0,630138	0,086334	0,013504	40,8701	9,78863	1,63E-10
HRH	15	91506637	rs8037137	C	T	0,865517	0,122467	0,019277	40,3594	9,67512	2,11E-10
HRH	15	91526275	rs7180016	A	G	0,826152	0,110343	0,017395	40,2386	9,64827	2,25E-10
HRH	15	91502849	rs8038661	A	G	0,860265	0,120115	0,019012	39,9139	9,57608	2,65E-10
HRH	15	91509140	15:91509140_CA_C	C	CA	0,63159	0,08536	0,013519	39,8665	9,56555	2,72E-10

HRH	15	91532869	rs6496746	A	C	0,823874	0,109198	0,017355	39,5893	9,50392	3,13E-10
HRH	15	91529452	rs60290423	C	T	0,867136	0,121064	0,019248	39,5584	9,49704	3,18E-10
HRH	15	91520131	rs12595025	T	C	0,866562	0,120576	0,019227	39,3277	9,44573	3,58E-10
HRH	15	91513157	rs2890156	A	T	0,867997	0,12099	0,019305	39,2786	9,43481	3,67E-10
HRH	15	91535938	15:91535938_GT_G	G	GT	0,823698	0,108494	0,017341	39,1452	9,40512	3,93E-10
HRH	15	91512267	rs2290202	T	G	0,867947	0,120693	0,019305	39,0858	9,39191	4,06E-10
HRH	15	91531352	rs3803563	A	C	0,827018	0,108244	0,01733	39,0128	9,37568	4,21E-10
HRH	15	91528961	rs8036115	A	G	0,827075	0,10797	0,017331	38,8133	9,3313	4,66E-10
HRH	15	91529122	rs8042518	C	T	0,827075	0,107941	0,017331	38,7907	9,32628	4,72E-10
HRH	15	91519642	rs8035844	T	C	0,826512	0,107831	0,017318	38,7685	9,32134	4,77E-10
HRH	15	91528070	rs2301825	G	C	0,868227	0,119945	0,019292	38,6558	9,29626	5,06E-10
HRH	15	91523713	rs1867226	G	C	0,826931	0,107605	0,017323	38,5836	9,28019	5,25E-10
HRH	15	91532565	rs7179428	A	C	0,826941	0,107553	0,017322	38,5519	9,27313	5,33E-10
HRH	15	91531995	rs12594752	T	C	0,86844	0,119817	0,019303	38,5272	9,26764	5,4E-10
HRH	15	91532630	rs6496745	T	C	0,826954	0,107451	0,017322	38,4792	9,25697	5,53E-10
HRH	15	91522253	rs8026714	A	G	0,866953	0,119295	0,019234	38,4703	9,25497	5,56E-10
HRH	15	91506422	rs8028409	T	A	0,867511	0,119769	0,019352	38,3014	9,21739	6,06E-10
HRH	15	91505779	rs79548680	C	G	0,86191	0,117736	0,019038	38,2442	9,20466	6,24E-10
HRH	15	91508579	15:91508579_AG_A	A	AG	0,775152	0,097217	0,01574	38,1492	9,18352	6,55E-10
HRH	15	91528875	rs150333865	A	G	0,866997	0,118651	0,019238	38,0393	9,15906	6,93E-10
HRH	15	91533816	rs8025699	G	A	0,826793	0,106744	0,017316	38,0027	9,15092	7,06E-10
HRH	15	91514707	rs59278520	A	G	0,826151	0,106716	0,017313	37,994	9,14897	7,1E-10
HRH	15	91530675	rs8026781	A	G	0,827556	0,106768	0,017331	37,9525	9,13974	7,25E-10
HRH	15	91507495	rs112770790	A	G	0,867605	0,11909	0,019333	37,9434	9,13772	7,28E-10
HRH	15	91520287	rs12594925	A	G	0,869271	0,119546	0,019417	37,9081	9,12984	7,42E-10
HRH	15	91515489	rs59902579	T	C	0,826059	0,1066	0,017315	37,9028	9,12868	7,44E-10

HRH	15	91527152	rs2285937	G	A	0,826459	0,106512	0,017333	37,7608	9,09706	8E-10
HRH	15	91527451	rs8031684	A	T	0,826504	0,106445	0,017333	37,7145	9,08674	8,19E-10
HRH	15	91536225	rs73502775	A	G	0,826585	0,106221	0,0173	37,6973	9,08293	8,26E-10
HRH	15	91535329	rs76119208	T	G	0,868522	0,118355	0,019299	37,6119	9,06391	8,63E-10
HRH	15	91551876	rs79253100	T	C	0,86876	0,118572	0,01938	37,4331	9,02409	9,46E-10
HRH	15	91502383	rs1105291	T	C	0,864184	0,117191	0,019192	37,2872	8,9916	1,02E-09
HRH	15	91534294	rs112553294	G	A	0,82641	0,105616	0,017309	37,2325	8,97943	1,05E-09
HRH	15	91538920	rs78029804	T	C	0,869567	0,117484	0,019344	36,8863	8,90232	1,25E-09
HRH	15	91523763	rs28584391	G	T	0,860544	0,115255	0,019002	36,7912	8,88113	1,31E-09
HRH	15	91523742	15:91523742_CTTTT_C	C	CTTTT	0,860558	0,115206	0,019001	36,7607	8,87433	1,34E-09
HRH	15	91509215	rs77554484	A	C	0,869734	0,117656	0,019428	36,6742	8,85507	1,4E-09
URU	11	61589481	rs28456	G	A	0,690084	0,107932	0,017891	36,3935	8,79252	1,61E-09
HRH	15	91548601	rs3826033	A	G	0,870841	0,117407	0,019472	36,3546	8,78387	1,64E-09
HRH	15	91528358	rs6496743	A	G	0,676996	0,083728	0,013967	35,9389	8,69121	2,04E-09
HRH	15	91536627	rs6496747	G	A	0,676314	0,083579	0,013956	35,865	8,67475	2,11E-09
HRH	15	91561182	rs3743445	T	C	0,877564	0,119847	0,020022	35,8298	8,66689	2,15E-09
URU	19	45395909	rs34404554	G	C	0,855338	0,142613	0,023872	35,6887	8,63543	2,32E-09
HRH	15	91521898	rs7167128	G	A	0,677269	0,083249	0,01396	35,5608	8,60693	2,47E-09
HRH	15	91520373	rs4544218	A	C	0,675005	0,083212	0,013967	35,4968	8,59265	2,55E-09
HRH	15	91502183	rs201055286	C	CA	0,780604	0,097217	0,01633	35,4422	8,58048	2,63E-09
HRH	15	91502166	15:91502166_TAA_T	T	TAA	0,780603	0,097213	0,01633	35,4394	8,57985	2,63E-09
HRH	15	91558457	rs7183740	A	G	0,877844	0,119354	0,020056	35,4149	8,57441	2,66E-09
HRH	15	91503871	rs1867219	G	T	0,67458	0,083164	0,014043	35,0727	8,49808	3,18E-09
HRH	15	91520319	rs12593261	C	T	0,678233	0,082761	0,013978	35,0582	8,49486	3,2E-09
URU	11	61582708	rs174561	C	T	0,696209	0,106693	0,018021	35,052	8,49348	3,21E-09
URU	11	61579760	rs174555	C	T	0,696283	0,106602	0,018016	35,0105	8,48421	3,28E-09

HRH	15	91522092	rs113343095	A	G	0,871699	0,115521	0,019527	34,9996	8,48179	3,3E-09
URU	11	61571382	rs174549	A	G	0,696112	0,106531	0,018012	34,9795	8,47729	3,33E-09
HRH	15	91516015	rs1443256	C	T	0,676152	0,082387	0,013957	34,845	8,4473	3,57E-09
URU	11	61581368	rs174557	G	A	0,696295	0,10626	0,01802	34,7724	8,43112	3,71E-09
URU	19	45388130	rs34342646	A	G	0,850675	0,139851	0,023718	34,7694	8,43044	3,71E-09
HRH	15	91503752	rs1867220	C	T	0,674572	0,082793	0,014044	34,7544	8,42708	3,74E-09
URU	11	61581764	rs174560	C	T	0,689613	0,105295	0,017861	34,7539	8,42697	3,74E-09
HRH	15	91520099	rs8037254	G	A	0,676613	0,082259	0,013954	34,7502	8,42614	3,75E-09
URU	19	45396144	rs11556505	T	C	0,855113	0,140441	0,023838	34,7081	8,41677	3,83E-09
URU	11	61571348	rs174548	G	C	0,689444	0,10518	0,017854	34,7069	8,4165	3,83E-09
URU	19	45395619	rs2075650	G	A	0,854836	0,139677	0,023798	34,4483	8,3588	4,38E-09
HRH	15	91514721	rs4508401	A	T	0,67641	0,081902	0,013959	34,423	8,35314	4,43E-09
URU	11	61575807	rs751229834	C	CA	0,698631	0,106395	0,018153	34,3507	8,33703	4,6E-09
HRH	15	91513832	15:91513832_GGGA_G	G	GGGA	0,676802	0,081693	0,013958	34,2575	8,31623	4,83E-09
HRH	15	91514797	rs11853024	A	C	0,678105	0,081531	0,013934	34,2385	8,31198	4,88E-09
HRH	15	91514814	rs7166743	A	C	0,677149	0,081586	0,013948	34,2132	8,30633	4,94E-09
URU	11	61580635	rs174556	T	C	0,697077	0,105432	0,018031	34,1923	8,30167	4,99E-09
HRH	15	91520159	rs12595029	G	C	0,676384	0,081493	0,013958	34,0866	8,27808	5,27E-09
HRH	15	91517780	rs12898311	C	G	0,675692	0,081362	0,013952	34,0054	8,25995	5,5E-09
URU	19	45394336	rs71352238	C	T	0,853519	0,138573	0,023771	33,9821	8,25475	5,56E-09
URU	11	61567753	rs174544	A	C	0,697168	0,105069	0,018029	33,9644	8,2508	5,61E-09
HRH	6	32620909	rs17843625	G	A	0,489189	0,075452	0,012954	33,9242	8,24184	5,73E-09
HRH	15	91537151	15:91537151_TA_T	T	TA	0,67696	0,081286	0,013972	33,8488	8,225	5,96E-09
HRH	15	91518800	rs36074798	ACT	A	0,676723	0,08113	0,013949	33,8276	8,22028	6,02E-09
HRH	15	91519291	rs8031915	G	A	0,676418	0,081148	0,01396	33,7883	8,2115	6,14E-09
HRH	19	45395844	rs34095326	A	G	0,888899	-0,11543	0,019864	33,7671	8,20677	6,21E-09

HRH	15	91506739	rs2034084	G	C	0,681702	0,081785	0,014078	33,7515	8,20328	6,26E-09
URU	11	61565908	rs174541	C	T	0,646884	0,100142	0,017284	33,5696	8,16266	6,88E-09
HRH	15	91502524	rs5814451	TA	T	0,67499	0,081543	0,01408	33,5416	8,15641	6,98E-09
URU	11	61603510	rs174576	A	C	0,651548	0,100122	0,017305	33,4746	8,14146	7,22E-09
URU	19	45387596	rs12972970	A	G	0,853694	0,138188	0,023917	33,3841	8,12124	7,56E-09
HRH	6	32628765	rs9273537	T	C	0,534426	0,075735	0,013118	33,3296	8,10909	7,78E-09
HRH	15	91521337	rs8042680	A	C	0,678214	0,080639	0,013974	33,3026	8,10304	7,89E-09
URU	19	45387459	rs12972156	G	C	0,853791	0,138058	0,023931	33,2809	8,0982	7,98E-09
HRH	11	61589481	rs28456	G	A	0,690084	-0,07825	0,013625	32,9853	8,03217	9,29E-09
URU	11	61564299	rs4246215	T	G	0,647292	0,099202	0,01729	32,9205	8,0177	9,6E-09

Appendix 7: Longevity associated variants

SNP	Symbol in study	Study metadata	Gene annotation	Type	Variable	BETA	SE	CHISQ	LOG10P
rs4420638	APOE	Age > 90 (10.1093/hmg/ddu139)	CD81	eQTL	HRH	-0,12784	0,015932	64,3787	14,9886
rs4420638	APOE	Age > 90 (10.1093/hmg/ddu139)	CCDC86	eQTL	HRH	-0,12784	0,015932	64,3787	14,9886
rs4420638	APOE	Age > 90 (10.1093/hmg/ddu139)	PTGDR2	eQTL	HRH	-0,12784	0,015932	64,3787	14,9886
rs4420638	APOE	Age > 90 (10.1093/hmg/ddu139)	MAD1L1	eQTL	HRH	-0,12784	0,015932	64,3787	14,9886
rs4420638	APOE	Age > 90 (10.1093/hmg/ddu139)	TOMM40	eQTL	HRH	-0,12784	0,015932	64,3787	14,9886
rs4420638	TOMM40	Age > 90 (10.1093/hmg/ddu139)	CD81	eQTL	HRH	-0,12784	0,015932	64,3787	14,9886
rs4420638	TOMM40	Age > 90 (10.1093/hmg/ddu139)	CCDC86	eQTL	HRH	-0,12784	0,015932	64,3787	14,9886
rs4420638	TOMM40	Age > 90 (10.1093/hmg/ddu139)	PTGDR2	eQTL	HRH	-0,12784	0,015932	64,3787	14,9886
rs4420638	TOMM40	Age > 90 (10.1093/hmg/ddu139)	MAD1L1	eQTL	HRH	-0,12784	0,015932	64,3787	14,9886
rs4420638	TOMM40	Age > 90 (10.1093/hmg/ddu139)	TOMM40	eQTL	HRH	-0,12784	0,015932	64,3787	14,9886
rs429358	APOE	Parental mortality (10.1038/ncomms11174)	TOMM40	eQTL	HRH	-0,17859	0,017078	109,351	24,8667
rs429358	APOE	Parental mortality (10.1038/ncomms11174)	APOE	Proxy	HRH	-0,17859	0,017078	109,351	24,8667
rs429358	TOMM40	Parental mortality (10.1038/ncomms11174)	TOMM40	eQTL	HRH	-0,17859	0,017078	109,351	24,8667
rs429358	TOMM40	Parental mortality (10.1038/ncomms11174)	APOE	Proxy	HRH	-0,17859	0,017078	109,351	24,8667
rs2075650	APOE	Exceptional longevity (10.1371/journal.pgen.1005728.t003)	TOMM40	eQTL	HRH	-0,13232	0,017647	56,2176	13,1879
rs2075650	APOE	Exceptional longevity (10.1371/journal.pgen.1005728.t003)	TOMM40	eQTL	URU	0,139677	0,023798	34,4483	8,3588
rs2075650	APOE	Exceptional longevity (10.1371/journal.pgen.1005728.t003)	PVRL2	eQTL	HRH	-0,13232	0,017647	56,2176	13,1879
rs2075650	APOE	Exceptional longevity (10.1371/journal.pgen.1005728.t003)	PVRL2	eQTL	URU	0,139677	0,023798	34,4483	8,3588
rs2075650	APOE	Exceptional longevity (10.1371/journal.pgen.1005728.t003)	TOMM40	Proxy	HRH	-0,13232	0,017647	56,2176	13,1879
rs2075650	APOE	Exceptional longevity (10.1371/journal.pgen.1005728.t003)	TOMM40	Proxy	URU	0,139677	0,023798	34,4483	8,3588
rs2075650	TOMM40	Exceptional longevity (10.1371/journal.pgen.1005728.t003)	TOMM40	eQTL	HRH	-0,13232	0,017647	56,2176	13,1879
rs2075650	TOMM40	Exceptional longevity (10.1371/journal.pgen.1005728.t003)	TOMM40	eQTL	URU	0,139677	0,023798	34,4483	8,3588
rs2075650	TOMM40	Exceptional longevity (10.1371/journal.pgen.1005728.t003)	PVRL2	eQTL	HRH	-0,13232	0,017647	56,2176	13,1879
rs2075650	TOMM40	Exceptional longevity (10.1371/journal.pgen.1005728.t003)	PVRL2	eQTL	URU	0,139677	0,023798	34,4483	8,3588
rs2075650	TOMM40	Exceptional longevity (10.1371/journal.pgen.1005728.t003)	TOMM40	Proxy	HRH	-0,13232	0,017647	56,2176	13,1879
rs2075650	TOMM40	Exceptional longevity (10.1371/journal.pgen.1005728.t003)	TOMM40	Proxy	URU	0,139677	0,023798	34,4483	8,3588

6 References

Acharjee, A., H. J. Finkers, R. G. F. Visser, and C. A. Maliepaard. 2013. 'Comparison of Regularized Regression Methods for ~Omics Data'. *Metabolomics : Open Access* 3 (3). https://doi.org/10.4172/2153-0769.1000126.

Acharjee, Animesh, Philippa Prentice, Carlo Acerini, James Smith, Ieuan A. Hughes, Ken Ong, Julian L. Griffin, David Dunger, and Albert Koulman. 2017. 'The Translation of Lipid Profiles to Nutritional Biomarkers in the Study of Infant Metabolism'. *Metabolomics* 13 (3): 25. https://doi.org/10.1007/s11306-017-1166-2.

Ahadi, Sara, Wenyu Zhou, Sophia Miryam Schüssler-Fiorenza Rose, M. Reza Sailani, Kévin Contrepois, Monika Avina, Melanie Ashland, Anne Brunet, and Michael Snyder. 2020. 'Personal Aging Markers and Ageotypes Revealed by Deep Longitudinal Profiling'. *Nature Medicine* 26 (1): 83–90. https://doi.org/10.1038/s41591-019-0719-5.

Ahmad, Tariq, Nihar R. Desai, Yu Yamamoto, Aditya Biswas, Lama Ghazi, Melissa Martin, Michael Simonov, et al. 2022. 'Alerting Clinicians to 1-Year Mortality Risk in Patients Hospitalized With Heart Failure: The REVEAL-HF Randomized Clinical Trial'. *JAMA Cardiology* 7 (9): 905–12. https://doi.org/10.1001/jamacardio.2022.2496.

Akker, Erik B. van den, Stella Trompet, Jurriaan J.H. Barkey Wolf, Marian Beekman, H. Eka D. Suchiman, Joris Deelen, Folkert W. Asselbergs, et al. 2020. 'Metabolic Age Based on the BBMRI-NL 1H-NMR Metabolomics Repository as Biomarker of Age-Related Disease'. *Circulation: Genomic and Precision Medicine* 13 (5): 541–47. https://doi.org/10.1161/CIRCGEN.119.002610.

Albert, Frank W., and Leonid Kruglyak. 2015. 'The Role of Regulatory Variation in Complex Traits and Disease'. *Nature Reviews Genetics* 16 (4): 197–212. https://doi.org/10.1038/nrg3891.

Ali, Moez. 2020. *PyCaret: An Open Source, Low-Code Machine Learning Library in Python*. https://www.pycaret.org.

Allen, Luke N., Simon Wigley, Hampus Holmer, and Pepita Barlow. 2023. 'Non-Communicable Disease Policy Implementation from 2014 to 2021: A Repeated Cross-Sectional Analysis of Global Policy Data for 194 Countries'. *The Lancet Global Health* 11 (4): e525–33. https://doi.org/10.1016/S2214-109X(23)00042-6.

Alpert, Ayelet, Yishai Pickman, Michael Leipold, Yael Rosenberg-Hasson, Xuhuai Ji, Renaud Gaujoux, Hadas Rabani, et al. 2019. 'A Clinically Meaningful Metric of Immune Age Derived from High-Dimensional Longitudinal Monitoring'. *Nature Medicine* 25 (3): 487–95. https://doi.org/10.1038/s41591-019-0381-y.

Altshuler, David, Mark J. Daly, and Eric S. Lander. 2008. 'Genetic Mapping in Human Disease'. *Science* 322 (5903): 881–88. https://doi.org/10.1126/science.1156409.

Ashley, Euan A. 2016. 'Towards Precision Medicine'. *Nature Reviews. Genetics* 17 (9): 507–22. https://doi.org/10.1038/nrg.2016.86.

Baghdadi, Maarouf, Helena M Hinterding, Linda Partridge, and Joris Deelen. 2022. 'From Mutation to Mechanism: Deciphering the Molecular Function of Genetic Variants Linked to Human Ageing'. *Briefings in Functional Genomics* 21 (1): 13–23. https://doi.org/10.1093/bfgp/elab005.

Bamba, Shigeki, Masaya Sasaki, Azusa Takaoka, Kenichiro Takahashi, Hirotsugu Imaeda, Atsushi Nishida, Osamu Inatomi, Mitsushige Sugimoto, and Akira Andoh. 2017. 'Sarcopenia Is a Predictive Factor for Intestinal Resection in Admitted Patients with

Crohn's Disease'. *PLOS ONE* 12 (6): e0180036. https://doi.org/10.1371/journal.pone.0180036.

Barardo, Diogo, Daniel Thornton, Harikrishnan Thoppil, Michael Walsh, Samim Sharifi, Susana Ferreira, Andreja Anžič, et al. 2017. 'The DrugAge Database of Aging-Related Drugs'. *Aging Cell* 16 (3): 594–97. https://doi.org/10.1111/acel.12585.

Barbeira, Alvaro N., Rodrigo Bonazzola, Eric R. Gamazon, Yanyu Liang, YoSon Park, Sarah Kim-Hellmuth, Gao Wang, et al. 2021. 'Exploiting the GTEx Resources to Decipher the Mechanisms at GWAS Loci'. *Genome Biology* 22 (1): 49. https://doi.org/10.1186/s13059-020-02252-4.

Barker, Michelle, Neil P. Chue Hong, Daniel S. Katz, Anna-Lena Lamprecht, Carlos Martinez-Ortiz, Fotis Psomopoulos, Jennifer Harrow, et al. 2022. 'Introducing the FAIR Principles for Research Software'. *Scientific Data* 9 (1): 622. https://doi.org/10.1038/s41597-022-01710-x.

Barone, Lindsay, Jason Williams, and David Micklos. 2017. 'Unmet Needs for Analyzing Biological Big Data: A Survey of 704 NSF Principal Investigators'. *PLOS Computational Biology* 13 (10): e1005755. https://doi.org/10.1371/journal.pcbi.1005755.

Barredo Arrieta, Alejandro, Natalia Díaz-Rodríguez, Javier Del Ser, Adrien Bennetot, Siham Tabik, Alberto Barbado, Salvador Garcia, et al. 2020. 'Explainable Artificial Intelligence (XAI): Concepts, Taxonomies, Opportunities and Challenges toward Responsible AI'. *Information Fusion* 58 (June): 82–115. https://doi.org/10.1016/j.inffus.2019.12.012.

Barretina, Jordi, Giordano Caponigro, Nicolas Stransky, Kavitha Venkatesan, Adam A. Margolin, Sungjoon Kim, Christopher J. Wilson, et al. 2012. 'The Cancer Cell Line Encyclopedia Enables Predictive Modelling of Anticancer Drug Sensitivity'. *Nature* 483 (7391): 603–7. https://doi.org/10.1038/nature11003.

Bates, Stephen, Trevor Hastie, and Robert Tibshirani. 2022. 'Cross-Validation: What Does It Estimate and How Well Does It Do It?' arXiv. https://doi.org/10.48550/arXiv.2104.00673.

Belkin, Mikhail, Daniel Hsu, Siyuan Ma, and Soumik Mandal. 2019. 'Reconciling Modern Machine-Learning Practice and the Classical Bias–Variance Trade-Off'. *Proceedings of the National Academy of Sciences* 116 (32): 15849–54. https://doi.org/10.1073/pnas.1903070116.

Ben Hassen, Céline, Aurore Fayosse, Benjamin Landré, Martina Raggi, Mikaela Bloomberg, Séverine Sabia, and Archana Singh-Manoux. 2022. 'Association between Age at Onset of Multimorbidity and Incidence of Dementia: 30 Year Follow-up in Whitehall II Prospective Cohort Study'. *BMJ (Clinical Research Ed.)* 376 (February): e068005. https://doi.org/10.1136/bmj-2021-068005.

Berger, Bonnie, Jian Peng, and Mona Singh. 2013. 'Computational Solutions for Omics Data'. *Nature Reviews. Genetics* 14 (5): 333–46. https://doi.org/10.1038/nrg3433.

Biecek, Przemyslaw. 2018. 'DALEX: Explainers for Complex Predictive Models in R'. *Journal of Machine Learning Research* 19 (84): 1–5.

Blair, David R., Christopher S. Lyttle, Jonathan M. Mortensen, Charles F. Bearden, Anders Boeck Jensen, Hossein Khiabanian, Rachel Melamed, et al. 2013. 'A Nondegenerate Code of Deleterious Variants in Mendelian Loci Contributes to Complex Disease Risk'. *Cell* 155 (1): 70–80. https://doi.org/10.1016/j.cell.2013.08.030.

Blasimme, Alessandro, and Effy Vayena. 2019. 'The Ethics of AI in Biomedical Research, Patient Care and Public Health'. SSRN Scholarly Paper ID 3368756. Rochester, NY: Social Science Research Network. https://doi.org/10.2139/ssrn.3368756.

Borchers, C.H., J. Kast, L.J. Foster, K.W.M. Siu, C.M. Overall, T.A. Binkowski, W.H. Hildebrand, A. Scherer, M. Mansoor, and P.A. Keown. 2014. 'The Human Proteome

Organization Chromosome 6 Consortium: Integrating Chromosome-Centric and Biology/Disease Driven Strategies'. *Journal of Proteomics* 100 (April): 60–67. https://doi.org/10.1016/j.jprot.2013.08.001.

Bou Sleiman, Maroun, Pooja Jha, Riekelt Houtkooper, Robert W. Williams, Xu Wang, and Johan Auwerx. 2020. 'The Gene-Regulatory Footprint of Aging Highlights Conserved Central Regulators'. *Cell Reports* 32 (13): 108203. https://doi.org/10.1016/j.celrep.2020.108203.

Boyle, Evan A., Yang I. Li, and Jonathan K. Pritchard. 2017. 'An Expanded View of Complex Traits: From Polygenic to Omnigenic'. *Cell* 169 (7): 1177–86. https://doi.org/10.1016/j.cell.2017.05.038.

Bravo-Merodio, Laura, Animesh Acharjee, Jon Hazeldine, Conor Bentley, Mark Foster, Georgios V. Gkoutos, and Janet M. Lord. 2019. 'Machine Learning for the Detection of Early Immunological Markers as Predictors of Multi-Organ Dysfunction'. *Scientific Data* 6 (1): 1–10. https://doi.org/10.1038/s41597-019-0337-6.

Bravo-Merodio, Laura, John A. Williams, Georgios V. Gkoutos, and Animesh Acharjee. 2019. '-Omics Biomarker Identification Pipeline for Translational Medicine'. *Journal of Translational Medicine* 17 (1): 155. https://doi.org/10.1186/s12967-019-1912-5.

Burton, Paul R., David G. Clayton, Lon R. Cardon, Nick Craddock, Panos Deloukas, Audrey Duncanson, Dominic P. Kwiatkowski, et al. 2007. 'Genome-Wide Association Study of 14,000 Cases of Seven Common Diseases and 3,000 Shared Controls'. *Nature* 447 (7145): 661–78. https://doi.org/10.1038/nature05911.

Bush, William S., Matthew T. Oetjens, and Dana C. Crawford. 2016. 'Unravelling the Human Genome–Phenome Relationship Using Phenome-Wide Association Studies'. *Nature Reviews Genetics* 17 (3): 129–45. https://doi.org/10.1038/nrg.2015.36.

Bycroft, Clare, Colin Freeman, Desislava Petkova, Gavin Band, Lloyd T. Elliott, Kevin Sharp, Allan Motyer, et al. 2018. 'The UK Biobank Resource with Deep Phenotyping and Genomic Data'. *Nature* 562 (7726): 203–9. https://doi.org/10.1038/s41586-018-0579-z.

Cabrera, Claudia P., Joanna Manson, Joanna M. Shepherd, Hew D. Torrance, David Watson, M. Paula Longhi, Mimoza Hoti, et al. 2017. 'Signatures of Inflammation and Impending Multiple Organ Dysfunction in the Hyperacute Phase of Trauma: A Prospective Cohort Study'. *PLOS Medicine* 14 (7): e1002352. https://doi.org/10.1371/journal.pmed.1002352.

Calderón-Larrañaga, Amaia, Davide L. Vetrano, Graziano Onder, Luis A. Gimeno-Feliu, Carlos Coscollar-Santaliestra, Angelo Carfí, Maria S. Pisciotta, et al. 2017. 'Assessing and Measuring Chronic Multimorbidity in the Older Population: A Proposal for Its Operationalization'. *The Journals of Gerontology. Series A, Biological Sciences and Medical Sciences* 72 (10): 1417–23. https://doi.org/10.1093/gerona/glw233.

Canet, Jaume, Lluís Gallart, Carmen Gomar, Guillem Paluzie, Jordi Vallès, Jordi Castillo, Sergi Sabaté, et al. 2010. 'Prediction of Postoperative Pulmonary Complications in a Population-Based Surgical Cohort'. *Anesthesiology* 113 (6): 1338–50. https://doi.org/10.1097/ALN.0b013e3181fc6e0a.

Canet, Jaume, Sergi Sabaté, Valentín Mazo, Lluís Gallart, Marcelo Gama de Abreu, Javier Belda, Olivier Langeron, Andreas Hoeft, Paolo Pelosi, and PERISCOPE group. 2015. 'Development and Validation of a Score to Predict Postoperative Respiratory Failure in a Multicentre European Cohort: A Prospective, Observational Study'. *European Journal of Anaesthesiology* 32 (7): 458–70. https://doi.org/10.1097/EJA.0000000000000223.

Canuel, Vincent, Bastien Rance, Paul Avillach, Patrice Degoulet, and Anita Burgun. 2015. 'Translational Research Platforms Integrating Clinical and Omics Data: A Review of

Publicly Available Solutions'. *Briefings in Bioinformatics* 16 (2): 280–90. https://doi.org/10.1093/bib/bbu006.

Caruana, Rich, Yin Lou, Johannes Gehrke, Paul Koch, Marc Sturm, and Noemie Elhadad. 2015. 'Intelligible Models for HealthCare: Predicting Pneumonia Risk and Hospital 30-Day Readmission'. In *Proceedings of the 21th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, 1721–30. KDD '15. New York, NY, USA: Association for Computing Machinery. https://doi.org/10.1145/2783258.2788613.

Cassidy, Sophie, Josephine Y. Chau, Michael Catt, Adrian Bauman, and Michael I. Trenell. 2016. 'Cross-Sectional Study of Diet, Physical Activity, Television Viewing and Sleep Duration in 233 110 Adults from the UK Biobank; the Behavioural Phenotype of Cardiovascular Disease and Type 2 Diabetes'. *BMJ Open* 6 (3): e010038. https://doi.org/10.1136/bmjopen-2015-010038.

Cezard, G., F. Sullivan, and K. Keenan. 2022. 'Understanding Multimorbidity Trajectories in Scotland Using Sequence Analysis'. *Scientific Reports* 12 (1): 16485. https://doi.org/10.1038/s41598-022-20546-4.

Chandak, Payal, Kexin Huang, and Marinka Zitnik. 2022. 'Building a Knowledge Graph to Enable Precision Medicine'. https://www.biorxiv.org/content/10.1101/2022.05.01.489928v2.

Chang, Winston, Joe Cheng, J. J. Allaire, Carson Sievert, Barret Schloerke, Yihui Xie, Jeff Allen, Jonathan McPherson, Alan Dipert, and Barbara Borges. 2023. *Shiny: Web Application Framework for R*. https://shiny.rstudio.com/.

Chen, Dijun, Liang-Yu Fu, Dahui Hu, Christian Klukas, Ming Chen, and Kerstin Kaufmann. 2018. 'The HTPmod Shiny Application Enables Modeling and Visualization of Large-Scale Biological Data'. *Communications Biology* 1 (1): 1–8. https://doi.org/10.1038/s42003-018-0091-x.

Chen, Edward Y., Christopher M. Tan, Yan Kou, Qiaonan Duan, Zichen Wang, Gabriela Vaz Meirelles, Neil R. Clark, and Avi Ma'ayan. 2013. 'Enrichr: Interactive and Collaborative HTML5 Gene List Enrichment Analysis Tool'. *BMC Bioinformatics* 14 (1): 128. https://doi.org/10.1186/1471-2105-14-128.

Chen, James C., Mariano J. Alvarez, Flaminia Talos, Harshil Dhruv, Gabrielle E. Rieckhof, Archana Iyer, Kristin L. Diefes, et al. 2014. 'Identification of Causal Genetic Drivers of Human Disease through Systems-Level Analysis of Regulatory Networks'. *Cell* 159 (2): 402–14. https://doi.org/10.1016/j.cell.2014.09.021.

Chen, Lu, Bing Ge, Francesco Paolo Casale, Louella Vasquez, Tony Kwan, Diego Garrido-Martín, Stephen Watt, et al. 2016. 'Genetic Drivers of Epigenetic and Transcriptional Variation in Human Immune Cells'. *Cell* 167 (5): 1398-1414.e24. https://doi.org/10.1016/j.cell.2016.10.026.

Chen, Po-Hsuan Cameron, Yun Liu, and Lily Peng. 2019. 'How to Develop Machine Learning Models for Healthcare'. *Nature Materials* 18 (5): 410–14. https://doi.org/10.1038/s41563-019-0345-0.

Chen, Tianqi, and Carlos Guestrin. 2016. 'XGBoost: A Scalable Tree Boosting System'. In *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, 785–94. https://doi.org/10.1145/2939672.2939785.

Chen, Xingdong, Jeffrey Gole, Athurva Gore, Qiye He, Ming Lu, Jun Min, Ziyu Yuan, et al. 2020. 'Non-Invasive Early Detection of Cancer Four Years before Conventional Diagnosis Using a Blood Test'. *Nature Communications* 11 (1): 3475. https://doi.org/10.1038/s41467-020-17316-z.

Chowdhury, Saifur Rahman, Dipak Chandra Das, Tachlima Chowdhury Sunna, Joseph Beyene, and Ahmed Hossain. 2023. 'Global and Regional Prevalence of Multimorbidity in the Adult Population in Community Settings: A Systematic Review

and Meta-Analysis'. *EClinicalMedicine* 57 (March). https://doi.org/10.1016/j.eclinm.2023.101860.

Cohen, Alan A., Luigi Ferrucci, Tamàs Fülöp, Dominique Gravel, Nan Hao, Andres Kriete, Morgan E. Levine, et al. 2022. 'A Complex Systems Approach to Aging Biology'. *Nature Aging* 2 (7): 580–91. https://doi.org/10.1038/s43587-022-00252-6.

Cohen, Mitchell J., Karim Brohi, Carolyn S. Calfee, Pamela Rahn, Brian B. Chesebro, Sarah C. Christiaans, Michel Carles, Marybeth Howard, and Jean-François Pittet. 2009. 'Early Release of High Mobility Group Box Nuclear Protein 1 after Severe Trauma in Humans: Role of Injury Severity and Tissue Hypoperfusion'. *Critical Care (London, England)* 13 (6): R174. https://doi.org/10.1186/cc8152.

Collaborative, COVIDSurg, and GlobalSurg Collaborative. 2021. 'Timing of Surgery Following SARS-CoV-2 Infection: An International Prospective Cohort Study'. *Anaesthesia* 76 (6): 748–58. https://doi.org/10.1111/anae.15458.

Collins, Gary S, and Karel G M Moons. 2019. 'Reporting of Artificial Intelligence Prediction Models'. *The Lancet* 393 (10181): 1577–79. https://doi.org/10.1016/S0140-6736(19)30037-6.

Connally, Noah, Sumaiya Nazeen, Daniel Lee, Huwenbo Shi, John Stamatoyannopoulos, Sung Chun, Chris Cotsapas, Christopher Cassa, and Shamil Sunyaev. 2021. 'The Missing Link between Genetic Association and Regulatory Function'. medRxiv. https://doi.org/10.1101/2021.06.08.21258515.

Consortium, The GTEx. 2015. 'The Genotype-Tissue Expression (GTEx) Pilot Analysis: Multitissue Gene Regulation in Humans'. *Science* 348 (6235): 648–60. https://doi.org/10.1126/science.1262110.

'CovidSurg – Globalsurg'. n.d. Accessed 14 March 2023. https://globalsurg.org/covidsurg/.

COVIDSurg Collaborative. 2020. 'Mortality and Pulmonary Complications in Patients Undergoing Surgery with Perioperative SARS-CoV-2 Infection: An International Cohort Study'. *Lancet (London, England)* 396 (10243): 27–38. https://doi.org/10.1016/S0140-6736(20)31182-X.

———. 2021. 'Machine Learning Risk Prediction of Mortality for Patients Undergoing Surgery with Perioperative SARS-CoV-2: The COVIDSurg Mortality Score'. *British Journal of Surgery* 108 (11): 1274–92. https://doi.org/10.1093/bjs/znab183.

———. 2022. 'Projecting COVID-19 Disruption to Elective Surgery'. *Lancet (London, England)* 399 (10321): 233–34. https://doi.org/10.1016/S0140-6736(21)02836-1.

Culos, Anthony, Amy S. Tsai, Natalie Stanley, Martin Becker, Mohammad S. Ghaemi, David R. McIlwain, Ramin Fallahzadeh, et al. 2020. 'Integration of Mechanistic Immunological Knowledge into a Machine Learning Pipeline Improves Predictions'. *Nature Machine Intelligence* 2 (10): 619–28. https://doi.org/10.1038/s42256-020-00232-8.

'Data_Exploration'. n.d. Accessed 11 February 2021. https://suryaprakash.shinyapps.io/Data_Exploration/.

Davis, Mark M, Cristina M Tato, and David Furman. 2017. 'Systems Immunology: Just Getting Started'. *Nature Immunology* 18 (7): 725–32. https://doi.org/10.1038/ni.3768.

De Jager, Philip L., Nir Hacohen, Diane Mathis, Aviv Regev, Barbara E. Stranger, and Christophe Benoist. 2015. 'ImmVar Project: Insights and Design Considerations for Future Studies of "Healthy" Immune Variation'. *Seminars in Immunology*, Global transcriptional regulation in the immune system, 27 (1): 51–57. https://doi.org/10.1016/j.smim.2015.03.003.

Debray, Thomas P. A., Yvonne Vergouwe, Hendrik Koffijberg, Daan Nieboer, Ewout W. Steyerberg, and Karel G. M. Moons. 2015. 'A New Framework to Enhance the Interpretation of External Validation Studies of Clinical Prediction Models'. *Journal of Clinical Epidemiology* 68 (3): 279–89. https://doi.org/10.1016/j.jclinepi.2014.06.018.

Deelen, Joris, Marian Beekman, Hae-Won Uh, Linda Broer, Kristin L. Ayers, Qihua Tan, Yoichiro Kamatani, et al. 2014. 'Genome-Wide Association Meta-Analysis of Human Longevity Identifies a Novel Locus Conferring Survival beyond 90 Years of Age'. *Human Molecular Genetics* 23 (16): 4420–32. https://doi.org/10.1093/hmg/ddu139.

Dhiman, Paula, Jie Ma, Constanza L. Andaur Navarro, Benjamin Speich, Garrett Bullock, Johanna A. A. Damen, Lotty Hooft, et al. 2022. 'Risk of Bias of Prognostic Models Developed Using Machine Learning: A Systematic Review in Oncology'. *Diagnostic and Prognostic Research* 6 (1): 13. https://doi.org/10.1186/s41512-022-00126-w.

Dixon, Emmanuel D., Alexander D. Nardo, Thierry Claudel, and Michael Trauner. 2021. 'The Role of Lipid Sensing Nuclear Receptors (PPARs and LXR) and Metabolic Lipases in Obesity, Diabetes and NAFLD'. *Genes* 12 (5): 645. https://doi.org/10.3390/genes12050645.

Dönertaş, Handan Melike, Daniel K. Fabian, Matías Fuentealba, Linda Partridge, and Janet M. Thornton. 2021. 'Common Genetic Associations between Age-Related Diseases'. *Nature Aging* 1 (4): 400–412. https://doi.org/10.1038/s43587-021-00051-5.

Dong, Guiying, Jianfeng Feng, Fengzhu Sun, Jingqi Chen, and Xing-Ming Zhao. 2021. 'A Global Overview of Genetically Interpretable Multimorbidities among Common Diseases in the UK Biobank'. *Genome Medicine* 13 (1): 110. https://doi.org/10.1186/s13073-021-00927-6.

Doshi-Velez, Finale, and Been Kim. 2017. 'Towards A Rigorous Science of Interpretable Machine Learning'. *ArXiv:1702.08608 [Cs, Stat]*, March. http://arxiv.org/abs/1702.08608.

Durinck, Steffen, Paul T. Spellman, Ewan Birney, and Wolfgang Huber. 2009. 'Mapping Identifiers for the Integration of Genomic Datasets with the R/Bioconductor Package BiomaRt'. *Nature Protocols* 4 (8): 1184–91. https://doi.org/10.1038/nprot.2009.97.

E, Christodoulou, Ma J, Collins Gs, Steyerberg Ew, Verbakel Jy, and Van Calster B. 2019. 'A Systematic Review Shows No Performance Benefit of Machine Learning over Logistic Regression for Clinical Prediction Models'. *Journal of Clinical Epidemiology* 110 (June). https://doi.org/10.1016/j.jclinepi.2019.02.004.

ElHabr, r on Tony. 2020. 'Comparing Variable Importance Functions (For Modeling) | R-Bloggers'. 13 July 2020. https://www.r-bloggers.com/2020/07/comparing-variable-importance-functions-for-modeling/.

Ellinghaus, David, Luke Jostins, Sarah L. Spain, Adrian Cortes, Jörn Bethune, Buhm Han, Yu Rang Park, et al. 2016. 'Analysis of Five Chronic Inflammatory Diseases Identifies 27 New Associations and Highlights Disease-Specific Patterns at Shared Loci'. *Nature Genetics* 48 (5): 510–18. https://doi.org/10.1038/ng.3528.

Emilsson, Valur, Valborg Gudmundsdottir, Alexander Gudjonsson, Thorarinn Jonmundsson, Brynjolfur G. Jonsson, Mohd A. Karim, Marjan Ilkov, et al. 2022. 'Coding and Regulatory Variants Are Associated with Serum Protein Levels and Disease'. *Nature Communications* 13 (1): 481. https://doi.org/10.1038/s41467-022-28081-6.

Emilsson, Valur, Marjan Ilkov, John R. Lamb, Nancy Finkel, Elias F. Gudmundsson, Rebecca Pitts, Heather Hoover, et al. 2018. 'Co-Regulatory Networks of Human Serum Proteins Link Genetics to Disease'. *Science* 361 (6404): 769–73. https://doi.org/10.1126/science.aaq1327.

ENCODE Project Consortium. 2012. 'An Integrated Encyclopedia of DNA Elements in the Human Genome'. *Nature* 489 (7414): 57–74. https://doi.org/10.1038/nature11247.

'Extending the Reach of the Lasso and Elastic Net Penalties: Methodology and Practice - ProQuest'. n.d. Accessed 21 February 2023. https://www.proquest.com/openview/bdcba6c113b0c33e98131e303fbf3054/1?pq-origsite=gscholar&cbl=18750&diss=y.

Fabbri, Elisa, Yang An, Marco Zoli, Eleanor M. Simonsick, Jack M. Guralnik, Stefania Bandinelli, Cynthia M. Boyd, and Luigi Ferrucci. 2015. 'Aging and the Burden of Multimorbidity: Associations With Inflammatory and Anabolic Hormonal Biomarkers'. *The Journals of Gerontology: Series A* 70 (1): 63–70. https://doi.org/10.1093/gerona/glu127.

Fabian, Daniel K., Matías Fuentealba, Handan Melike Dönertaş, Linda Partridge, and Janet M. Thornton. 2021. 'Functional Conservation in Genes and Pathways Linking Ageing and Immunity'. *Immunity & Ageing* 18 (1): 23. https://doi.org/10.1186/s12979-021-00232-1.

Fairfax, Benjamin P, and Julian C Knight. 2014. 'Genetics of Gene Expression in Immunity to Infection'. *Current Opinion in Immunology*, Immunogenetics and transplantation * Special section: Effects of endogenous immune stimulants, 30 (October): 63–71. https://doi.org/10.1016/j.coi.2014.07.001.

Farh, Kyle Kai-How, Alexander Marson, Jiang Zhu, Markus Kleinewietfeld, William J. Housley, Samantha Beik, Noam Shoresh, et al. 2015. 'Genetic and Epigenetic Fine Mapping of Causal Autoimmune Disease Variants'. *Nature* 518 (7539): 337–43. https://doi.org/10.1038/nature13835.

Fawcett, Tom. 2006. 'An Introduction to ROC Analysis'. *Pattern Recognition Letters*, ROC Analysis in Pattern Recognition, 27 (8): 861–74. https://doi.org/10.1016/j.patrec.2005.10.010.

Feng, Jinwen, Chen Ding, Naiqi Qiu, Xiaotian Ni, Dongdong Zhan, Wanlin Liu, Xia Xia, et al. 2017. 'Firmiana: Towards a One-Stop Proteomic Cloud Platform for Data Processing and Analysis'. *Nature Biotechnology* 35 (5): 409–12. https://doi.org/10.1038/nbt.3825.

Fernández-Torras, Adrià, Miquel Duran-Frigola, Martino Bertoni, Martina Locatelli, and Patrick Aloy. 2022. 'Integrating and Formatting Biomedical Data as Pre-Calculated Knowledge Graph Embeddings in the Bioteque'. *Nature Communications* 13 (1): 5304. https://doi.org/10.1038/s41467-022-33026-0.

Finlayson, Samuel G., Andrew L. Beam, and Maarten van Smeden. 2023. 'Machine Learning and Statistics in Clinical Research Articles—Moving Past the False Dichotomy'. *JAMA Pediatrics*, March. https://doi.org/10.1001/jamapediatrics.2023.0034.

Fortney, Kristen, Edgar Dobriban, Paolo Garagnani, Chiara Pirazzini, Daniela Monti, Daniela Mari, Gil Atzmon, et al. 2015. 'Genome-Wide Scan Informed by Age-Related Disease Identifies Loci for Exceptional Human Longevity'. *PLOS Genetics* 11 (12): e1005728. https://doi.org/10.1371/journal.pgen.1005728.

Fotouhi, Babak, Naghmeh Momeni, Maria A. Riolo, and David L. Buckeridge. 2018. 'Statistical Methods for Constructing Disease Comorbidity Networks from Longitudinal Inpatient Data'. *Applied Network Science* 3 (1): 1–34. https://doi.org/10.1007/s41109-018-0101-4.

Fowler, Joseph Wayne M, Rong Zhang, Bo Tao, Nabil E Boutagy, and William C Sessa. 2022. 'Inflammatory Stress Signaling via NF-KB Alters Accessible Cholesterol to Upregulate SREBP2 Transcriptional Activity in Endothelial Cells'. Edited by Edward A Fisher and Mone Zaidi. *ELife* 11 (August): e79529. https://doi.org/10.7554/eLife.79529.

Franceschi, Claudio, Massimiliano Bonafè, Silvana Valensin, Fabiola Olivieri, Maria De Luca, Enzo Ottaviani, and Giovanna De Benedictis. 2000. 'Inflamm-Aging: An Evolutionary Perspective on Immunosenescence'. *Annals of the New York Academy of Sciences* 908 (1): 244–54. https://doi.org/10.1111/j.1749-6632.2000.tb06651.x.

Fraser, Helen C., Valerie Kuan, Ronja Johnen, Magdalena Zwierzyna, Aroon D. Hingorani, Andreas Beyer, and Linda Partridge. 2022. 'Biological Mechanisms of Aging Predict Age-Related Disease Co-Occurrence in Patients'. *Aging Cell* 21 (4): e13524. https://doi.org/10.1111/acel.13524.

Friedman, Jerome, Trevor Hastie, and Robert Tibshirani. 2010. ‘Regularization Paths for Generalized Linear Models via Coordinate Descent’. *Journal of Statistical Software* 33 (1): 1–22. https://doi.org/10.18637/jss.v033.i01.

Furman, David, Judith Campisi, Eric Verdin, Pedro Carrera-Bastos, Sasha Targ, Claudio Franceschi, Luigi Ferrucci, et al. 2019. ‘Chronic Inflammation in the Etiology of Disease across the Life Span’. *Nature Medicine* 25 (12): 1822–32. https://doi.org/10.1038/s41591-019-0675-0.

Gamazon, Eric R., Ayellet V. Segrè, Martijn van de Bunt, Xiaoquan Wen, Hualin S. Xi, Farhad Hormozdiari, Halit Ongen, et al. 2018. ‘Using an Atlas of Gene Regulation across 44 Human Tissues to Inform Complex Disease- and Trait-Associated Variation’. *Nature Genetics* 50 (7): 956–67. https://doi.org/10.1038/s41588-018-0154-4.

Gao, Xiaoyi Raymond, Marion Chiariglione, Ke Qin, Karen Nuytemans, Douglas W. Scharre, Yi-Ju Li, and Eden R. Martin. 2023. ‘Explainable Machine Learning Aggregates Polygenic Risk Scores and Electronic Health Records for Alzheimer’s Disease Prediction’. *Scientific Reports* 13 (1): 450. https://doi.org/10.1038/s41598-023-27551-1.

Gao, Yue, Shipeng Shang, Shuang Guo, Xinyue Wang, Hanxiao Zhou, Yue Sun, Jing Gan, et al. 2022. ‘AgingBank: A Manually Curated Knowledgebase and High-Throughput Analysis Platform That Provides Experimentally Supported Multi-Omics Data Relevant to Aging in Multiple Species’. *Briefings in Bioinformatics* 23 (6): bbac438. https://doi.org/10.1093/bib/bbac438.

Gerin, Isabelle, Vernon W. Dolinsky, Jonathan G. Shackman, Robert T. Kennedy, Shian-Huey Chiang, Charles F. Burant, Knut R. Steffensen, Jan-Åke Gustafsson, and Ormond A. MacDougald. 2005. ‘LXRβ Is Required for Adipocyte Growth, Glucose Homeostasis, and β Cell Function *’. *Journal of Biological Chemistry* 280 (24): 23024–31. https://doi.org/10.1074/jbc.M412564200.

Giannoula, Alexia, Emilio Centeno, Miguel-Angel Mayer, Ferran Sanz, and Laura I. Furlong. 2021. ‘A System-Level Analysis of Patient Disease Trajectories Based on Clinical, Phenotypic and Molecular Similarities’. *Bioinformatics (Oxford, England)* 37 (10): 1435–43. https://doi.org/10.1093/bioinformatics/btaa964.

Giannoula, Alexia, Alba Gutierrez-Sacristán, Álex Bravo, Ferran Sanz, and Laura I. Furlong. 2018. ‘Identifying Temporal Patterns in Patient Disease Trajectories Using Dynamic Time Warping: A Population-Based Study’. *Scientific Reports* 8 (1): 1–14. https://doi.org/10.1038/s41598-018-22578-1.

Gibbs, Richard A. 2020. ‘The Human Genome Project Changed Everything’. *Nature Reviews Genetics* 21 (10): 575–76. https://doi.org/10.1038/s41576-020-0275-3.

Gibbs, Richard A., John W. Belmont, Paul Hardenbol, Thomas D. Willis, Fuli Yu, Huanming Yang, Lan-Yang Ch’ang, et al. 2003. ‘The International HapMap Project’. *Nature* 426 (6968): 789–96. https://doi.org/10.1038/nature02168.

Glasbey, James, Adesoji Ademuyiwa, Adewale Adisa, Ehab AlAmeer, Alexis P. Arnaud, Faris Ayasra, José Azevedo, et al. 2021. ‘Effect of COVID-19 Pandemic Lockdowns on Planned Cancer Surgery for 15 Tumour Types in 61 Countries: An International, Prospective, Cohort Study’. *The Lancet Oncology* 22 (11): 1507–17. https://doi.org/10.1016/S1470-2045(21)00493-9.

Glasbey, James C., Dmitri Nepogodiev, Joana F. F. Simoes, Omar Omar, Elizabeth Li, Mary L. Venn, null Pgdme, et al. 2021. ‘Elective Cancer Surgery in COVID-19-Free Surgical Pathways During the SARS-CoV-2 Pandemic: An International, Multicenter, Comparative Cohort Study’. *Journal of Clinical Oncology: Official Journal of the American Society of Clinical Oncology* 39 (1): 66–78. https://doi.org/10.1200/JCO.20.01933.

Goh, Kwang-Il, Michael E. Cusick, David Valle, Barton Childs, Marc Vidal, and Albert-László Barabási. 2007. 'The Human Disease Network'. *Proceedings of the National Academy of Sciences* 104 (21): 8685–90. https://doi.org/10.1073/pnas.0701361104.

Goorbergh, Ruben van den, Maarten van Smeden, Dirk Timmerman, and Ben Van Calster. 2022. 'The Harm of Class Imbalance Corrections for Risk Prediction Models: Illustration and Simulation Using Logistic Regression'. *Journal of the American Medical Informatics Association* 29 (9): 1525–34. https://doi.org/10.1093/jamia/ocac093.

GTEx Consortium. 2020. 'The GTEx Consortium Atlas of Genetic Regulatory Effects across Human Tissues'. *Science (New York, N.Y.)* 369 (6509): 1318–30. https://doi.org/10.1126/science.aaz1776.

Gupta, Rishi K., Ewen M. Harrison, Antonia Ho, Annemarie B. Docherty, Stephen R. Knight, Maarten van Smeden, Ibrahim Abubakar, et al. 2021. 'Development and Validation of the ISARIC 4C Deterioration Model for Adults Hospitalised with COVID-19: A Prospective Cohort Study'. *The Lancet Respiratory Medicine* 9 (4): 349–59. https://doi.org/10.1016/S2213-2600(20)30559-2.

Hanlon, Peter, Bhautesh D. Jani, Barbara Nicholl, Jim Lewsey, David A. McAllister, and Frances S. Mair. 2022. 'Associations between Multimorbidity and Adverse Health Outcomes in UK Biobank and the SAIL Databank: A Comparison of Longitudinal Cohort Studies'. *PLOS Medicine* 19 (3): e1003931. https://doi.org/10.1371/journal.pmed.1003931.

Hannum, Gregory, Justin Guinney, Ling Zhao, Li Zhang, Guy Hughes, SriniVas Sadda, Brandy Klotzle, et al. 2013. 'Genome-Wide Methylation Profiles Reveal Quantitative Views of Human Aging Rates'. *Molecular Cell* 49 (2): 359–67. https://doi.org/10.1016/j.molcel.2012.10.016.

Harold, Denise, Richard Abraham, Paul Hollingworth, Rebecca Sims, Amy Gerrish, Marian L. Hamshere, Jaspreet Singh Pahwa, et al. 2009. 'Genome-Wide Association Study Identifies Variants at CLU and PICALM Associated with Alzheimer's Disease'. *Nature Genetics* 41 (10): 1088–93. https://doi.org/10.1038/ng.440.

Hastie, Trevor, Jerome Friedman, and Robert Tibshirani. 2001. *The Elements of Statistical Learning*. Springer Series in Statistics. New York, NY: Springer. https://doi.org/10.1007/978-0-387-21606-5.

Hastie, Trevor, Robert Tibshirani, and Jerome Friedman. 2009. *The Elements of Statistical Learning*. Springer Series in Statistics. New York, NY: Springer. https://doi.org/10.1007/978-0-387-84858-7.

Haug, Nils, Johannes Sorger, Teresa Gisinger, Michael Gyimesi, Alexandra Kautzky-Willer, Stefan Thurner, and Peter Klimek. 2021. 'Decompression of Multimorbidity Along the Disease Trajectories of Diabetes Mellitus Patients'. *Frontiers in Physiology* 11. https://www.frontiersin.org/articles/10.3389/fphys.2020.612604.

Hazeldine, Jon, David N. Naumann, Emma Toman, David Davies, Jonathan R. B. Bishop, Zhangjie Su, Peter Hampson, et al. 2017. 'Prehospital Immune Responses and Development of Multiple Organ Dysfunction Syndrome Following Traumatic Injury: A Prospective Cohort Study'. *PLoS Medicine* 14 (7): e1002338. https://doi.org/10.1371/journal.pmed.1002338.

Hernansaiz-Ballesteros, Rosa, Christian H. Holland, Aurelien Dugourd, and Julio Saez-Rodriguez. 2021. 'FUNKI: Interactive Functional Footprint-Based Analysis of Omics Data'. arXiv. https://doi.org/10.48550/arXiv.2109.05796.

Hidalgo, César A., Nicholas Blumm, Albert-László Barabási, and Nicholas A. Christakis. 2009. 'A Dynamic Network Approach for the Study of Human Phenotypes'. *PLOS Computational Biology* 5 (4): e1000353. https://doi.org/10.1371/journal.pcbi.1000353.

Hietbrink, Falco, Leo Koenderman, Martje Althuizen, and Luke P. H. Leenen. 2009. 'Modulation of the Innate Immune Response after Trauma Visualised by a Change in Functional PMN Phenotype'. *Injury* 40 (8): 851–55. https://doi.org/10.1016/j.injury.2008.11.002.

Ho, Tin Kam. 1995. 'Random Decision Forests'. In *Proceedings of 3rd International Conference on Document Analysis and Recognition*, 1:278–82 vol.1. https://doi.org/10.1109/ICDAR.1995.598994.

Holzscheck, Nicholas, Cassandra Falckenhayn, Jörn Söhle, Boris Kristof, Ralf Siegner, André Werner, Janka Schössow, et al. 2021. 'Modeling Transcriptomic Age Using Knowledge-Primed Artificial Neural Networks'. *Npj Aging and Mechanisms of Disease* 7 (1). https://doi.org/10.1038/s41514-021-00068-5.

Hood, Leroy, and Mauricio Flores. 2012. 'A Personal View on Systems Medicine and the Emergence of Proactive P4 Medicine: Predictive, Preventive, Personalized and Participatory'. *New Biotechnology* 29 (6): 613–24. https://doi.org/10.1016/j.nbt.2012.03.004.

Hormozdiari, Farhad, Steven Gazal, Bryce van de Geijn, Hilary K. Finucane, Chelsea J.-T. Ju, Po-Ru Loh, Armin Schoech, et al. 2018. 'Leveraging Molecular Quantitative Trait Loci to Understand the Genetic Architecture of Diseases and Complex Traits'. *Nature Genetics* 50 (7): 1041–47. https://doi.org/10.1038/s41588-018-0148-2.

Horvath, Steve. 2013. 'DNA Methylation Age of Human Tissues and Cell Types'. *Genome Biology* 14 (10). https://doi.org/10.1186/gb-2013-14-10-r115.

Hu, Jessica Xin, Cecilia Engel Thomas, and Søren Brunak. 2016. 'Network Biology Concepts in Complex Disease Comorbidities'. *Nature Reviews Genetics* 17 (10): 615–29. https://doi.org/10.1038/nrg.2016.87.

Huang, Yingxiang, Wentao Li, Fima Macheret, Rodney A. Gabriel, and Lucila Ohno-Machado. 2020. 'A Tutorial on Calibration Measurements and Calibration Models for Clinical Prediction Models'. *Journal of the American Medical Informatics Association* 27 (4): 621–33. https://doi.org/10.1093/jamia/ocz228.

Huber-Lang, Markus, John D. Lambris, and Peter A. Ward. 2018. 'Innate Immune Responses to Trauma'. *Nature Immunology* 19 (4): 327. https://doi.org/10.1038/s41590-018-0064-8.

Ito, Ayaka, Cynthia Hong, Xin Rong, Xuewei Zhu, Elizabeth J. Tarling, Per Niklas Hedde, Enrico Gratton, John Parks, and Peter Tontonoz. 2015. 'LXRs Link Metabolism to Inflammation through Abca1-Dependent Regulation of Membrane Composition and TLR Signaling'. *ELife* 4 (July): e08009. https://doi.org/10.7554/eLife.08009.

Jagadeesh, Karthik A., Kushal K. Dey, Daniel T. Montoro, Rahul Mohan, Steven Gazal, Jesse M. Engreitz, Ramnik J. Xavier, Alkes L. Price, and Aviv Regev. 2022. 'Identifying Disease-Critical Cell Types and Cellular Processes by Integrating Single-Cell RNA-Sequencing and Human Genetics'. *Nature Genetics* 54 (10): 1479–92. https://doi.org/10.1038/s41588-022-01187-9.

Jakobsen, Janus Christian, Christian Gluud, Jørn Wetterslev, and Per Winkel. 2017. 'When and How Should Multiple Imputation Be Used for Handling Missing Data in Randomised Clinical Trials – a Practical Guide with Flowcharts'. *BMC Medical Research Methodology* 17 (1): 162. https://doi.org/10.1186/s12874-017-0442-1.

Jalali, Mohammad S., Catherine DiGennaro, and Devi Sridhar. 2020. 'Transparency Assessment of COVID-19 Models'. *The Lancet Global Health* 0 (0). https://doi.org/10.1016/S2214-109X(20)30447-2.

Jensen, Anders Boeck, Pope L. Moseley, Tudor I. Oprea, Sabrina Gade Ellesøe, Robert Eriksson, Henriette Schmock, Peter Bjødstrup Jensen, Lars Juhl Jensen, and Søren Brunak. 2014. 'Temporal Disease Trajectories Condensed from Population-Wide

Registry Data Covering 6.2 Million Patients'. *Nature Communications* 5 (1): 4022. https://doi.org/10.1038/ncomms5022.

Joshi, Peter K., Krista Fischer, Katharina E. Schraut, Harry Campbell, Tõnu Esko, and James F. Wilson. 2016. 'Variants near CHRNA3/5 and APOE Have Age- and Sex-Related Effects on Human Lifespan'. *Nature Communications* 7 (1): 11174. https://doi.org/10.1038/ncomms11174.

Kaczorowski, Kevin J., Karthik Shekhar, Dieudonné Nkulikiyimfura, Cornelia L. Dekker, Holden Maecker, Mark M. Davis, Arup K. Chakraborty, and Petter Brodin. 2017. 'Continuous Immunotypes Describe Human Immune Variation and Predict Diverse Responses'. *Proceedings of the National Academy of Sciences* 114 (30): E6097–6106. https://doi.org/10.1073/pnas.1705065114.

Kamat, Mihir A., James A. Blackshaw, Robin Young, Praveen Surendran, Stephen Burgess, John Danesh, Adam S. Butterworth, and James R. Staley. 2019. 'PhenoScanner V2: An Expanded Tool for Searching Human Genotype-Phenotype Associations'. *Bioinformatics (Oxford, England)* 35 (22): 4851–53. https://doi.org/10.1093/bioinformatics/btz469.

Kanehisa, M., and S. Goto. 2000. 'KEGG: Kyoto Encyclopedia of Genes and Genomes'. *Nucleic Acids Research* 28 (1): 27–30. https://doi.org/10.1093/nar/28.1.27.

Kang, Chanhee, Qikai Xu, Timothy D. Martin, Mamie Z. Li, Marco Demaria, Liviu Aron, Tao Lu, Bruce A. Yankner, Judith Campisi, and Stephen J. Elledge. 2015. 'The DNA Damage Response Induces Inflammation and Senescence by Inhibiting Autophagy of GATA4'. *Science (New York, N.Y.)* 349 (6255): aaa5612. https://doi.org/10.1126/science.aaa5612.

Kennedy, Brian K., Shelley L. Berger, Anne Brunet, Judith Campisi, Ana Maria Cuervo, Elissa S. Epel, Claudio Franceschi, et al. 2014a. 'Aging: A Common Driver of Chronic Diseases and a Target for Novel Interventions'. *Cell* 159 (4): 709–13. https://doi.org/10.1016/j.cell.2014.10.039.

———. 2014b. 'Geroscience: Linking Aging to Chronic Disease'. *Cell* 159 (4): 709–13. https://doi.org/10.1016/j.cell.2014.10.039.

Kerimov, Nurlan, James D. Hayhurst, Kateryna Peikova, Jonathan R. Manning, Peter Walter, Liis Kolberg, Marija Samoviča, et al. 2021. 'A Compendium of Uniformly Processed Human Gene Expression and Splicing Quantitative Trait Loci'. *Nature Genetics* 53 (9): 1290–99. https://doi.org/10.1038/s41588-021-00924-w.

Kibbe, Melina R. 2020. 'Surgery and COVID-19'. *JAMA* 324 (12): 1151–52. https://doi.org/10.1001/jama.2020.15191.

Kim, Seung-Soo, Adam D. Hudgins, Brenda Gonzalez, Sofiya Milman, Nir Barzilai, Jan Vijg, Zhidong Tu, and Yousin Suh. 2021. 'A Compendium of Age-Related PheWAS and GWAS Traits for Human Genetic Association Studies, Their Networks and Genetic Correlations'. *Frontiers in Genetics* 12. https://www.frontiersin.org/articles/10.3389/fgene.2021.680560.

Kirkwood, Thomas B. L. 2005. 'Understanding the Odd Science of Aging'. *Cell* 120 (4): 437–47. https://doi.org/10.1016/j.cell.2005.01.027.

Kitano, Hiroaki. 2002. 'Computational Systems Biology'. *Nature* 420 (6912): 206–10. https://doi.org/10.1038/nature01254.

Knight, Stephen R., Antonia Ho, Riinu Pius, Iain Buchan, Gail Carson, Thomas M. Drake, Jake Dunning, et al. 2020. 'Risk Stratification of Patients Admitted to Hospital with Covid-19 Using the ISARIC WHO Clinical Characterisation Protocol: Development and Validation of the 4C Mortality Score'. *BMJ* 370 (September): m3339. https://doi.org/10.1136/bmj.m3339.

Kotas, Maya E., and Ruslan Medzhitov. 2015. 'Homeostasis, Inflammation, and Disease Susceptibility'. *Cell* 160 (5): 816–27. https://doi.org/10.1016/j.cell.2015.02.010.

Kou, Hao-Wei, Chih-Hua Yeh, Hsin-I. Tsai, Chih-Chieh Hsu, Yi-Chung Hsieh, Wei-Ting Chen, Hao-Tsai Cheng, Ming-Chin Yu, and Chao-Wei Lee. 2019. 'Sarcopenia Is an Effective Predictor of Difficult-to-Wean and Mortality among Critically Ill Surgical Patients'. *PLOS ONE* 14 (8): e0220699. https://doi.org/10.1371/journal.pone.0220699.

Krispin, Rami. (2017) 2022. 'MLstudio'. R. https://github.com/RamiKrispin/MLstudio.

Kuan Po Ai, Valerie. 2020. 'Exploring the Relationship between Age and Health Conditions Using Electronic Health Records: From Single Diseases to Multimorbidities'. *Doctoral Thesis, UCL (University College London).* Doctoral, UCL (University College London). https://discovery.ucl.ac.uk/id/eprint/10094022/.

Kuan, Valerie, Spiros Denaxas, Praveetha Patalay, Dorothea Nitsch, Rohini Mathur, Arturo Gonzalez-Izquierdo, Reecha Sofat, et al. 2023. 'Identifying and Visualising Multimorbidity and Comorbidity Patterns in Patients in the English National Health Service: A Population-Based Study'. *The Lancet Digital Health* 5 (1): e16–27. https://doi.org/10.1016/S2589-7500(22)00187-X.

Kubben, Nard, and Tom Misteli. 2017. 'Shared Molecular and Cellular Mechanisms of Premature Ageing and Ageing-Associated Diseases'. *Nature Reviews Molecular Cell Biology* 18 (10): 595–609. https://doi.org/10.1038/nrm.2017.68.

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

Kundaje, Anshul, Wouter Meuleman, Jason Ernst, Misha Bilenky, Angela Yen, Alireza Heravi-Moussavi, Pouya Kheradpour, et al. 2015. 'Integrative Analysis of 111 Reference Human Epigenomes'. *Nature* 518 (7539): 317–30. https://doi.org/10.1038/nature14248.

Kuo, Chia-Ling, Luke C. Pilling, Zuyun Liu, Janice L. Atkins, and Morgan E. Levine. 2021. 'Genetic Associations for Two Biological Age Measures Point to Distinct Aging Phenotypes'. *Aging Cell* 20 (6): e13376. https://doi.org/10.1111/acel.13376.

Kuo, Pei-Lun, Jennifer A. Schrack, Morgan E. Levine, Michelle D. Shardell, Eleanor M. Simonsick, Chee W. Chia, Ann Zenobia Moore, et al. 2022. 'Longitudinal Phenotypic Aging Metrics in the Baltimore Longitudinal Study of Aging'. *Nature Aging* 2 (7): 635–43. https://doi.org/10.1038/s43587-022-00243-7.

Kusnadi, Anthony, Sung Ho Park, Ruoxi Yuan, Tania Pannellini, Eugenia Giannopoulou, David Oliver, Theresa Lu, Kyung-Hyun Park-Min, and Lionel B. Ivashkiv. 2019. 'The Cytokine TNF Promotes Transcription Factor SREBP Activity and Binding to Inflammatory Genes to Activate Macrophages and Limit Tissue Repair'. *Immunity* 51 (2): 241-257.e9. https://doi.org/10.1016/j.immuni.2019.06.005.

Lawrence, Michael, Wolfgang Huber, Hervé Pagès, Patrick Aboyoun, Marc Carlson, Robert Gentleman, Martin T. Morgan, and Vincent J. Carey. 2013. 'Software for Computing and Annotating Genomic Ranges'. *PLOS Computational Biology* 9 (8): e1003118. https://doi.org/10.1371/journal.pcbi.1003118.

LeCun, Yann, Yoshua Bengio, and Geoffrey Hinton. 2015. 'Deep Learning'. *Nature* 521 (7553): 436–44. https://doi.org/10.1038/nature14539.

LeDell, Erin, Maya Petersen, and Mark van der Laan. 2022. 'CvAUC: Cross-Validated Area Under the ROC Curve Confidence Intervals'. https://CRAN.R-project.org/package=cvAUC.

Lee, Mark N., Chun Ye, Alexandra-Chloé Villani, Towfique Raj, Weibo Li, Thomas M. Eisenhaure, Selina H. Imboywa, et al. 2014. 'Common Genetic Variants Modulate Pathogen-Sensing Responses in Human Dendritic Cells'. *Science* 343 (6175): 1246980. https://doi.org/10.1126/science.1246980.

Lehallier, Benoit, David Gate, Nicholas Schaum, Tibor Nanasi, Song Eun Lee, Hanadie Yousef, Patricia Moran Losada, et al. 2019. 'Undulating Changes in Human Plasma

Proteome Profiles across the Lifespan'. *Nature Medicine* 25 (12): 1843–50. https://doi.org/10.1038/s41591-019-0673-2.

Leipzig, Jeremy. 2017. 'A Review of Bioinformatic Pipeline Frameworks'. *Briefings in Bioinformatics* 18 (3): 530–36. https://doi.org/10.1093/bib/bbw020.

Leslie, David, Anjali Mazumder, Aidan Peppin, Maria K. Wolters, and Alexa Hagerty. 2021. 'Does "AI" Stand for Augmenting Inequality in the Era of Covid-19 Healthcare?' *BMJ* 372 (March): n304. https://doi.org/10.1136/bmj.n304.

Levine, Morgan E., Ake T. Lu, Austin Quach, Brian H. Chen, Themistocles L. Assimes, Stefania Bandinelli, Lifang Hou, et al. 2018. 'An Epigenetic Biomarker of Aging for Lifespan and Healthspan'. *Aging* 10 (4): 573–91. https://doi.org/10.18632/aging.101414.

Li, Binglan, and Marylyn D. Ritchie. 2021. 'From GWAS to Gene: Transcriptome-Wide Association Studies and Other Methods to Functionally Understand GWAS Discoveries'. *Frontiers in Genetics* 12. https://www.frontiersin.org/articles/10.3389/fgene.2021.713230.

Li, Xia, Alexander Ploner, Yunzhang Wang, Patrik KE Magnusson, Chandra Reynolds, Deborah Finkel, Nancy L Pedersen, Juulia Jylhävä, and Sara Hägg. 2020. 'Longitudinal Trajectories, Correlations and Mortality Associations of Nine Biological Ages across 20-Years Follow-Up'. Edited by Diane M Harper, Eduardo Franco, and Alexey Moskalev. *ELife* 9 (February): e51507. https://doi.org/10.7554/eLife.51507.

Li, Xia, Alexander Ploner, Yunzhang Wang, Yiqiang Zhan, Nancy L. Pedersen, Patrik KE Magnusson, Juulia Jylhävä, and Sara Hägg. 2021. 'Clinical Biomarkers and Associations with Healthspan and Lifespan: Evidence from Observational and Genetic Data'. *EBioMedicine* 66 (April). https://doi.org/10.1016/j.ebiom.2021.103318.

Ligthart, Symen, Ahmad Vaez, Yi-Hsiang Hsu, Inflammation Working Group of the CHARGE Consortium, PMI-WG-XCP, LifeLines Cohort Study, Ronald Stolk, et al. 2016. 'Bivariate Genome-Wide Association Study Identifies Novel Pleiotropic Loci for Lipids and Inflammation'. *BMC Genomics* 17 (June): 443. https://doi.org/10.1186/s12864-016-2712-4.

Liu, Zuyun, Pei-Lun Kuo, Steve Horvath, Eileen Crimmins, Luigi Ferrucci, and Morgan Levine. 2018. 'A New Aging Measure Captures Morbidity and Mortality Risk across Diverse Subpopulations from NHANES IV: A Cohort Study'. *PLOS Medicine* 15 (12): e1002718. https://doi.org/10.1371/journal.pmed.1002718.

Lonsdale, John, Jeffrey Thomas, Mike Salvatore, Rebecca Phillips, Edmund Lo, Saboor Shad, Richard Hasz, et al. 2013. 'The Genotype-Tissue Expression (GTEx) Project'. *Nature Genetics* 45 (6): 580–85. https://doi.org/10.1038/ng.2653.

López-Otín, Carlos, Maria A. Blasco, Linda Partridge, Manuel Serrano, and Guido Kroemer. 2013. 'The Hallmarks of Aging'. *Cell* 153 (6): 1194–1217. https://doi.org/10.1016/j.cell.2013.05.039.

———. 2022. 'Hallmarks of Aging: An Expanding Universe'. *Cell*, December, S0092-8674(22)01377-0. https://doi.org/10.1016/j.cell.2022.11.001.

Lu, Ake T., Austin Quach, James G. Wilson, Alex P. Reiner, Abraham Aviv, Kenneth Raj, Lifang Hou, et al. 2019. 'DNA Methylation GrimAge Strongly Predicts Lifespan and Healthspan'. *Aging (Albany NY)* 11 (2). https://doi.org/10.18632/aging.101684.

Ma, Xiaotu, Yu Liu, Yanling Liu, Ludmil B. Alexandrov, Michael N. Edmonson, Charles Gawad, Xin Zhou, et al. 2018. 'Pan-Cancer Genome and Transcriptome Analyses of 1,699 Paediatric Leukaemias and Solid Tumours'. *Nature* 555 (7696): 371–76. https://doi.org/10.1038/nature25795.

Manolio, Teri A., Francis S. Collins, Nancy J. Cox, David B. Goldstein, Lucia A. Hindorff, David J. Hunter, Mark I. McCarthy, et al. 2009. 'Finding the Missing Heritability of Complex Diseases'. *Nature* 461 (7265): 747–53. https://doi.org/10.1038/nature08494.

Manson, Joanna, Elaine Cole, Henry D. De'Ath, Paul Vulliamy, Ute Meier, Dan Pennington, and Karim Brohi. 2016. 'Early Changes within the Lymphocyte Population Are Associated with the Development of Multiple Organ Dysfunction Syndrome in Trauma Patients'. *Critical Care* 20. https://doi.org/10.1186/s13054-016-1341-2.

Mateen, Bilal A., James Liley, Alastair K. Denniston, Chris C. Holmes, and Sebastian J. Vollmer. 2020. 'Improving the Quality of Machine Learning in Health Applications and Clinical Research'. *Nature Machine Intelligence* 2 (10): 554–56. https://doi.org/10.1038/s42256-020-00239-1.

Mayr, Viktoria D, Martin W Dünser, Veronika Greil, Stefan Jochberger, Günter Luckner, Hanno Ulmer, Barbara E Friesenecker, Jukka Takala, and Walter R Hasibeder. 2006. 'Causes of Death and Determinants of Outcome in Critically Ill Patients'. *Critical Care* 10 (6): R154. https://doi.org/10.1186/cc5086.

Mbatchou, Joelle, Leland Barnard, Joshua Backman, Anthony Marcketta, Jack A. Kosmicki, Andrey Ziyatdinov, Christian Benner, et al. 2021. 'Computationally Efficient Whole-Genome Regression for Quantitative and Binary Traits'. *Nature Genetics* 53 (7): 1097–1103. https://doi.org/10.1038/s41588-021-00870-7.

McLaren, William, Laurent Gil, Sarah E. Hunt, Harpreet Singh Riat, Graham R. S. Ritchie, Anja Thormann, Paul Flicek, and Fiona Cunningham. 2016. 'The Ensembl Variant Effect Predictor'. *Genome Biology* 17 (1): 122. https://doi.org/10.1186/s13059-016-0974-4.

Medzhitov, Ruslan. 2021. 'The Spectrum of Inflammatory Responses'. *Science* 374 (6571): 1070–75. https://doi.org/10.1126/science.abi5200.

Meizlish, Matthew L., Ruth A. Franklin, Xu Zhou, and Ruslan Medzhitov. 2021. 'Tissue Homeostasis and Inflammation'. *Annual Review of Immunology* 39 (1): 557–81. https://doi.org/10.1146/annurev-immunol-061020-053734.

Mills, Ken I., Alexander Kohlmann, P. Mickey Williams, Lothar Wieczorek, Wei-min Liu, Rachel Li, Wen Wei, et al. 2009. 'Microarray-Based Classifiers and Prognosis Models Identify Subgroups with Distinct Clinical Outcomes and High Risk of AML Transformation of Myelodysplastic Syndrome'. *Blood* 114 (5): 1063–72. https://doi.org/10.1182/blood-2008-10-187203.

Moons, Karel G.M., Douglas G. Altman, Johannes B. Reitsma, John P.A. Ioannidis, Petra Macaskill, Ewout W. Steyerberg, Andrew J. Vickers, David F. Ransohoff, and Gary S. Collins. 2015. 'Transparent Reporting of a Multivariable Prediction Model for Individual Prognosis Or Diagnosis (TRIPOD): Explanation and Elaboration'. *Annals of Internal Medicine* 162 (1): W1. https://doi.org/10.7326/M14-0698.

Munz, Matthias, Inken Wohlers, Eric Simon, Tobias Reinberger, Hauke Busch, Arne S. Schaefer, and Jeanette Erdmann. 2020. 'Qtlizer: Comprehensive QTL Annotation of GWAS Results'. *Scientific Reports* 10 (1): 20417. https://doi.org/10.1038/s41598-020-75770-7.

Murdoch, W. James, Chandan Singh, Karl Kumbier, Reza Abbasi-Asl, and Bin Yu. 2019. 'Definitions, Methods, and Applications in Interpretable Machine Learning'. *Proceedings of the National Academy of Sciences* 116 (44): 22071–80. https://doi.org/10.1073/pnas.1900654116.

Naumann, David N., Jon Hazeldine, Robert J. Dinsdale, Jon R. Bishop, Mark J. Midwinter, Paul Harrison, Sam D. Hutchings, and Janet M. Lord. 2017. 'Endotheliopathy Is Associated with Higher Levels of Cell-Free DNA Following Major Trauma: A Prospective Observational Study'. *PLOS ONE* 12 (12): e0189870. https://doi.org/10.1371/journal.pone.0189870.

Neph, Shane, M. Scott Kuehn, Alex P. Reynolds, Eric Haugen, Robert E. Thurman, Audra K. Johnson, Eric Rynes, et al. 2012. 'BEDOPS: High-Performance Genomic Feature

Operations'. *Bioinformatics* 28 (14): 1919–20. https://doi.org/10.1093/bioinformatics/bts277.

Niccoli, Teresa, and Linda Partridge. 2012. 'Ageing as a Risk Factor for Disease'. *Current Biology* 22 (17): R741–52. https://doi.org/10.1016/j.cub.2012.07.024.

Nordling, Linda. 2019. 'A Fairer Way Forward for AI in Health Care'. *Nature* 573 (7775): S103–5. https://doi.org/10.1038/d41586-019-02872-2.

Novakovsky, Gherman, Nick Dexter, Maxwell W. Libbrecht, Wyeth W. Wasserman, and Sara Mostafavi. 2022. 'Obtaining Genetics Insights from Deep Learning via Explainable Artificial Intelligence'. *Nature Reviews Genetics*, October, 1–13. https://doi.org/10.1038/s41576-022-00532-2.

Obenchain, Valerie, Michael Lawrence, Vincent Carey, Stephanie Gogarten, Paul Shannon, and Martin Morgan. 2014. 'VariantAnnotation : A Bioconductor Package for Exploration and Annotation of Genetic Variants'. *Bioinformatics* 30 (14): 2076–78. https://doi.org/10.1093/bioinformatics/btu168.

Ouzounis, Christos A., and Alfonso Valencia. 2003. 'Early Bioinformatics: The Birth of a Discipline--a Personal View'. *Bioinformatics (Oxford, England)* 19 (17): 2176–90. https://doi.org/10.1093/bioinformatics/btg309.

P, Royston, Altman Dg, and Sauerbrei W. 2006. 'Dichotomizing Continuous Predictors in Multiple Regression: A Bad Idea'. *Statistics in Medicine* 25 (1). https://doi.org/10.1002/sim.2331.

Paik, Hyojung, and Jimin Kim. 2021. 'Condensed Trajectory of the Temporal Correlation of Diseases and Mortality Extracted from over 300,000 Patients in Hospitals'. *PLOS ONE* 16 (10): e0257894. https://doi.org/10.1371/journal.pone.0257894.

Panagiotou, Orestis A., John P. A. Ioannidis, and Genome-Wide Significance Project. 2012. 'What Should the Genome-Wide Significance Threshold Be? Empirical Replication of Borderline Genetic Associations'. *International Journal of Epidemiology* 41 (1): 273–86. https://doi.org/10.1093/ije/dyr178.

Pearson-Stuttard, Jonathan, Sara Holloway, Rosie Polya, Rebecca Sloan, Linxuan Zhang, Edward W. Gregg, Katy Harrison, Jamie Elvidge, Pall Jonsson, and Thomas Porter. 2022. 'Variations in Comorbidity Burden in People with Type 2 Diabetes over Disease Duration: A Population-Based Analysis of Real World Evidence'. *EClinicalMedicine* 52 (October). https://doi.org/10.1016/j.eclinm.2022.101584.

Pietzner, Maik, Isobel D. Stewart, Johannes Raffler, Kay-Tee Khaw, Gregory A. Michelotti, Gabi Kastenmüller, Nicholas J. Wareham, and Claudia Langenberg. 2021. 'Plasma Metabolites to Profile Pathways in Noncommunicable Disease Multimorbidity'. *Nature Medicine* 27 (3): 471–79. https://doi.org/10.1038/s41591-021-01266-0.

Prados-Torres, Alexandra, Amaia Calderón-Larrañaga, Jorge Hancco-Saavedra, Beatriz Poblador-Plou, and Marjan van den Akker. 2014. 'Multimorbidity Patterns: A Systematic Review'. *Journal of Clinical Epidemiology* 67 (3): 254–66. https://doi.org/10.1016/j.jclinepi.2013.09.021.

Prentice, Philippa, Albert Koulman, Lee Matthews, Carlo L. Acerini, Ken K. Ong, and David B. Dunger. 2015. 'Lipidomic Analyses, Breast- and Formula-Feeding, and Growth in Infants'. *The Journal of Pediatrics* 166 (2): 276-281.e6. https://doi.org/10.1016/j.jpeds.2014.10.021.

Purcell, Shaun, Benjamin Neale, Kathe Todd-Brown, Lori Thomas, Manuel A. R. Ferreira, David Bender, Julian Maller, et al. 2007. 'PLINK: A Tool Set for Whole-Genome Association and Population-Based Linkage Analyses'. *American Journal of Human Genetics* 81 (3): 559–75. https://doi.org/10.1086/519795.

R Core Team. 2020. *R: A Language and Environment for Statistical Computing*. Vienna, Austria: R Foundation for Statistical Computing. https://www.R-project.org/.

‘Radiant - A Shiny App for Statistics and Machine Learning’. n.d. Accessed 28 February 2023. https://shiny.rstudio.com/gallery/radiant.html.

Rajamani, Deepa, and Manoj K. Bhasin. 2016. ‘Identification of Key Regulators of Pancreatic Cancer Progression through Multidimensional Systems-Level Analysis’. *Genome Medicine* 8 (1): 38. https://doi.org/10.1186/s13073-016-0282-3.

Rajkomar, Alvin, Jeffrey Dean, and Isaac Kohane. 2019. ‘Machine Learning in Medicine’. *New England Journal of Medicine* 380 (14): 1347–58. https://doi.org/10.1056/NEJMra1814259.

Ramos, Erin M., Douglas Hoffman, Heather A. Junkins, Donna Maglott, Lon Phan, Stephen T. Sherry, Mike Feolo, and Lucia A. Hindorff. 2014. ‘Phenotype–Genotype Integrator (PheGenI): Synthesizing Genome-Wide Association Study (GWAS) Data with Existing Genomic Resources’. *European Journal of Human Genetics* 22 (1): 144–47. https://doi.org/10.1038/ejhg.2013.96.

Rea, Irene Maeve, David S. Gibson, Victoria McGilligan, Susan E. McNerlan, H. Denis Alexander, and Owen A. Ross. 2018. ‘Age and Age-Related Diseases: Role of Inflammation Triggers and Cytokines’. *Frontiers in Immunology* 9 (April): 586. https://doi.org/10.3389/fimmu.2018.00586.

Reiter, Taylor, Phillip T Brooks†, Luiz Irber†, Shannon E K Joslin†, Charles M Reid†, Camille Scott†, C Titus Brown, and N Tessa Pierce-Ward. 2021. ‘Streamlining Data-Intensive Biology with Workflow Systems’. *GigaScience* 10 (1): giaa140. https://doi.org/10.1093/gigascience/giaa140.

Rodríguez-Pérez, Raquel, and Jürgen Bajorath. 2020. ‘Interpretation of Machine Learning Models Using Shapley Values: Application to Compound Potency and Multi-Target Activity Predictions’. *Journal of Computer-Aided Molecular Design*, May. https://doi.org/10.1007/s10822-020-00314-0.

Rosenfeld, Avi, David G. Graham, Sarah Jevons, Jose Ariza, Daryl Hagan, Ash Wilson, Samuel J. Lovat, et al. 2020. ‘Development and Validation of a Risk Prediction Model to Diagnose Barrett’s Oesophagus (MARK-BE): A Case-Control Machine Learning Approach’. *The Lancet Digital Health* 2 (1): e37–48. https://doi.org/10.1016/S2589-7500(19)30216-X.

Royen, Florien S. van, Karel G. M. Moons, Geert-Jan Geersing, and Maarten van Smeden. 2022. ‘Developing, Validating, Updating and Judging the Impact of Prognostic Models for Respiratory Diseases’. *European Respiratory Journal* 60 (3). https://doi.org/10.1183/13993003.00250-2022.

RStudio Team. 2020. *RStudio: Integrated Development Environment for R*. Boston, MA: RStudio, PBC. http://www.rstudio.com/.

Rudin, Cynthia. 2019. ‘Stop Explaining Black Box Machine Learning Models for High Stakes Decisions and Use Interpretable Models Instead’. *Nature Machine Intelligence* 1 (5): 206–15. https://doi.org/10.1038/s42256-019-0048-x.

Rutledge, Jarod, Hamilton Oh, and Tony Wyss-Coray. 2022. ‘Measuring Biological Age Using Omics Data’. *Nature Reviews Genetics*, June, 1–13. https://doi.org/10.1038/s41576-022-00511-7.

Sachdeva, Gagandeep, Diana Han, Pearse A. Keane, Alastair K. Denniston, and Xiaoxuan Liu. 2023. ‘Demonstrating Clinical Impact for AI Interventions’. In *AI in Clinical Medicine*, 459–68. John Wiley & Sons, Ltd. https://doi.org/10.1002/9781119790686.ch42.

Said, Saredo, Raha Pazoki, Ville Karhunen, Urmo Võsa, Symen Ligthart, Barbara Bodinier, Fotios Koskeridis, et al. 2022. ‘Genetic Analysis of over Half a Million People Characterises C-Reactive Protein Loci’. *Nature Communications* 13 (1): 2198. https://doi.org/10.1038/s41467-022-29650-5.

Sánchez-Valle, Jon, Héctor Tejero, José María Fernández, David Juan, Beatriz Urda-García, Salvador Capella-Gutiérrez, Fátima Al-Shahrour, et al. 2020. 'Interpreting Molecular Similarity between Patients as a Determinant of Disease Comorbidity Relationships'. *Nature Communications* 11 (1): 2854. https://doi.org/10.1038/s41467-020-16540-x.

Satagopam, Venkata, Wei Gu, Serge Eifes, Piotr Gawron, Marek Ostaszewski, Stephan Gebel, Adriano Barbosa-Silva, Rudi Balling, and Reinhard Schneider. 2016. 'Integration and Visualization of Translational Medicine Data for Better Understanding of Human Diseases'. *Big Data* 4 (2): 97–108. https://doi.org/10.1089/big.2015.0057.

Sayed, Nazish, Yingxiang Huang, Khiem Nguyen, Zuzana Krejciova-Rajaniemi, Anissa P. Grawe, Tianxiang Gao, Robert Tibshirani, et al. 2021. 'An Inflammatory Aging Clock (iAge) Based on Deep Learning Tracks Multimorbidity, Immunosenescence, Frailty and Cardiovascular Aging'. *Nature Aging* 1 (7): 598–615. https://doi.org/10.1038/s43587-021-00082-y.

Schmidt, Philipp, and Felix Biessmann. 2019. 'Quantifying Interpretability and Trust in Machine Learning Systems'. *ArXiv:1901.08558 [Cs, Stat]*, January. http://arxiv.org/abs/1901.08558.

Schultze, Joachim L., and Philip Rosenstiel. 2018. 'Systems Medicine in Chronic Inflammatory Diseases'. *Immunity* 48 (4): 608–13. https://doi.org/10.1016/j.immuni.2018.03.022.

Schüssler-Fiorenza Rose, Sophia Miryam, Kévin Contrepois, Kegan J. Moneghetti, Wenyu Zhou, Tejaswini Mishra, Samson Mataraso, Orit Dagan-Rosenfeld, et al. 2019. 'A Longitudinal Big Data Approach for Precision Health'. *Nature Medicine* 25 (5): 792–804. https://doi.org/10.1038/s41591-019-0414-6.

Seshadri, Sudha, Annette L. Fitzpatrick, M. Arfan Ikram, Anita L. DeStefano, Vilmundur Gudnason, Merce Boada, Joshua C. Bis, et al. 2010. 'Genome-Wide Analysis of Genetic Loci Associated with Alzheimer Disease'. *JAMA* 303 (18): 1832–40. https://doi.org/10.1001/jama.2010.574.

Shah, Pratik, Francis Kendall, Sean Khozin, Ryan Goosen, Jianying Hu, Jason Laramie, Michael Ringel, and Nicholas Schork. 2019. 'Artificial Intelligence and Machine Learning in Clinical Development: A Translational Perspective'. *Npj Digital Medicine* 2 (1): 1–5. https://doi.org/10.1038/s41746-019-0148-3.

Shao, Wei, and Peter J. Espenshade. 2012. 'Expanding Roles for SREBP in Metabolism'. *Cell Metabolism* 16 (4): 414–19. https://doi.org/10.1016/j.cmet.2012.09.002.

Shen, Xiaotao, Ryan Kellogg, Daniel J. Panyard, Nasim Bararpour, Kevin Erazo Castillo, Brittany Lee-McMullen, Alireza Delfarah, et al. 2023. 'Multi-Omics Microsampling for the Profiling of Lifestyle-Associated Changes in Health'. *Nature Biomedical Engineering*, January, 1–19. https://doi.org/10.1038/s41551-022-00999-8.

Shepherd, Joanna M., Elaine Cole, and Karim Brohi. 2017. 'Contemporary Patterns of Multiple Organ Dysfunction in Trauma'. *Shock (Augusta, Ga.)* 47 (4): 429–35. https://doi.org/10.1097/SHK.0000000000000779.

Sierra, Felipe. 2016. 'The Emergence of Geroscience as an Interdisciplinary Approach to the Enhancement of Health Span and Life Span'. *Cold Spring Harbor Perspectives in Medicine* 6 (4). https://doi.org/10.1101/cshperspect.a025163.

Simoes, J., A. Bhangu, and CovidSurg Collaborative. 2020. 'Should We Be Re-Starting Elective Surgery?' *Anaesthesia* 75 (12): 1563–65. https://doi.org/10.1111/anae.15296.

Sinnott-Armstrong, Nasa, Sahin Naqvi, Manuel Rivas, and Jonathan K Pritchard. 2021. 'GWAS of Three Molecular Traits Highlights Core Genes and Pathways alongside a Highly Polygenic Background'. Edited by Jonathan Flint, Patricia J Wittkopp, Vincent J Lynch, Naomi Wray, and Aravinda Chakravarti. *ELife* 10 (February): e58615. https://doi.org/10.7554/eLife.58615.

Sinnott-Armstrong, Nasa, Yosuke Tanigawa, David Amar, Nina Mars, Christian Benner, Matthew Aguirre, Guhan Ram Venkataraman, et al. 2021. 'Genetics of 35 Blood and Urine Biomarkers in the UK Biobank'. *Nature Genetics* 53 (2): 185–94. https://doi.org/10.1038/s41588-020-00757-z.

Sollis, Elliot, Abayomi Mosaku, Ala Abid, Annalisa Buniello, Maria Cerezo, Laurent Gil, Tudor Groza, et al. 2023. 'The NHGRI-EBI GWAS Catalog: Knowledgebase and Deposition Resource'. *Nucleic Acids Research* 51 (D1): D977–85. https://doi.org/10.1093/nar/gkac1010.

Souza, Viviane Angelina de, Dílmerson Oliveira, Sérgio Ribeiro Barbosa, José Otávio do Amaral Corrêa, Fernando Antônio Basile Colugnati, Henrique Novais Mansur, Natália Maria da Silva Fernandes, and Marcus Gomes Bastos. 2017. 'Sarcopenia in Patients with Chronic Kidney Disease Not yet on Dialysis: Analysis of the Prevalence and Associated Factors'. *PLOS ONE* 12 (4): e0176230. https://doi.org/10.1371/journal.pone.0176230.

Spruijt, Nicole E., Tjaakje Visser, and Luke Ph Leenen. 2010. 'A Systematic Review of Randomized Controlled Trials Exploring the Effect of Immunomodulative Interventions on Infection, Organ Failure, and Mortality in Trauma Patients'. *Critical Care (London, England)* 14 (4): R150. https://doi.org/10.1186/cc9218.

STARSurg Collaborative. 2020. 'REspiratory COmplications after AbdomiNal Surgery (RECON): Study Protocol for a Multi-Centre, Observational, Prospective, International Audit of Postoperative Pulmonary Complications after Major Abdominal Surgery'. *British Journal of Anaesthesia* 124 (1): e13–16. https://doi.org/10.1016/j.bja.2019.10.005.

STARSurg Collaborative and COVIDSurg Collaborative. 2021. 'Death Following Pulmonary Complications of Surgery before and during the SARS-CoV-2 Pandemic'. *The British Journal of Surgery* 108 (12): 1448–64. https://doi.org/10.1093/bjs/znab336.

STARSurg Collaborative and TASMAN Collaborative. 2022. 'Evaluation of Prognostic Risk Models for Postoperative Pulmonary Complications in Adult Patients Undergoing Major Abdominal Surgery: A Systematic Review and International External Validation Cohort Study'. *The Lancet. Digital Health* 4 (7): e520–31. https://doi.org/10.1016/S2589-7500(22)00069-3.

Steyerberg, Ewout W., and Yvonne Vergouwe. 2014. 'Towards Better Clinical Prediction Models: Seven Steps for Development and an ABCD for Validation'. *European Heart Journal* 35 (29): 1925–31. https://doi.org/10.1093/eurheartj/ehu207.

Subramanian, Aravind, Pablo Tamayo, Vamsi K. Mootha, Sayan Mukherjee, Benjamin L. Ebert, Michael A. Gillette, Amanda Paulovich, et al. 2005. 'Gene Set Enrichment Analysis: A Knowledge-Based Approach for Interpreting Genome-Wide Expression Profiles'. *Proceedings of the National Academy of Sciences* 102 (43): 15545–50. https://doi.org/10.1073/pnas.0506580102.

Sundararajan, Mukund, and Amir Najmi. 2020. 'The Many Shapley Values for Model Explanation'. In *Proceedings of the 37th International Conference on Machine Learning*, 9269–78. ICML'20. JMLR.org.

Tacutu, Robi, Daniel Thornton, Emily Johnson, Arie Budovsky, Diogo Barardo, Thomas Craig, Eugene Diana, et al. 2018. 'Human Ageing Genomic Resources: New and Updated Databases'. *Nucleic Acids Research* 46 (D1): D1083–90. https://doi.org/10.1093/nar/gkx1042.

Tanaka, Toshiko, Angelique Biancotto, Ruin Moaddel, Ann Zenobia Moore, Marta Gonzalez-Freire, Miguel A. Aon, Julián Candia, et al. 2018. 'Plasma Proteomic Signature of Age in Healthy Humans'. *Aging Cell* 17 (5): e12799. https://doi.org/10.1111/acel.12799.

Tansey, Katherine E., Darren Cameron, and Matthew J. Hill. 2018. 'Genetic Risk for Alzheimer's Disease Is Concentrated in Specific Macrophage and Microglial Transcriptional Networks'. *Genome Medicine* 10 (1): 14. https://doi.org/10.1186/s13073-018-0523-8.

Tebani, Abdellah, Anders Gummesson, Wen Zhong, Ina Schuppe Koistinen, Tadepally Lakshmikanth, Lisa M. Olsson, Fredrik Boulund, et al. 2020. 'Integration of Molecular Profiles in a Longitudinal Wellness Profiling Cohort'. *Nature Communications* 11 (1): 4487. https://doi.org/10.1038/s41467-020-18148-7.

Teschendorff, Andrew E. 2019. 'Avoiding Common Pitfalls in Machine Learning Omic Data Science'. *Nature Materials* 18 (5): 422–27. https://doi.org/10.1038/s41563-018-0241-z.

Tibshirani, Robert. 1994. 'Regression Shrinkage and Selection Via the Lasso'. *Journal of the Royal Statistical Society, Series B* 58: 267–88.

Timmers, Paul RHJ, Ninon Mounier, Kristi Lall, Krista Fischer, Zheng Ning, Xiao Feng, Andrew D Bretherick, et al. 2019. 'Genomics of 1 Million Parent Lifespans Implicates Novel Pathways and Common Diseases and Distinguishes Survival Chances'. *ELife* 8 (January): e39856. https://doi.org/10.7554/eLife.39856.

Tomic, Adriana, Ivan Tomic, Levi Waldron, Ludwig Geistlinger, Max Kuhn, Rachel L. Spreng, Lindsay C. Dahora, et al. 2021. 'SIMON: Open-Source Knowledge Discovery Platform'. *Patterns (New York, N.Y.)* 2 (1): 100178. https://doi.org/10.1016/j.patter.2020.100178.

Tomic, Dunya, Jonathan E. Shaw, and Dianna J. Magliano. 2022. 'The Burden and Risks of Emerging Complications of Diabetes Mellitus'. *Nature Reviews Endocrinology* 18 (9): 525–39. https://doi.org/10.1038/s41574-022-00690-7.

'Turing-NLG: A 17-Billion-Parameter Language Model by Microsoft'. 2020. *Microsoft Research* (blog). 10 February 2020. https://www.microsoft.com/en-us/research/blog/turing-nlg-a-17-billion-parameter-language-model-by-microsoft/.

Uffelmann, Emil, Qin Qin Huang, Nchangwi Syntia Munung, Jantina de Vries, Yukinori Okada, Alicia R. Martin, Hilary C. Martin, Tuuli Lappalainen, and Danielle Posthuma. 2021. 'Genome-Wide Association Studies'. *Nature Reviews Methods Primers* 1 (1): 1–21. https://doi.org/10.1038/s43586-021-00056-9.

Van Calster, Ben, David J. McLernon, Maarten van Smeden, Laure Wynants, Ewout W. Steyerberg, Patrick Bossuyt, Gary S. Collins, et al. 2019. 'Calibration: The Achilles Heel of Predictive Analytics'. *BMC Medicine* 17 (1): 230. https://doi.org/10.1186/s12916-019-1466-7.

Van Calster, Ben, Ewout W. Steyerberg, Laure Wynants, and Maarten van Smeden. 2023. 'There Is No Such Thing as a Validated Prediction Model'. *BMC Medicine* 21 (1): 70. https://doi.org/10.1186/s12916-023-02779-w.

Vetrano, Davide L., Giulia Grande, Alessandra Marengoni, Amaia Calderón-Larrañaga, and Debora Rizzuto. 2021. 'Health Trajectories in Swedish Centenarians'. *The Journals of Gerontology Series A: Biological Sciences and Medical Sciences* 76 (1): 157. https://doi.org/10.1093/gerona/glaa152.

Vetrano, Davide L., Albert Roso-Llorach, Sergio Fernández, Marina Guisado-Clavero, Concepción Violán, Graziano Onder, Laura Fratiglioni, Amaia Calderón-Larrañaga, and Alessandra Marengoni. 2020. 'Twelve-Year Clinical Trajectories of Multimorbidity in a Population of Older Adults'. *Nature Communications* 11 (1): 3223. https://doi.org/10.1038/s41467-020-16780-x.

Violan, Concepció, Quintí Foguet-Boreu, Gemma Flores-Mateo, Chris Salisbury, Jeanet Blom, Michael Freitag, Liam Glynn, Christiane Muth, and Jose M. Valderas. 2014. 'Prevalence, Determinants and Patterns of Multimorbidity in Primary Care: A

Systematic Review of Observational Studies'. *PloS One* 9 (7): e102149. https://doi.org/10.1371/journal.pone.0102149.

Walraven, Carl van. 2014. 'The Hospital-Patient One-Year Mortality Risk Score Accurately Predicted Long-Term Death Risk in Hospitalized Patients'. *Journal of Clinical Epidemiology* 67 (9): 1025–34. https://doi.org/10.1016/j.jclinepi.2014.05.003.

Watanabe, Kyoko, Sven Stringer, Oleksandr Frei, Maša Umićević Mirkov, Christiaan de Leeuw, Tinca J. C. Polderman, Sophie van der Sluis, Ole A. Andreassen, Benjamin M. Neale, and Danielle Posthuma. 2019. 'A Global Overview of Pleiotropy and Genetic Architecture in Complex Traits'. *Nature Genetics* 51 (9): 1339–48. https://doi.org/10.1038/s41588-019-0481-0.

Watanabe, Kyoko, Erdogan Taskesen, Arjen van Bochoven, and Danielle Posthuma. 2017. 'Functional Mapping and Annotation of Genetic Associations with FUMA'. *Nature Communications* 8 (1): 1826. https://doi.org/10.1038/s41467-017-01261-5.

Wijst, Monique G. P. van der, Harm Brugge, Dylan H. de Vries, Patrick Deelen, Morris A. Swertz, and Lude Franke. 2018. 'Single-Cell RNA Sequencing Identifies Celltype-Specific Cis-EQTLs and Co-Expression QTLs'. *Nature Genetics* 50 (4): 493–97. https://doi.org/10.1038/s41588-018-0089-9.

Wolff, Robert F., Karel G.M. Moons, Richard D. Riley, Penny F. Whiting, Marie Westwood, Gary S. Collins, Johannes B. Reitsma, Jos Kleijnen, and Sue Mallett. 2019. 'PROBAST: A Tool to Assess the Risk of Bias and Applicability of Prediction Model Studies'. *Annals of Internal Medicine* 170 (1): 51–58. https://doi.org/10.7326/M18-1376.

Wolpert, D.H., and W.G. Macready. 1997. 'No Free Lunch Theorems for Optimization'. *IEEE Transactions on Evolutionary Computation* 1 (1): 67–82. https://doi.org/10.1109/4235.585893.

Wynants, Laure, Ben Van Calster, Gary S. Collins, Richard D. Riley, Georg Heinze, Ewoud Schuit, Marc M. J. Bonten, et al. 2020. 'Prediction Models for Diagnosis and Prognosis of Covid-19: Systematic Review and Critical Appraisal'. *BMJ* 369 (April). https://doi.org/10.1136/bmj.m1328.

Xia, Jianguo, and David S. Wishart. 2016. 'Using MetaboAnalyst 3.0 for Comprehensive Metabolomics Data Analysis'. *Current Protocols in Bioinformatics* 55 (1): 14.10.1-14.10.91. https://doi.org/10.1002/cpbi.11.

Xiao, Wenzhong, Michael N. Mindrinos, Junhee Seok, Joseph Cuschieri, Alex G. Cuenca, Hong Gao, Douglas L. Hayden, et al. 2011. 'A Genomic Storm in Critically Injured Humans'. *The Journal of Experimental Medicine* 208 (13): 2581–90. https://doi.org/10.1084/jem.20111354.

Yao, Douglas W., Luke J. O'Connor, Alkes L. Price, and Alexander Gusev. 2020. 'Quantifying Genetic Effects on Disease Mediated by Assayed Gene Expression Levels'. *Nature Genetics* 52 (6): 626–33. https://doi.org/10.1038/s41588-020-0625-2.

Yengo, Loïc, Sailaja Vedantam, Eirini Marouli, Julia Sidorenko, Eric Bartell, Saori Sakaue, Marielisa Graff, et al. 2022. 'A Saturated Map of Common Genetic Variants Associated with Human Height'. *Nature* 610 (7933): 704–12. https://doi.org/10.1038/s41586-022-05275-y.

Yim, Jason, Reena Chopra, Terry Spitz, Jim Winkens, Annette Obika, Christopher Kelly, Harry Askham, et al. 2020. 'Predicting Conversion to Wet Age-Related Macular Degeneration Using Deep Learning'. *Nature Medicine* 26 (6): 892–99. https://doi.org/10.1038/s41591-020-0867-7.

Zenin, Aleksandr, Yakov Tsepilov, Sodbo Sharapov, Evgeny Getmantsev, L. I. Menshikov, Peter O. Fedichev, and Yurii Aulchenko. 2019. 'Identification of 12 Genetic Loci Associated with Human Healthspan'. *Communications Biology* 2 (1): 1–11. https://doi.org/10.1038/s42003-019-0290-0.

Zhang, Angela, Lei Xing, James Zou, and Joseph C. Wu. 2022. 'Shifting Machine Learning for Healthcare from Development to Deployment and from Models to Data'. *Nature Biomedical Engineering* 6 (12): 1330–45. https://doi.org/10.1038/s41551-022-00898-y.

Zhang, Chiyuan, Samy Bengio, Moritz Hardt, Benjamin Recht, and Oriol Vinyals. 2016. 'Understanding Deep Learning Requires Rethinking Generalization'. ArXiv.Org. 10 November 2016. https://arxiv.org/abs/1611.03530v2.

Zhang, Xinyuan, Anastasia M. Lucas, Yogasudha Veturi, Theodore G. Drivas, William P. Bone, Anurag Verma, Wendy K. Chung, et al. 2022. 'Large-Scale Genomic Analyses Reveal Insights into Pleiotropy across Circulatory System Diseases and Nervous System Disorders'. *Nature Communications* 13 (1): 3428. https://doi.org/10.1038/s41467-022-30678-w.

Zhou, Wenyu, M. Reza Sailani, Kévin Contrepois, Yanjiao Zhou, Sara Ahadi, Shana R. Leopold, Martin J. Zhang, et al. 2019. 'Longitudinal Multi-Omics of Host–Microbe Dynamics in Prediabetes'. *Nature* 569 (7758): 663–71. https://doi.org/10.1038/s41586-019-1236-x.

Zhu, Xudong, Zhiyang Chen, Weiyan Shen, Gang Huang, John M. Sedivy, Hu Wang, and Zhenyu Ju. 2021. 'Inflammation, Epigenetics, and Metabolism Converge to Cell Senescence and Ageing: The Regulation and Intervention'. *Signal Transduction and Targeted Therapy* 6 (June): 245. https://doi.org/10.1038/s41392-021-00646-9.

Zou, Hui, and Trevor Hastie. 2005. 'Regularization and Variable Selection via the Elastic Net'. *Journal of the Royal Statistical Society. Series B: Statistical Methodology* 67 (2): 301–20. https://doi.org/10.1111/j.1467-9868.2005.00503.x.