PREDICTIVE MODELING OF DAMAGE AND REPAIR FOR DISEASE AND ACTIVITY OF DAILY LIVING STATUS IN ELSA DATASET USING MACHINE LEARNING MODELS

by

Emre Dil

Submitted in partial fulfillment of the requirements
for the degree of Master of Science

at

Dalhousie University
Halifax, Nova Scotia
November 2023

Table of Contents

List of Tables . . . . . **iv**

List of Figures . . . . . **v**

Abstract . . . . . **vii**

Acknowledgements . . . . . **viii**

Chapter 1 Introduction **1**

1.1 Background 1

1.2 Our Study: Motivation and Objectives 3

Chapter 2 Literature Review **5**

2.1 Overview of Longitudinal Datasets 5
2.1.1 ELSA 6

2.2 Previous Studies on Neural Networks, Random Forest, and Logistic Regression in Health State Prediction 7
2.2.1 Neural Networks 7
2.2.2 Random Forests 8
2.2.3 Logistic Regression 9

2.3 Analyzing Damage and Repair Probabilities and Its Implications . . . 10

Chapter 3 Research Methodology **12**

3.1 Details of ELSA and Variable Selection 12

3.2 Neural Networks 14

3.3 Random Forest 15

3.4 Logistic Regression 17

3.5 Imputation, Model and Feature Selection, and Analysis 18

Chapter 4 Results and Analysis **20**

4.1 Motivation of Paper 20

4.2 Credits to Authors 21

4.3 Paper . . . 21
4.3.1 Abstract . . . 21
4.3.2 Introduction . . . 21
4.3.3 Methods and Model Selection . . . 23
4.3.4 Results . . . 29
4.3.5 Discussion . . . 32
4.3.6 Acknowledgements . . . 34
4.3.7 Supplementary Material . . . 34

4.4 Perspective . . . 58

Chapter 5 Additional Results and Analysis . . . 59

5.1 Repair Transition Probabilities . . . 59

5.2 Male-Female Hazard Ratios . . . 64

5.3 Correlations of Transition Probabilities . . . 67

5.4 Calibration of Correlations for Health States and Damage Transition Probabilities . . . 70

5.5 Relations between Prevalences, Exponents of Transition Probabilities and HRs . . . 74

Chapter 6 Conclusion and Discussion . . . 78

Bibliography . . . 81

Appendix A Supplementary Material . . . 91

List of Tables

3.1 Summary of ELSA Data 13

4.1 Summary of ELSA Data 25

4.2 Performance of imputation methods 26

4.3 Model performance for diseases and ADLs 30

4.4 Percentage prevalence of 19 disease outputs 36

4.5 Percentage prevalence of 25 ADL outputs 37

4.6 Used Core and Nurse Variables 39

4.7 Performance of imputation methods for LR predictions 40

4.8 Selected features using the filter feature selection method . . . 41

4.9 Model selection for DNNs 42

4.10 Comparison of random forest and the best DNN 42

4.11 Comparison of the prediction performance of logistic regression and the best DNN 42

4.12 Comparison of the prediction performance of logistic regression and the best DNN for transition probabilities 43

4.13 Calibration scores 43

4.14 Individual disease calibration scores 43

4.15 Individual ADL calibration scores 44

4.16 Observed and predicted HRs for bimodal distributions 44

5.1 Individual disease calibration scores of repair transition probabilities 60

5.2 Individual ADL calibration scores of repair transition probabilities 61

S1 Percentage prevalence of 19-Disease outputs 92

S2 Percentage prevalence of 25-ADL outputs 93

List of Figures

4.1 Study flowchart . . . 24

4.2 Calibration curves for damage transition probabilities . . . 30

4.3 Hazard ratios . . . 31

4.4 Prediction performance in terms of AUC vs #n features . . . 45

4.5 Predicted versus observed average damage and repair probabilities for Diseases . . . 46

4.6 Predicted versus observed average damage and repair probabilities for ADLs . . . 47

4.7 Individual disease calibration curves for damage transition probabilities . . . 48

4.8 Individual ADL calibration curves for damage transition probabilities . . . 49

4.9 Prediction performance vs age . . . 50

4.10 Hazard ratio versus prevalence of diseases and ADLs . . . 50

4.11 HR versus all diseases and ADLs . . . 51

4.12 Histograms of predicted disease damage transition probabilities 52

4.13 Histograms of predicted disease damage transition probabilities in Waves 2-3 in log-log scale . . . 53

4.14 Histograms of predicted ADL damage transition probabilities . 54

4.15 Histograms of predicted ADL damage transition probabilities in log-log scale . . . 55

4.16 Correlations between observed health states . . . 56

4.17 Correlations between predicted health states . . . 57

5.1 Individual disease calibration curves for repair transition probabilities . . . 62

5.2 Individual ADL calibration curves for repair transition probabilities . . . 63

5.3 Hazard ratios with males and females 65
5.4 HR versus all diseases and ADLs for all, male and females . . 66
5.5 Correlations between observed damage transition probabilities 68
5.6 Correlations between predicted damage transition probabilities 69
5.7 Calibration of observed, predicted and simulated (once) health states and damage transition probability correlations 71
5.8 Calibration of observed, predicted and simulated (100 times) health states and damage transition probability correlations . 72
5.9 Calibration of observed, predicted and simulated (100 times and averaged) health states and damage transition probability correlations 73
5.10 Hazard ratios for diseases in wave transition from 2 to 3 . . . 75
5.11 Hazard ratios for ADLs in wave transition from 2 to 3 76
5.12 Relations between hazard ratios, exponents in damage transition probability distributions, and prevalences 77

Abstract

A good predictive model is useful in health sciences for predicting onset of disease, as well as damage or repair of health deficits. One can predict one or more of these quantities depending on the nature of the collected data. In this thesis, we predict the future binary health states between successive waves of the English Longitudinal Study of Aging (ELSA) dataset. The predicted health states are 19 diseases and 25 activities of daily living states (ADLs) of individuals in the ELSA study. While we can directly predict those states with a high prediction quality, we cannot directly predict damage and repair probabilities or individual binary damage transition probabilities with the similar high prediction quality. However, we could predictively model damage and repair probabilities using the predicted health states. We applied model selection between deep neural networks (DNN), random forests, and logistic regression, then found that a simple one-hidden layer 128-node DNN was best able to predict future health states (AUC ≥ 0.91) and average damage and repair probabilities ($R^2 \geq 0.92$). We applied feature selection for 134 full explanatory variables and found that 33 variables are sufficient to predict all disease and ADL states well. The prediction quality of individual damage transition probabilities are analyzed by the deciles of the probabilities and found to be well calibrated. We also studied the correlations between predicted health states which were stronger than the observed correlations. The hazard ratios (HRs) between high-risk deciles and the average were between 3 and 10 where high prevalence damage transitions typically had smaller HRs. We did not find a significant relation between model predictions versus individual ages.

Acknowledgements

I sincerely thank to my supervisor Professor Andrew Rutenberg who gave me the opportunity to be a master student and to do research in his group on aging studies based on machine learning methods. I am also thankful to Glen Pridham being Professor Rutenberg's PhD student and my office mate for his help and suggestions on scientific and technical points. Moreover, I would like to thank to committee members Professor Joanna Mills Flemming, and Professor Laurent Kreplak for their valuable guidance and help. I finally thank to Dalhousie University, Physics Department for giving me this opportunity to attend its community and Graduate Coordinator Assistant Tanya Timmins for her help in any period of my journey at Dal.

Chapter 1

Introduction

1.1 Background

Aging research is a wide-ranging scientific endeavor aimed at comprehending the processes of aging and enhancing the duration of one's healthy and active lifespan. Aging is a complex process with many factors in its causes and effects. Various genetic, molecular, cellular and environmental elements can contribute to the decline of different bodily functions as one ages [1].

Aging can be described as the decline of an organism's healthy functioning over time [2]. Age plays a crucial role in influencing the susceptibility and development, and outcome of various health conditions. The prevalence of numerous chronic diseases and impairments rises as individuals grow older [3, 4]. It is an open question whether age itself determines the dysfunction, or whether the things associated with age determines it (e.g. dysfunction and disease).

The significance of aging is becoming more pronounced as the global population continues to age. By 2050, approximately 25% of the population in Western countries is projected to be over 65 years old, impacting healthcare, social security, and the economy [5, 6, 7, 8]. This is one reason why aging research is getting more attention in developed countries recently, leading to the emergence of high dimensional and larger scale datasets.

Age plays an important role in how diseases affect people and develop over time. It is crucial in health science research to understand the reasons behind damage propagation during aging. To understand the aging process, the decline of health can be defined or summarized in a quantitative manner, and then can be analyzed by various methodologies, such as predictive ML models, more interpretable regression models or exploratory data analysis.

Predictive ML models help the biological or health data of individuals to be analyzed to understand the process of how individuals change as they age. There are

various aspects of the datasets, such as measured biomarkers, asked survey questions for previous health conditions, or lifestyle questions for each individual. Those variables in the dataset may be taken once or several times over a period of time. A dataset where each individual is assessed at a single time point is called cross-sectional, while datasets including multiple time assessments of individuals are called longitudinal datasets. From the cross-sectional data, we can only get the current correlations between measured variables and change in health of individuals, not the dynamics. Acquiring extensive longitudinal data, where individuals are assessed at different or periodic time points, is much more useful to understand the aging dynamics.

Numerous countries carry out longitudinal aging research by gathering information on senior individuals within the population. Some examples are SHARE (Survey of Health Ageing and Retirement in Europe), TILDA (Irish Longitudinal Study on Ageing), CLHLS (Chinese Longitudinal Healthy Longevity Survey), WLS (The Wisconsin Longitudinal Study), KloSA (Korean Longitudinal Study of Aging), CLSA (Canadian Longitudinal Study on Aging), NSJE (National Survey of the Japanese Elderly), LASI (Longitudinal Aging Study in India), HRS (Health and Retirement Study in USA) and ELSA (English Longitudinal Study of Ageing). These datasets encompass various variables, such as physical and mental health conditions, daily activity capabilities, as well as demographic and socioeconomic data, collected across multiple visits or waves over time. Population studies typically involve tens of thousands of participants (instances) with hundreds of variables (dimensions) [9]. In this thesis study, we will use ELSA [10] dataset since our group studied some different but related aspects of ELSA data [11, 12].

Machine learning (ML) methods have been applied to health and medicine data since the 1980s [13, 14] due to the complex and high-dimensional nature of the data. These applications have proven useful in various clinical implementations, such as risk prediction, extracting key information from patient records, generating clinical notes, real-time detection of patients at risk for deterioration, and providing personalized predictions [15]. ML models excel at handling hundreds or thousands of variables simultaneously, including multi-dimensional target variables [16, 17, 18]. While predictive ML models applied to cross-sectional data provide insights into current health status, their use on longitudinal data offers valuable information about future health

outcomes [19, 20].

Aging research is a multifaceted scientific pursuit aimed at understanding the complex processes of aging and extending the duration of a healthy and active lifespan. It encompasses various factors that contribute to age-related decline in bodily functions. As individuals age, they become more susceptible to chronic diseases and health conditions, raising questions about whether age itself or associated factors drive dysfunction and disease. The global aging population highlights the increasing significance of aging research. Our approach to address these challenges focuses on predictive machine learning (ML) models employed to analyze biological and health data, particularly in longitudinal datasets, to better understand how individuals change over time. ML methods have proven valuable in handling the high-dimensional and complex nature of health data, offering insights into the age related health states, as well as damage and repair occurrences that will be dealt with in the next section.

1.2 Our Study: Motivation and Objectives

There are three studies motivating us to investigate the dynamics of aging concerning the occurrence of disease and dysfunction on the ELSA dataset individuals. From our group, Farrell et al. [11] developed a computational dynamic joint interpretable network (DJIN) model predicting aging trajectories related to continuous health variables in the ELSA dataset. Farrell et al. [12] also explored the damage and repair rates in discrete ADLs variables of ELSA. In the third study, Buergel et al. [21] investigated risk stratification for high-risk individuals using the UK Biobank Dataset by specifically examining the nuclear magnetic resonance (NMR) spectroscopy-derived metabolomic profiles for 24 common conditions and obtained hazard ratios (HR) from calibrated event rate predictions generated by a neural network. We will explore the binary disease and ADL states in ELSA and develop a ML model to predict those binary health states, and their damage and repair probabilities.

We will focus on the ELSA dataset, specifically the core and nurse data variables. The study considers different age groups to see how age is linked to disease and ADL damage probabilities, revealing the diverse risks faced by people as they get older. We aim to develop an ML model for the prediction of future health states of individuals using explanatory data from previous waves (visits). From these health

states, we calculate the occurrence of average damage and repair probabilities in future waves for both predicted and observed health states. Using the best predictive model, we explore the age dependency of the model by comparing the evaluation metrics across different age groups. Additionally, we analyze the model's performance separately for diseases and ADLs to determine if there are significant differences in prediction accuracy. We also investigate the calibration of our model by examining the rank-ordered individual damage and repair transition probabilities in deciles for both predicted and observed values. By calculating hazard ratios for 19 diseases and 25 ADLs, we assess the risk stratification and explore any potential relationship between the prevalence of health outputs and hazard ratios. Finally, we study the correlations between the damage transition probabilities of all health outputs.

We want to investigate two important questions in this study. The first one is to see if a simple DNN model can well predict the discrete health states and damage and/or repair transitions using in the ELSA dataset, also if this model can give better prediction results than simpler logistic-regression models (see e.g. [22, 23]) or not [24, 25]. The second one is to find the most important variables from the high-dimensional input data to predict those binary states and transitions, such as discrete or continuous input variables.

Chapter 2

Literature Review

In this chapter, we will overview the longitudinal datasets. We also introduce the examples of used machine learning models for those health state predictions. Lastly, we talk about the product of health state predictions as damage and repair probabilities.

2.1 Overview of Longitudinal Datasets

In the literature, one can find the detailed information about various longitudinal aging datasets in aging research. A review paper for 51 longitudinal studies of aging [26] provides an overview of the key discoveries and points out the areas that require further investigation within six overarching research themes: cognitive function, socioeconomic status, health and physical capabilities, factors predicting illness and mortality, healthcare expenditures, and genetics. Another review paper [27] examines the objectives of each study, methodological comparisons, central policy themes, advantages, interactions with other datasets, areas that need improvement, challenges, and the principal findings achieved up to this point. Also the harmonized versions of many longitudinal aging studies in terms of the harmonized names, types and wave number of the variables can be found with the access links on the Gateway to Global Aging Data platform for population survey data on aging around the world [28].

Because our group previously studied ELSA data in terms of predictive and exploratory analysis [11, 12] and because of our easy access to ELSA, we use the original non-harmonized ELSA dataset in this study. The majority of publicly accessible ELSA data can be obtained through the UK Data Service (UKDS), which offers three tiers of access: 1) End User License (EUL), 2) Special License (SL), and 3) Secure Access (SA). EUL data, which we used, is anonymized and necessitates only registration with UKDS. SL data contain variables that pose a slightly higher risk of identification and thus require additional approval. SA data are regarded as sensitive

and can only be accessed through a Secure Lab or encrypted remote access.

Health state prediction in longitudinal datasets is an important approach to understand the future health outcomes of individuals based on their previous health records. Longitudinal datasets are particularly advantageous as they include data from the same individuals over multiple time points, allowing researchers to investigate changes in health over time and make predictions about future health states. Researchers typically employ machine learning techniques to model the relationship between previous wave data (explanatory variables) and subsequent health states. By utilizing the rich information available in the dataset, these predictive models can offer valuable insights into the future health outcomes of individuals [29, 30, 31]

2.1.1 ELSA

The longitudinal dataset used in this study is the English Longitudinal Study of Ageing (ELSA) dataset which is a large-scale, nationally representative survey of adults aged 50 years and older in England. The dataset collects comprehensive information on various aspects of participants' lives, including health, socio-economic status, and lifestyle factors. ELSA's design incorporates multiple waves of data collection, with follow-up visits occurring every two years, enabling researchers to track participants' health trajectories over an extended period [32, 33].

Researchers have used health state prediction on ELSA data for various purposes, including risk prediction for specific diseases (e.g., long-term cholesterol risk, hypertension risk, dementia) [34, 35, 36] and the identification of factors that contribute to functional decline or disabilities in activities of daily living (ADLs). Additionally, predictive models have been developed to investigate the age dependency of health state predictions, exploring how prediction performance varies across different age groups [11, 12].

In the next section, we will give the machine learning models that will be employed on ELSA dataset in our study. By employing machine learning techniques and sophisticated analyses, valuable insights into disease occurrence, functional decline, and age-related health changes can be gained.

2.2 Previous Studies on Neural Networks, Random Forest, and Logistic Regression in Health State Prediction

In recent years, health state prediction has garnered significant attention as a critical task in medical research and clinical decision-making. Various machine learning algorithms have been explored to predict health states accurately. Among them, three popular approaches are neural networks, random forest, and logistic regression. This section provides an overview of previous studies that have utilized these algorithms for health state prediction and highlights their respective strengths and limitations.

2.2.1 Neural Networks

Neural networks are a class of machine learning models originally inspired by the structure and functioning of the human brain. They consist of interconnected artificial neurons, organized into layers, and are used for various tasks, including predictions or classifications by adjusting the strengths of connections (synaptic weights) between neurons.

Popular architectures are deep neural networks (DNNs), having more than one hidden layers, and convolutional neural networks (CNNs) performing convolution operations on input data, applying filters to detect local patterns or features and mostly used in image processing. A Recurrent Neural Network (RNN) is another type of neural network architecture specifically designed for processing sequences of data having a memory of past inputs, and maintaining a hidden state so that RNNs are suitable for natural language processing, speech recognition, and complex time series predictions. Neural networks have been widely explored in research and practical applications, as pattern recognition, image and speech processing, and data analysis. [37].

Neural networks, especially deep learning models, have demonstrated remarkable success in diverse healthcare applications. Due to its relatively simplistic architecture, there has been many research in the realm of DNNs [38, 39]. Khanam and Foo [40] executed a neural network (NN) model for forecasting diabetes, utilizing 1, 2, and 3 hidden layers with varying epochs set at 200, 400, and 800, respectively. Notably, the second hidden layer, operating at 400 epochs, achieved an accuracy of

88.6%, surpassing other machine learning models like Decision Trees, Random Forest and Logistic Regression. In 2021, Soundarya et al. [41] utilized DNN for Alzheimer's Disease (AD) detection, showcasing its superior accuracy when compared to various machine learning models, given adequate data. Additionally, Pasha et al. [42] utilized DNN to enhance the precision of cardiovascular disease prediction. Traditional machine learning models struggle with large datasets, whereas DNN exhibits an advantage in handling such extensive data. These instances collectively signal DNN as a forthcoming trend, suggesting that deep learning, particularly through DNN, will likely emerge as the primary algorithm for disease prediction.

Researchers have also leveraged convolutional neural networks (CNNs) for medical image analysis, recurrent neural networks (RNNs) for time-series data, and transformer-based (RNN like) models for natural language processing tasks in the medical domain. For instance, Rajpurkar et al. [43] applied CNNs to chest X-rays for pneumonia detection, achieving accuracy comparable to radiologists. Meanwhile, Choi et al. [44] used RNNs to predict patient health deterioration using electronic health records, demonstrating the potential of recurrent architectures in clinical settings. While neural networks exhibit excellent predictive capabilities, their black-box nature and high computational demands raise interpretability and resource concerns [45, 46]. The details of DNN working are given in Section 3.2

2.2.2 Random Forests

Random Forest (RF) is an ensemble learning method widely used for classification and regression tasks. It leverages decision trees as base learners and employs bagging to create multiple bootstrapped subsets of the dataset for training individual trees. Notably, it introduces random feature selection, considering only a subset of features at each node in each tree, which helps reduce overfitting and decorrelates the trees. When making predictions, Random Forest combines the results from individual trees by majority voting in classification tasks and averaging in regression. This method is highly regarded for its robustness to overfitting, suitability for large and high-dimensional datasets, and its ability to provide feature importance scores in the final ensemble [47, 48, 49].

Random forest has been widely adopted in healthcare for its ability to handle high-dimensional data and nonlinear relationships effectively. Researchers have employed random forest models for predicting various medical conditions, including diabetes, cancer, and heart disease. For instance, Li et al. [50] utilized random forest to predict diabetic retinopathy, achieving high sensitivity and specificity rates. Similarly, Wang et al. [51] applied the random forest algorithm to detect lung cancer, showcasing its potential as a reliable diagnostic tool. However, random forest models may struggle with overfitting on noisy data and can be computationally intensive for large datasets [52]. Therefore, we expect DNN models to present better prediction quality than the RFs since we will be considering the large ELSA dataset as mentioned in a previous section.

2.2.3 Logistic Regression

Logistic regression (LR) is a classic linear classifier which is a type of machine learning model that aims to classify data into different categories or classes based on a decision boundary. LR is a statistical method used for binary classification, which means it's employed to predict the probability of an outcome falling into one of two categories. It's an extension of linear regression that models the relationship between a dependent binary variable and one or more independent variables, applying a sigmoidal logistic function to predict the probability of the binary outcome. Logistic regression is commonly used in fields such as medicine, social sciences, and machine learning for tasks like spam detection and disease diagnosis.

Logistic regression has been extensively used in medical research due to its simplicity, interpretability, and efficiency. This method is particularly suitable for binary classification tasks and has found applications in predicting outcomes such as mortality, readmission, and disease presence [53]. For example, Ross et al. [54] employed logistic regression to predict readmission within 30 days for heart failure patients, facilitating targeted interventions. Likewise, Barghi and Azadeh-Fard [55] utilized logistic regression with some other ML models for early detection of sepsis, showcasing its effectiveness in time-critical scenarios. Despite its simplicity, logistic regression may struggle to capture complex nonlinear relationships in data [56] and we expect LR to give weaker prediction performance than DNN for our complex ELSA dataset

with 19 binary diseases and 25 binary ADL target variables.

The prediction of health states has witnessed significant advancements through the application of neural networks, random forest, and logistic regression. Neural networks offer powerful and flexible models capable of handling diverse data modalities but come with challenges related to interpretability and computational resources. Random forest excels in managing complex data structures but may face issues with overfitting and computational efficiency. On the other hand, logistic regression remains a reliable and interpretable choice for binary classification tasks, although it may not capture intricate nonlinear patterns. We will investigate the best prediction model for binary health states prediction in the ELSA dataset, we expect DNN to predict better than RF and LR.

2.3 Analyzing Damage and Repair Probabilities and Its Implications

In medicine, understanding the probabilities of damage occurrence and the effectiveness of repair mechanisms is crucial for various applications, ranging from disease prognosis and treatment to the development of personalized medicine strategies [57]. This section explores the significance of analyzing damage and repair probabilities in medical contexts and its implications on patient care and research.

Molecular and cellular damage and dysfunction are the first scales of the 'hallmarks' of aging [2]. DNA damage is a fundamental driver of various diseases, including cancer and genetic disorders. Analyzing the probabilities of DNA damage occurrence due to external factors (e.g., radiation, chemicals) or endogenous processes (e.g., oxidative stress) is essential to comprehend disease risk and development. Several studies have investigated DNA repair mechanisms and their efficiency in maintaining genomic integrity [58].

Cellular damage caused by inflammation, injury, or aging can lead to tissue dysfunction and organ failure. Analyzing the probabilities of cellular damage and the regenerative capacity of tissues is essential for understanding the progression of diseases and designing therapeutic interventions [59, 60]. By understanding the repair capacity of cancer cells, researchers can anticipate the efficacy of chemotherapy and radiation therapy [61, 62]. Similarly, studies on repair pathways in specific diseases, such as BRCA1/2 mutations in breast cancer [63], provide valuable information for

targeted therapies and personalized treatment decisions. Predicting diagnosis and treatment responses based on damage and repair capacity helps optimize strategies and improve patient outcomes.

Damage and repair at the level of whole organisms may exhibit unique behaviors [64, 57, 1]. The damage and repair dynamics can be related to the summary health measures, Frailty Index (FI) [65] and the Frailty Phenotype [66] due to the increasing health deficits with worsening health [67, 68]. Moreover, from studies on resilience [69, 70, 71] and robustness [72, 73], sustaining health of the organism during aging can be related to organismal damage and repair [12].

Farrel et. al. [12] proposed a novel analytical method to examine both damage and repair for binary health indicators since the majority of health state data is binary. This method leverages longitudinal data to generate summarized measures of organismal damage and repair processes over time, corresponding to discrete shifts in those binary health state variables. In our study, we adopt the methodology in [12] for damage and repair discussed below.

In this thesis, we obtain damage and repair probabilities by predicting binary health state variables of 19 diseases and 25 ADLs. Analyzing damage and repair probabilities can aid in exploring the risk stratification and finding the hazard ratios over different health states.

Chapter 3

Research Methodology

3.1 Details of ELSA and Variable Selection

The dataset of the English Longitudinal Study of Ageing (ELSA) [10] has been ongoing since 1998 and has gone through 9 waves of data collection so far, with each wave approximately 2 years apart. Certain waves of the study involve nurse visits, which occurred in waves 2, 4, 6, 8, and 9. During these visits, nurses collected various health-related data, including blood pressure, glucose levels, cholesterol levels, and physical measurements such as height, weight, and hand grip strength, by visiting the participants' homes. Due to attrition from the study, caused by factors such as death or personal preferences of individuals, there were both incoming and outgoing participants in each wave. The total number of unique individuals across all waves in core and nurse data is 22,964, with an average of approximately 10,000 individuals per wave.

To account for possible correlations between health states and the percentage of missing data in the variables, our study considered a total of 134 different variables from thousands of core and nurse data variables. We chose 134 variables according to the missingness percentage and health-related variables. We proceeded by removing variables with a missingness percentage exceeding 30% [74]. Those 134 variables consist of 78 binary and 14 continuous variables from the core data, and 42 continuous variables from the nurse data, for each wave. The variables encompass a wide range of aspects, including the presence of certain diseases, difficulties in daily activities, medication and treatment history, self-reported health status, activity levels, smoking and alcohol usage history, as well as various biomarker measurements, and physical measurements like height, weight, and blood pressure.

In this study, we focus on the even-numbered waves (2, 4, 6, and 8) of the ELSA dataset in order to combine core data with the nurse data. From these waves, we select a total of 134 explanatory variables, consisting of both binary and continuous

	W0	W1	W2 → W3		W4 → W5		W6 → W7		W8 → W9	
Individuals #	8267	12099	9432	9771	11050	10274	10601	9666	8445	8736
Same Individuals #	-	5053	9324	7680	7908	9460	8999	8866	8088	7146
New Individuals #	-	7046	108	2091	3142	814	1602	800	357	1590
Women %	57	56	56	56	55	56	55	56	56	56
Men %	43	44	44	44	45	44	45	44	44	44
Ever Smoked %	71	70	70	70	66	58	57	54	52	51
Core Missing %	29	30	27	28	27	26	25	25	22	25
Nurse Missing %	-	-	8	-	9	-	9	-	10	10

Table 3.1: Summary of ELSA Data: W0, W1, ..., W9 stand for the Wave 0, Wave 1, ..., Wave 9 in ELSA dataset. Individuals # is the number of patients in each wave. Same individuals # is the number of same patients in a previous wave. New Individuals # is the number of newly added individuals to the current wave when compared to previous wave. Core Missing % and Nurse Missing % are the missing data percentage of core data and nurse (lab) data in corresponding wave. We consider the transitions from waves 2 to 3, 4 to 5, 6 to 7 and 8 to 9. We used W0 and W1 for imputation and model evaluation purposes.

variables. Our objective is to use these explanatory variables to predict the binary health state variables in the subsequent odd-numbered waves (3, 5, 7, and 9). The binary health state variables are categorized into two sets: 19 diseases and 25 Activities of Daily Living (ADLs), as outlined in Supplementary Tables S1 and S2.

Table 3.1 provides a concise overview of the dataset, including the number of individuals, the percentage of women and men, the percentage of smokers, and the percentage of missing data. From the second and third rows of Table 3.1, one can see the complexity of dataset due to the variability of the population in a single wave. There are generally a significant amount of individuals are going out and coming into the study in each wave. We also show, the prevalence of each health state considered as target variables, pertaining to the 19 diseases and 25 ADLs, in each wave is presented in Supplementary Tables S1 and S2. These prevalence values offer valuable insights into the distribution of health states across different waves of the study.

Overall, this dataset and the selected explanatory variables should hold significant potential for predicting the binary health state variables in the future waves, contributing to a better understanding of health dynamics and paving the way for improved medical research and patient care.

Below, we conduct a systematic selection process for imputation algorithms to identify the most suitable method and missingness mechanism for handling the missing values. Once the best method and mechanism are determined, we proceed to impute all missing variables in all the waves under consideration, utilizing the selected approach. The further details of the imputation algorithm selection procedure

will be provided in Section 3.5.

3.2 Neural Networks

Neural networks (NNs) are computational models composed of interconnected nodes, inspired by the structure and functioning of biological neurons in the human brain. They are widely used in machine learning and artificial intelligence tasks due to their ability to learn complex patterns from data [75]. NNs can handle large datasets and high-dimensional feature spaces effectively. A typical NN consists of layers of interconnected nodes, also known as neurons or units. These layers include [76, 77, 78]:

1) Input Layer: This layer receives the initial data or features for the network.

2) Hidden Layers: These intermediate layers process the input data through weighted connections and activation functions. Deep neural networks (DNNs) have multiple hidden layers, allowing them to capture intricate relationships in the data.

3) Output Layer: The final layer produces the network's predictions or outcomes.

As every machine learning models, NNs are based on training and testing steps, where training step adjusts the model parameters by passing input and output variables of a certain amount of total data (say 80%) while test steps predicts the outputs from the inputs of the remaining amount of total data (say 20%). One can see overfit of the model if test error is relatively higher than the training error, meaning model memorized the structures but not learned it. NNs work by passing information through the network's layers and during training, NNs adjust weights to minimize the difference between predicted and actual outputs. The training process involves two main steps: feedforward and backpropagation [76].

Feedforward: The input data is propagated through the network's layers, with each neuron's output contributing to the next layer's inputs. The weighted sum of inputs is then passed through an activation function, which introduces non-linearity and allows the network to model complex relationships.

Backpropagation: The network's predictions are compared to the actual outcomes, and an error is calculated. This error is then propagated backward through the network. The weights of connections are adjusted using optimization algorithms like gradient descent, aiming to minimize the error and improve predictions.

In the context of our study, a simple one-hidden layer deep neural network (DNN)– strictly speaking not DNN, two-hidden layer DNN in Keras and some Autokeras variations of DNNs with different hyperparameters in Python are being employed to predict the different health states, average damage and repair probabilities for the corresponding health states for future ELSA wave transitions. The input layer receives the selected features from a current ELSA wave, the hidden layer processes and transforms this information, and the output layer provides the predicted disease and ADL results. Because our health states to be predicted are binary variables, we use a sigmoid function as the activation function in output layer. The sigmoid function, often referred to as the logistic sigmoid function, is a mathematical function used in machine learning and NNs. It's primarily used to introduce non-linearity into the model and squash the output of a neuron or a neural network layer into a specific range bounded between 0 and 1. We also use binary cross entropy as the loss function, which measures the difference between the predicted probabilities and the actual binary (0 or 1) output values. For the training and test split, we chose the common ratio that is 80% of data as training data, and 20% as test data for the one and two hidden layer DNNs. For Autokeras, it adjusts all its hyperparameters by itself.

3.3 Random Forest

Random Forest (RF) is a powerful ensemble learning technique widely used in machine learning for classification and regression tasks. It is composed of multiple decision trees (DTs), which are non-parametric ML models, meaning they don't make strong assumptions about the underlying data distribution. DTs are versatile and can be used for both classification and regression tasks. The hierarchical tree structure, with root, internal, and leaf nodes, is a key feature of decision trees, making them easy to interpret and visualize. The DTs in RFs work collaboratively to make predictions. Each decision tree in the forest is built on a subset of the training data, and the final prediction is determined by aggregating the predictions of individual trees. This is generally how RF works:

1) Bootstrap Sampling: RF begins by randomly selecting subsets of the training data (with replacement) to create multiple training datasets. Each subset is used to

train an individual decision tree [79].

2) Feature Selection: At each node of a decision tree, a random subset of features is considered for splitting. This randomness reduces the risk of overfitting and encourages diversity among the trees [79].

3) Decision Tree Building: Multiple decision trees are built using the subsets of training data. These trees can be deep or shallow, depending on the hyperparameters chosen. Each tree is trained to predict the target variable based on the selected features [80].

4) Voting or Averaging: For classification tasks, the mode (most frequent prediction) among all decision trees' predictions is taken as the final prediction. For regression tasks, the average of all predictions is considered [80].

RF reduces the overfitting by training multiple trees on different subsets of data, and results in improved generalization to unseen data. It is also robust when handling missing values and noisy data and also provides robustness against outliers. RF provides a measure for feature importance, helping in understanding the most influential features in making predictions. Like DNN, RF can handle large datasets and high-dimensional feature spaces effectively.

Random Forests, while a powerful ensemble method, have several weaknesses. They can lack interpretability when combining multiple DTs, making it challenging to understand their decision-making process. Additionally, RFs can be computationally expensive, particularly for large datasets and a high number of trees. They can still overfit noisy data, and addressing class imbalance is necessary for imbalanced classification tasks where one of the classes (0 or 1) is rare.

We utilize the imbalanced random forest model for the prediction of 19 binary diseases and 25 binary ADLs due to the low prevalence of these diseases and ADLs, as evident in Tables 4.4 and 4.5. We address the class imbalance using the imbalanced dataset model from the Python imblearn library. Specifically, we employ the Easy Ensemble Classifier with a stratified 10-Fold cross-validation approach. Stratification is the sampling of same percentage of minority class for each 10 folded train-test sets.

3.4 Logistic Regression

Logistic Regression (LR) is a statistical method used for binary classification tasks, where the objective is to predict the probability that an instance belongs to a particular class. It models the relationship between the dependent variable (binary outcome) and one or more independent variables (features) [81]. Logistic Regression transforms the linear combination of predictor variables and their coefficients using the logistic function,

$$p(x) = \frac{1}{1 + e^{-(x-\mu)/s}} \tag{3.1}$$

also known as the sigmoid function, to produce a value between 0 and 1. Here, μ is the location parameter being the midpoint where $p(\mu) = 1/2$, and s is scale parameter defining the spread of the distribution. $p(x)$ value represents the predicted probability of the positive class. If the predicted probability is greater than or equal to a chosen threshold (usually 0.5), the instance is classified as the positive class. Otherwise, it's classified as the negative class.

The logistic regression model is trained using a training dataset, where the LR coefficients are estimated to minimize the difference between predicted and actual outcomes. This is usually done using optimization techniques such as Maximum Likelihood Estimation (MLE). To prevent overfitting, regularization techniques like L1 (Lasso) or L2 (Ridge) regularization can be applied, introducing penalty terms on the coefficients during training [82]. LR finds applications in various fields, including diagnosis of diseases in medical research including aging, credit scoring in finance, and customer churn prediction in marketing.

We explore whether logistic regression exhibits superior performance compared to our DNN model. To do that, we predict 19 disease outputs and 25 ADL outputs by using LR to compare with DNN. Moreover, we predict the damage and repair transition probabilities for those diseases and ADLs for a second comparison to our DNN model. While logistic regression is a potent approach for predicting binary target variables when explanatory variables are linearly dependent, the DNN model is typically better on complex and high-dimensional datasets.

3.5 Imputation, Model and Feature Selection, and Analysis

Data imputation is the process of filling in missing or incomplete data with estimated values to maintain the integrity and utility of a dataset. We utilize the **MICE** package in R for imputing missing values in our dataset, as summarized in Table 3.1. To evaluate different imputation methods, we compare them to a reference method where continuous variables are imputed using their mean and categorical variables using their mode [83, 84, 85]. After evaluating various imputation approaches, we proceed to construct a predictive model to test the best imputation method for our dataset. We choose the suitable one of DNN and LR models to test the best imputation method.

We consider two missing data mechanisms: missing at random (MAR) assumes that the missing data is independent of the unobserved data, and missing not at random (MNAR) assumes that the missing data is systematically dependent of the unobserved data. While MAR is built in **MICE**, MNAR is applied by hand to some missing variables then the remaining missing variables are considered as MAR. For example, if a person never had a cancer and "ever had cancer medication" variable is missing for this person, this means the value of this missing variable is in fact "No". We also employ two imputation methods: predictive mean matching (PMM) and classification and regression trees (CART) imputation methods within the **MICE** package. Our evaluation focuses on predicting 19 disease health states in each wave using the imputed predictors. We compare the average evaluation metrics for all binary predictions derived from the imputed predictors across all waves. Based on these measures, we select the best imputation method.

Once we have a complete dataset obtained through the best imputation model over all waves, we can proceed with feature selection to predict our target variables. For feature selection, we employ two different feature selection methods: model agnostic Filter Method [86] which means it does not use a predictive model but uses the correlation and variance between features and targets, and recursive feature elimination (RFE) Method [87] which is not model agnostic and is computationally burdensome; so we only considered it for the (simple) LR model. RFE is an iterative technique that eliminates less significant features in order to create a subset of features that maximizes predictive accuracy.

After obtaining the best features, we evaluate different predictive models for the health state, or damage and repair transition probabilities. We first compare different DNN models, from a simple one-hidden layer DNN to more complex automatically generated Autokeras DNN models. We also, repeat the same evaluation by removing nurse data from the explanatory data completely to see its effect on predictions. We then compare the best DNN with LR and RF models. With the best predictive model, we go on full analysis of the thesis by predicting multiple health states, average damage and repair probabilities, and individual damage transition probabilities.

The benefit of good prediction scores for multiple health states and their calibrated individual damage transition probabilities is to stratify risks between different health states. Risk stratification is the process of sorting individuals into groups according to their chances of encountering specific events or outcomes. This approach enables the tailored interventions and treatments for different risk categories. A metric commonly used for risk stratification is the Hazard Ratio (HR), which measures the ratio of hazard rates between two groups. Generally, HR values assist in gauging the extent of risk association, with an HR exceeding 1 indicating a greater risk in the exposed group compared to the reference group

Chapter 4

Results and Analysis

We present our results and analysis as the written manuscript submitted to journal in Mechanisms of Ageing and Development as a research paper.

4.1 Motivation of Paper

Application of machine learning (ML) methods to high dimensional and complex health and medicine datasets can go in many directions. Clinical implementations of different ML models can predict risk in terms of disease or disability occurrence with damage transitions, identify key information from a patient's chart, providing clinical notes, enable real-time detection of patients at risk for clinical deterioration and perform personalized predictions [15]. Good high-dimensional prediction performance of ML models can draw the physicians' attention to specific cases, even while targeting multiple health outputs simultaneously. That said, our purpose is not to develop a clinical application. Rather it is to explore ML models for health state predictions.

We explore the best predictive model for the most suitable health variable among binary health states (19 diseases and 25 ADLs in ELSA data), continuous average damage and repair probabilities over whole population for each health state, and binary individual damage and repair transition probabilities for each health state during the different waves (visits) of individuals in ELSA study. One may obtain a very good predictive model for all of the three types of health variables, or one very good model for one of the variables and may obtain other variables indirectly from the first predicted variable. After obtaining damage and repair probabilities, we will be able to analyze risk stratification by obtaining the hazard ratio in terms of damage transition probabilities. We check if there is a relation between prevalence of health outputs and HRs. We will also be able to analyze the correlations between different health states for observed and predicted cases to see if there is consistency and utility.

4.2 Credits to Authors

The manuscript is authored by myself and my thesis advisor Dr. Andrew Rutenberg. The codes was written by me while the flow, direct editing and feedback to the manuscript was shaped by Dr. Rutenberg. I did all of the numerical analysis, data wrangling and figure preparation. The manuscript is included here in full. The format for the figure and table names, equations, and citations have been adapted for this document to be in accordance with the rest of the thesis.

4.3 Paper

4.3.1 Abstract

We predictively model damage transition probabilities for binary health outputs of 19 diseases and 25 activities of daily living states (ADLs) between successive waves of the English Longitudinal Study of Aging (ELSA). Model selection between deep neural networks (DNN), random forests, and logistic regression found that a simple one-hidden layer 128-node DNN was best able to predict future health states (AUC ≥ 0.91) and average damage probabilities ($R^2 \geq 0.92$). Feature selection from 134 explanatory variables found that 33 variables are sufficient to predict all disease and ADL states well. Deciles of predicted damage transition probabilities were well calibrated, but correlations between predicted health states were stronger than observed. The hazard ratios (HRs) between high-risk deciles and the average were between 3 and 10; high prevalence damage transitions typically had smaller HRs. Model predictions were good across all individual ages. A simple one-hidden layer DNN predicts multiple binary diseases and ADLs with well calibrated damage and repair transition probabilities.

4.3.2 Introduction

Aging is the decline in healthy functioning of an organism with time [2]. Disease and disability occurrences are important discrete events during aging. Age is also a significant factor that impacts disease susceptibility, progression, and outcomes across various health conditions. The prevalence of many chronic diseases and dysfunction

increase with age [3, 4]. Underlying this age dependence is a high-dimensional biological process of increasing dysfunction [88, 89, 1, 90]. Using this high-dimensional information is important for understanding how to better predict and mitigate transitions such as disease and disability.

Using machine learning (ML) methods on high-dimensional health and medicine datasets dates is increasingly popular. ML models can perform very well on hundreds or thousands of explanatory variables simultaneously, and even for multi-dimensional target variables [16, 17]. While training models with cross-sectional data can give insights about relationships among the present health states, training with longitudinal data is needed to give insights about *future* health states [91]. Determining the "best" predictive model depends on various factors, including the nature of the data and the specific outcomes predicted. Different machine learning algorithms may be appropriate depending on the characteristics of the data and the specific predictive task [82].

We are motivated to address questions raised by three recent studies investigating aging dynamics. In the first, Farrell et. al. [11] built a computational dynamic joint interpretable network (DJIN) model to predict aging trajectories of continuous health variables from the English Longitudinal Study of Ageing (ELSA). Though comprehensive, the approach was complex and difficult to replicate – furthermore only continuous variables were predicted. We hypothesize that predictive models built around binary health-states could be significantly simpler. In the second study, Farrell et. al. [12] characterized damage and repair rates of discrete activities of daily living (ADLs) in the ELSA dataset. This study showed significant heterogeneity among damage rates, indicating the possible need for ML methods for predictive studies of binary health states. Disease transitions were not characterized. In the third study, Buergel et. al. [21] used deep-neural networks (DNN) to risk-stratify individuals for 24 common conditions, including diseases, using UK Biobank metabolomic profiles. We wanted to explore similar questions with ELSA data – without metabolomic inputs but including ADLs.

Risk stratification categorizes individuals based on their likelihood of experiencing certain events or outcomes. This facilitates tailoring interventions and treatments to different risk groups. One metric of risk stratification is the Hazard Ratio (HR),

which quantifies the ratio of the hazard rates between two groups. Typically, HR values help in identifying the degree of risk association, where an HR greater than 1 signifies a higher risk in the exposed group compared to the reference group. Calibration, also important, assesses the agreement between predicted and observed risks by damage transition probabilities [92]. A well-calibrated model provides accurate risk predictions, facilitating decision-making.

ELSA [10] includes many binary, continuous, and categorical variables describing physical and mental health states, ability to do daily activities, and demographic and socioeconomic information that are repeatedly measured over different waves. As with many population studies [8], there are tens of thousands of individuals with hundreds of variables (dimensions). Our focus will be on initial waves that have both core (typically binary) and nurse (typically continuous lab) variables. We will predict subsequent individual health states by using explanatory data from the previous waves. From these health states, we will obtain predicted individual damage probabilities (i.e. individual risks).

There are two questions we want to address. The first is whether a simple deep-learning pipeline performs well for discrete health states and transitions using the available ELSA data – in particular whether it represents a significant improvement over simpler logistic-regression models (see e.g. [22, 23]) or not [24, 25]. The second is what aspects of high-dimensional input data are useful for these predictions – in particular whether discrete or continuous inputs are more useful.

4.3.3 Methods and Model Selection

We consider the core and nurse data from the ELSA dataset for our study. Because the even numbered waves include the nurse data we used variables in even numbered waves as explanatory data, and health states variables in odd numbered waves as target data to predict. We consider different machine learning (ML) architectures: deep neural networks (DNN), random forest (RF) and logistic regression (LR) as predictive models. We use LR as a baseline comparison model because it is standard, interpretable, less prone to overfitting, and can sometimes surpass ML models as a clinical prediction model [24]. In contrast, DNN and RF are computationally efficient

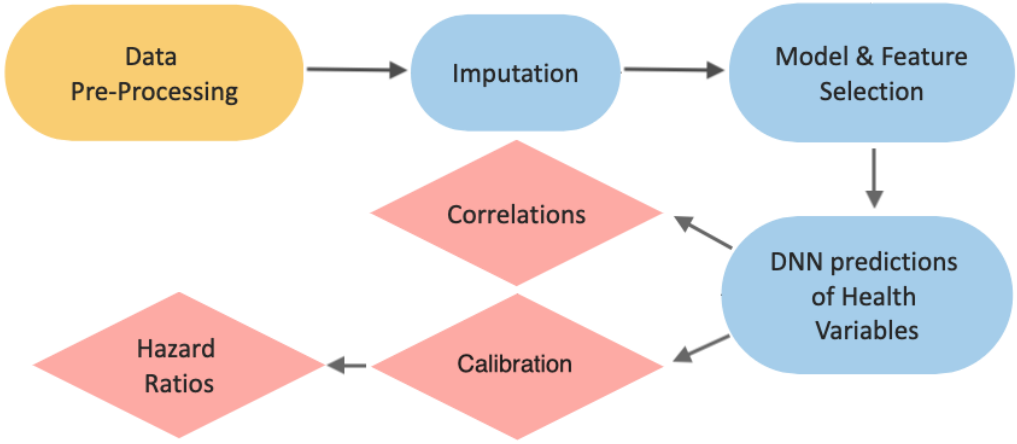

Figure 4.1: **Study flowchart.** In pre-processing (yellow oval), we choose target and explanatory variables, convert string variables to numbers, and normalize continuous data. We then determine the best imputation method using the MICE package in R and apply it to all data. With imputed data we apply feature and model selection to determine the best model, model architecture, and a sufficient feature-set for good predictions. Using the best model (DNN) we then predict the 19 disease and 25 ADL health states. For these predictions, we specifically explore (red diamonds) calibration, hazard ratios, and correlations.

predictive models appropriate for more complex non-linear tasks – such as (we hypothesize) disease and disability during aging. We follow the study flowchart given in Fig. 4.1.

Dataset

The English Longitudinal Study of Ageing (ELSA) [10] provides information on the health dynamics and well being of an English population over 50 year old. Since 1998, there have been 9 waves separated by approximately 2 years. A summary of the dataset can be found in Table 4.1. Waves 2, 4, 6, 8 and 9 are associated with nurse visits ("nurse" waves) where continuous lab data such as blood pressure, glucose level, or cholesterol level is collected and body measurements like height, weight, hand grip strength are taken. Data pre-processing included harmonizing variable names across waves, converting string data types to discrete numerical values, and normalizing continuous variables.

We considered 134 explanatory variables with missingness [74] less than 30% from core and nurse data: 78 binary variables (including diseases and ADLs), 14 continuous variables from the core data and 42 variables mostly continuous from the nurse data. We used these 134 binary and continuous explanatory variables to predict the

	W0	W1	W2 → W3		W4 → W5		W6 → W7		W8 → W9	
Individuals #	8267	12099	9432	9771	11050	10274	10601	9666	8445	8736
Same Individuals #	-	5053	9324	7680	7908	9460	8999	8866	8088	7146
New Individuals #	-	7046	108	2091	3142	814	1602	800	357	1590
Women %	57	56	56	56	55	56	55	56	56	56
Men %	43	44	44	44	45	44	45	44	44	44
Ever Smoked %	71	70	70	70	66	58	57	54	52	51
Core Missing %	23	23	22	22	22	21	20	20	19	20
Nurse Missing %	-	-	8	-	9	-	9	-	10	10

Table 4.1: **Summary of ELSA data.** WN stands for Wave N of the ELSA dataset. We model four transitions that start with waves that have nurse (lab) data (waves 2, 4, 6, and 8), as indicated by the arrows. Individuals # is the number of study participants in each wave. Same individuals # is the number of participants who were also in a previous wave. New Individuals # is the number of newly added participants in the indicated wave. Core Missing % and Nurse Missing % are the missing data percentage of core data and nurse (lab) data in the corresponding wave after choosing our 134 explanatory variables. We used W0 and W1 for additional imputation input.

target health state variables in subsequent odd numbered waves. The target variables (19 diseases and 25 ADLs) are listed in Supplementary Tables 4.4-4.5, along with their prevalences for waves 1-9, and the remaining explanatory variables are in Supplementary Table 4.6.

Imputation

We first perform a selection procedure to find the best imputation method and missingness mechanism for our 134 explanatory variables. We evaluate different imputation methods within the multiple imputation by chained equations (MICE) package in R version 4.2.0 (2022-04-22)[93], together with a simple reference method of imputing continuous variables by means or categorical variables by modes.

We consider both missing at random (MAR) and missing not at random (MNAR) mechanisms for MICE imputation methods. For both mechanisms, we consider both predictive mean matching (PMM) and classification and regression trees (CART) as imputation methods. We tested the best imputation method by predicting the 19 diseases' binary health states in each wave from the imputed predictors. We compare the average accuracy, Youden's index (J) and the area under curve (AUC) values for all binary predictions over all waves. Note that $AUC \approx (J+1)/2$ [94] and ranges from 0 to 1 (best). Youden's index ranges from -1 to 1 (best).

We used simple one-hidden layer DNN and LR as the predictive model for selecting the best imputation method. By passing the imputed variables in all waves (from 0-th to 9-th waves), we predict the 19 disease states in the same wave. Then, we obtain the accuracy, Youden's index, and AUC values for 9 waves and take their averages

for both DNN and LR used as predictive models. We see in Table 4.2 for DNN and in Supplementary Table 4.7 for LR that MICE with MNAR is significantly better than both MICE with MAR and with the simple mean/mode matching. PMM and CART methods are quite similar. The CART method has been shown to perform better in imputation of NHANES aging data [95]. We will therefore use MICE with MNAR and the CART method to impute missing data of all input variables, and for all models.

	Accuracy		**Youden's J**		**AUC**	
Mean/Mode	0.91		0.43		0.76	
	PMM	**CART**	**PMM**	**CART**	**PMM**	**CART**
MICE-MAR	0.90	0.90	0.35	0.36	0.70	0.70
MICE-MNAR	0.93	0.93	0.53	0.52	0.80	0.79

Table 4.2: **Performance of imputation methods.** Performance is evaluated with accuracy, Youden's index, and AUC (higher is better) for DNN predictions of 19 disease states in all waves. Mean/Mode is a simple reference imputation method using mode for categorical and mean for continuous missing variables. For imputation with MICE, we considered MAR and MNAR missingness mechanisms, and PMM (default) and CART methods.

Model and Feature Selection

Our prediction models generally determine the predicted binary health variables X_i (at the next wave) and its transition probability P_i , for a given individual i. Given the binary health variables X_i at the current wave, we obtain damage probability $D_i = (1 - X_i)P_i$ when $X_i = 0$ and repair probability $R_i = X_i(1 - P_i)$ when $X_i = 1$. To assess model and feature selection, we average damage and repair probabilities over the entire population:

$$D = \frac{\sum_i (1 - X_i)P_i}{\sum_i (1 - X_i)}$$
$$R = \frac{\sum_i X_i(1 - P_i)}{\sum_i X_i}. \tag{4.1}$$

We can compare these with observed averages, in which case P_i and X_i are binary. For feature and model selection, we only assess the transitions from wave 2 to 3. In this paper, our focus is on damage transitions.

a) Feature Selection

We performed feature selection for wave transition from 2 to 3 by choosing various

number of features for each disease and ADL. We considered both filter [86] and recursive feature elimination (RFE) methods [87]. For our preferred, model agnostic, filter method, correlations or variance between features and targets are used to select features. For binary predictions, we used a f_classif metric. The SelectKBest function in the Python Sklearn library was used to select a specified number of best features. The RFE method is not model agnostic and is computationally burdensome; we only considered it for the (simple) LR model.

By increasing the number of features per disease or ADL, we found that AUC rapidly saturated vs the total number of features for both DNN (see supplemental Figs. 4.4a and 4.4d) and LR (Figs. 4.4b and 4.4e) models. DNN predictions were better than LR predictions. While the AUC of disease prediction quickly plateaued at approximately 40 total features for both DNN and LR, the AUC for ADL prediction slowly increased (note small range of AUC values) as the number of features continued to increase. The performance of the RFE method with LR was worse than the filter method with LR in terms of reaching saturation very late (compare Figs. 4.4(b,e) and Figs. 4.4(c,f)). Accordingly, we used filter feature selection for all models.

The filter selected features for both disease and ADL predictions are given in Table 4.8 for both $N = 33$ and $N = 41$ total features. Remarkably, the same total features are selected for both disease and ADL predictions at these stages. $N = 33$ arise from selecting $k = 2$ and $k = 9$ features per disease or ADL, respectively, and dropping the duplicates. $N = 41$ arise from selecting $k = 3$ and $k = 15$, respectively. Most of the $N = 33$ features are ADL states themselves, but with cognitive diseases, pain, and grip strength as well. Added features for $N = 41$ include activity level, self-reported health, and systolic blood pressure and pulse. Significantly, very few disease states are selected – even for the goal of disease state prediction. We suspect that the small prevalence of diseases makes disease states themselves inefficient predictors. Strikingly, individual age is not selected. In Fig. 4.4, we observed that the reduced number of features are almost as good as full features in predictive performance for both diseases and ADLs.

b) Model Selection

We considered a simple 1-hidden Layer DNN, a 2-hidden layer DNN, and an automatically generated DNN using Autokeras. In all cases, we directly predicted

health states, and obtained average damage and repair probabilities using the full feature set ($N = 134$). We used binary crossentropy [96] as our loss function for DNN. We train models by using a randomly selected 80% of the total population, and use the remaining 20% as test data. We obtained the best model as the one-hidden layer DNN (1-Layer DNN) with the highest AUC (for binary health predictions) and best R^2 (for average damage probabilities, by disease or ADL), see Supplementary Table 4.9. A 2-Hidden Layer DNN gives comparable AUC results for health state predictions. An Autokeras DNN model gave the poorest prediction results in AUC (see Table 4.9). Although the AUC values of health state predictions of one and two hidden layer DNNs are close to each other, the R^2 values of average damage and repair probabilities estimations are slightly higher (better) for 1-Layer DNN models, see Supplementary Table 4.9.

We also investigated a random forest (RF) model to predict diseases and ADLs. Because prevalences are small, see Tables 4.4 and 4.5, we use the imbalanced dataset model in the imblearn [97] library in Python – we also used the Easy Ensemble Classifier with a stratified 10-Fold cross validation. For all wave transitions, our best DNN model always performs better than RF, see Supplementary Table 4.10. We also considered logistic regression (LR). We obtained poorer results than DNN both when predicting diseases and ADLs directly (Table 4.11), and when predicting damage and repair transition probabilities (Table 4.12). We obtained similar (worse than DNN) results even when using an LR-tuned imputation and feature selection pipeline (see Tables 4.2 and 4.7, and Fig. 4.4).

We therefore use a simple one-hidden layer DNN, as it was the best model to predict damage and repair probabilities for all wave transitions.

Calibration and Hazard Ratio

We investigate the calibration of our best model using 5-fold cross validation over all waves. For a given disease or ADL we rank ordered the population with respect to predicted transition probabilities. We grouped the rank-ordered population in deciles, and compared the averages within the deciles between the model and the observed transitions. A well calibrated model has predicted and observed probabilities that are comparable [98]. A quantitative measure of calibration is the Brier score [99], which

is the mean squared difference between observed and predicted probabilities. It takes values between 0 and 1, where 0 indicates the best calibration and 1 for the perfect inaccuracy in the calibration .

The hazard ratio (HR) can be used to help explain the relationship between exposures, interventions, and the occurrence of specific events. HR in our study is the ratio of the risk of the top decile over the mean risk. If all of the risk is concentrated in the top decile, then the HR will be 10. Accordingly, the HR must be between 1 and 10.

4.3.4 Results

We apply our best model, the one-hidden layer DNN with 128 nodes for the full features – for the wave transitions $2 \to 3$, $4 \to 5$, $6 \to 7$, and $8 \to 9$. The target variables are binary health states X_i for the ith individual. The transition probabilities of health states P_i can be obtained by comparing the previous and predicted future states X_i, and hence average damage and repair probabilities D and R of 19 diseases and 25 ADLs can be obtained according to Eq. 4.1. We use AUC values to evaluate the prediction performance of binary health states, while R^2 values are used to evaluate the average damage and repair probabilities of the same health variables. For diseases we find that all AUC are above 0.9 for all wave transitions, while R^2 are close to 1 – see Table 4.3. For ADLs the performance is comparable, though slightly worse. Our AUC scores are substantially better than recent deep-learning models predicting disease risk [100, 101], including a recent study by Google researchers using a large language model to analyze UK Biobank data [16]. Remarkably, our prediction performance has very little dependence on the age of individuals – see supplemental Fig. 4.9.

In Fig. 4.2, calibration curves [92] for average of damage transition probabilities over all health states show that our model is well calibrated for both diseases and ADLs. Here, we have rank ordered predicted transition probabilities of each disease or ADL, grouped the individuals in deciles, and then averaged both predicted damage probabilities and observed damage prevalences within the decile and across diseases or ADLs. In supplemental Figs. 4.5 and 4.6 we have instead averaged the damage probabilities within individual diseases or ADLs – and also for repair probabilities. You

DISEASES		W23		W45		W67		W89	
		Damage	Repair	Damage	Repair	Damage	Repair	Damage	Repair
1 Layer	AUC	0.92		0.93		0.93		0.91	
	R^2	0.99	0.98	0.96	0.96	0.92	0.94	0.99	0.98

ADLs		W23		W45		W67		W89	
1 Layer	AUC	0.90		0.90		0.91		0.90	
	R^2	0.94	0.89	0.97	0.87	0.97	0.94	0.96	0.87

Table 4.3: **Model performance for diseases and ADLs.** The evaluation performance of the best (one-layer) DNN model for full 134 features. AUC is the metric for the predictions of binary health outputs, R^2 is for the average damage and repair probabilities obtained from predicted health outputs. W23 stands for the wave transition from 2 to 3, and so on.

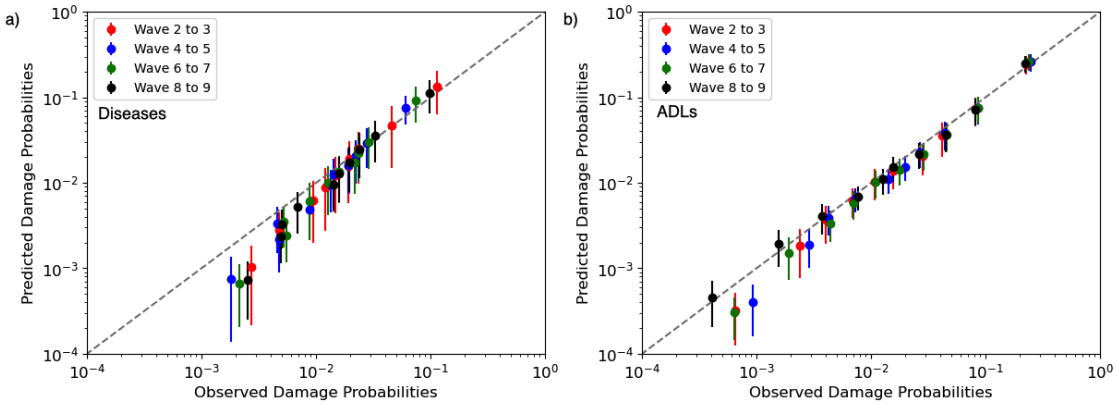

Figure 4.2: **Calibration curves for damage transition probabilities.** The points represent the deciles of the damage probabilities. a) For diseases, we average each decile across all diseases. See supplemental Fig. 4.7 for individual calibrations and supplemental Fig. 4.5 for averages across all deciles. b) For ADLs, we average each decile across all ADLs. See supplemental Fig. 4.8 for individual calibrations and supplemental Fig. 4.6 for averages across all deciles. Dashed line indicates perfect calibration.

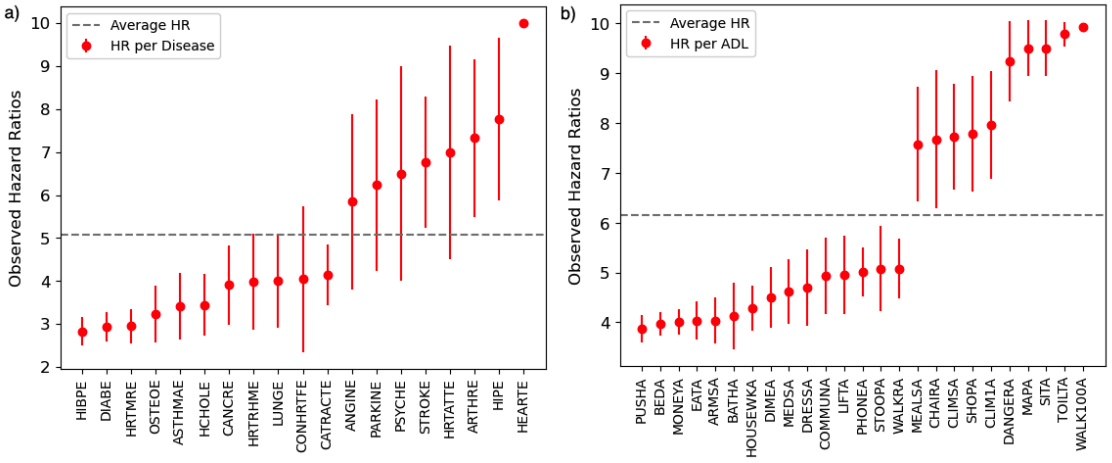

Figure 4.3: **Hazard ratios.** Observed hazard ratios (HR) between the highest decile of transition probability to the median for a) Diseases and b) ADLs, averaged over all wave transitions. For individual waves see Supplementary Fig. 4.11. Error bars are standard errors from five-fold cross-validation. The dashed line indicates the average HR.

can also see the individual disease and ADL calibration curves for damage transition probabilities of indicated waves in supplemental Figs. 4.7 and 4.8, respectively.

We obtain a quantitative calibration measure [92] for our predictions in Supplementary Table 4.13 using the Brier score, which takes values between 0 and 1, with smaller values better. As seen in Table 4.13 the average Brier score is low for both diseases (≤ 0.020) and ADL (≤ 0.035), but diseases are on average better calibrated than ADLs. This is also reflected in the Brier scores of individual diseases and ADLs in Supplementary Tables 4.14 and 4.15.

From deciles of rank ordered damage transition probabilities, we obtain the hazard ratio (HR) values of each health variable by dividing the average probability of the maximum decile to the mean probability. These are shown in Fig. 4.3 for all diseases and ADLs, averaged over all wave transitions, together with standard errors from five-fold cross-validation. The dashed lines indicate the average HR. Some health conditions have low HRs, indicating that less discrimination between high and low-risk individuals was possible. Other health conditions have much higher HRs, indicating that more discrimination was possible. These results are broadly consistent across different waves (see supplemental Fig. 4.11). Interestingly, we find that the disease HRs are inversely correlated with disease prevalence (Spearman's -0.43, see supplemental Fig. 4.10) – though the same is not seen with ADLs (Spearman's 0.06).

For the wave $2 \rightarrow 3$ transition, we have analyzed the distribution of predicted

damage transition probabilities P_i's for all diseases (see supplementary Fig. 4.12) and ADLs (supplementary Fig. 4.14). We generally find long-tailed, right-skewed distributions. With log-log plots (Figs. 4.13 and 4.15) it becomes clear that the larger probabilities are well approximated by a power-law tail with exponents typically between $2-3$. From these distributions, we also predict bimodal character in HIBPE, DIABE and ANGINE in Fig. 4.12 that naturally identifies lower vs higher risk groups. We see similar high versus low risk character for HIBPE, DIABE and ANGINE from the observed and predicted HRs of high risk versus low risk group for these diseases in Supplementary Table 4.16 – though with HR ≤ 2.

We have also considered correlations between observed (supplementary Fig. 4.16) and predicted (supplementary Fig. 4.17) health states. We observe strong correlations between ADLs, and weaker correlations between disease states. Generally the predicted correlations are substantially stronger than observed correlations, which indicates that joint-risk is not well calibrated. However, the hierarchical clustering is similar. The DNN model may therefore be using some aspects of the correlations to improve prediction performance of individual health states and health transitions.

4.3.5 Discussion

In this study, we predict the damage and repair probabilities for different health states using the ELSA dataset. Starting with 134 explanatory variables from a given wave, we predict future binary health states of 19 diseases and 25 ADLs in the subsequent wave. We considered several candidate models. A simple deep-neural network (DNN) with 1-hidden layer predicted future states best both with the full feature set and a selected set of 33 features (see Fig. 4.4(a,d), compared to other more complex DNN models, a random forest (RF) model, or a logistic regression (LR) model. Our results did not support the claim that LR is better at clinical prediction than machine learning [24]. Our best DNN always exhibited better predictive performance than LR (see Tables 4.11, 4.12, and Fig. 4.4). We used our best DNN model to obtain individual damage and repair probabilities for each health state. We obtained good AUC scores (approximately 0.90) for the binary health state predictions, and excellent R^2 for average damage and repair probabilities.

Our selected features were largely binary health variables, with very few continuous lab/nurse variables (see supplemental Table 4.8 and Fig. 4.4). This is consistent with excellent performance when nurse data was omitted altogether (see the second section in Table 4.9). While initially surprising, it is consistent with the very strong correlations between health states observed in the training data (supplemental Fig. 4.16). Notably, while correlations among ADL variables are the strongest – the correlations between ADL and disease states are comparable to those between different disease states. We hypothesize that our DNN model has learned from this correlation structure to better predict health states – which is an advantage of developing a model that predicts all health states (disease and ADL) at once.

Our DNN damage transition probabilities for both diseases and ADLs were well calibrated (see Fig. 4.2 and Table 4.13). This means that predicted transition probabilities corresponded to observed transition probabilities when considered in rank-ordered deciles of the predicted probabilities. We used this to determine hazard-ratios (HRs) of the maximum decile to the mean. We find HRs between 3 and 10 (the largest possible for deciles) – indicating substantial risk stratification is possible with our approach. Our well-calibrated model provides accuracy in risk prediction results which would provide reliability for decision-making [102, 103]. Since we had calibrated probabilities, we also considered the distributions of transition probabilities. We found that the shape of the damage transition probability distribution was long-tailed and right-skewed – with a power-law tail with exponents typically between 2 and 3. We do not suggest any mechanism for these power-law tails.

In the feature selection of our study, age was not selected as one of the best predictors during feature selection (see Table. 4.8. While age is a significant risk factor for most diseases and ADLs [3, 4], the disease and ADL context indirectly provide sufficient information about that risk. Remarkably, our predictive quality does not strongly depend on age (see Fig. 4.9). The age at disease onset may nevertheless influence disease trajectory and long-term outcomes.

Our approach has limitations. We trained each wave transition separately. We have not validated with other datasets. While damage transition probabilities were well calibrated, the repair transition probabilities were not (hence we have postponed extensive discussion of them) – and neither were the correlations between health

states. It is also important to note that we predicted the disease and ADL states as reported within the data. The connection between these observed health states and the onset of the "real" underlying diseases or health conditions could not be explored.

We investigated how much information about the future health states can be found in current states, and what sort of models are best designed to make the prediction. We found that the simple one-hidden layer DNN worked best, which is encouraging for development of interpretable and useful (i.e. translational) ML models in the future. Our model is simpler than some competing ML approaches, but seems to perform as well or better. It will be interesting to explore whether our simple approach can be substantially improved, for example by using information from multiple previous waves. It will also be important to explore whether readily available clinical information can lead to sufficiently good predictions to be useful. We have shown that the most significant features were generally the ADL health states together with cognitive function, pain level and grip strength.

4.3.6 Acknowledgements

ADR thanks the Natural Sciences and Engineering Research Council (NSERC) for an operating Grant (RGPIN 2019-05888).

4.3.7 Supplementary Material

Tables 4.4-4.5: Percentage prevalence of Diseases and ADLs

Table 4.6: Used core and nurse variables with their code and real name.

Table 4.7: Performance of imputation methods based on LR predictions

Table 4.8: Selected features in feature selection.

Table 4.9: Evaluation of best DNN model selection.

Table 4.10: Comparison of the performances of RF and the best DNN model.

Tables 4.11-4.12: Comparison of the performances of LR and the best DNN model for health states and transition probabilities (2 times target variables), respectively.

Table 4.13: Calibration scores for the average of all diseases and ADLs.

Tables 4.14-4.15: Individual disease and ADL calibration scores.

Table 4.16: Observed and predicted HRs of high versus low risk groups for the bimodal distributions in Fig. 4.12.

Figure 4.4: AUC vs #n features for filter DNN, filter LR and RFE LR feature selection methods.

Figures 4.5-4.6: Scatter plot for disease and ADL damage repair probabilities.

Figures 4.7-4.8: Individual disease and ADL calibration curves for damage transition probabilities.

Figure 4.9: Prediction performances of diseases and ADLs according to age.

Figure 4.10: Hazard ratio versus prevalence of diseases and ADLs.

Figure 4.11: HR versus all diseases and ADLs over all individual waves.

Figures 4.12-4.13: Histograms of predicted damage transition probabilities for each disease in log-lin and log-log scales.

Figures 4.14-4.15: Histograms of predicted damage transition probabilities for each ADL in log-lin and log-log scales.

Figures 4.16-4.17: Correlation map between observed and predicted health states.

	Health Output	W1	W2	W3	W4	W5	W6	W7	W8	W9	Average
PARKINE	ever had Parkinson's disease	0	1	1	1	1	1	1	1	1	1
CONHRTFE	ever had congestive heart failure	1	1	1	1	1	1	1	1	1	1
HIPE	ever had hip fracture	2	1	1	1	1	1	1	1	1	1
HEARTE	ever had heart problems	0	1	2	3	3	4	2	2	6	2
HRTMRE	ever had heart murmur	5	4	5	4	4	4	4	4	3	4
HRTATTE	ever had heart attack	0	6	6	6	6	5	5	5	1	4
STROKE	ever had stroke	4	5	5	4	5	5	5	5	3	5
LUNGE	ever had lung disease	6	7	5	5	5	5	5	5	5	5
CANCRE	ever had cancer	6	7	5	5	6	6	6	7	8	6
ANGINE	ever had angina	9	8	8	7	9	6	3	3	1	6
OSTEOE	ever had osteoporosis	5	6	6	7	8	7	8	8	8	7
HRTRHME	ever had abnormal heart rhythm	6	7	8	7	8	8	9	10	8	8
PSYCHE	ever had psych problems	8	9	8	9	10	10	10	10	8	9
DIABE	ever had diabetes	7	6	9	9	11	11	12	13	12	10
ASTHMAE	ever had asthma	12	13	11	11	11	11	11	11	11	11
CATRACTE	ever had cataracts	13	17	17	17	20	22	25	30	30	21
HCHOLE	ever had high cholesterol	0	18	29	32	36	34	35	36	32	28
ARTHRE	ever had arthritis	32	36	35	34	37	37	37	40	38	36
HIBPE	ever had high blood pressure	37	32	40	38	40	38	38	39	36	37

Table 4.4: Percentage prevalence of 19 disease outputs: WN represents Wave N in ELSA data. Prevalences are given for waves 1-9 together with the Average prevalence over all waves. Diseases have been rank ordered in increasing average prevalence.

	Health Output	W1	W2	W3	W4	W5	W6	W7	W8	W9	Average
DANGERA	Some Diff-Recognizing when in physical danger	0	0	0	2	2	2	2	2	2	1
EATA	Some Diff-Eating	2	2	2	2	2	3	2	3	3	2
MEDSA	Some Diff-Take medications	1	2	2	2	3	3	3	3	3	3
COMMUNA	Some Diff-Communication (speech, hearing, eye)	0	0	0	4	4	4	4	4	4	3
PHONEA	Some Diff-Use a telephone	2	2	3	3	3	3	3	3	3	3
MONEYA	Some Diff-Managing money	2	3	4	3	4	4	4	4	4	4
WALKRA	Some Diff-Walk across room	3	4	3	3	3	4	4	4	4	4
TOILTA	Some Diff-Using the toilet	3	3	4	3	4	4	4	4	4	4
MEALSA	Some Diff-Prepare hot meal	4	5	5	5	5	5	5	6	6	5
MAPA	Some Diff-Use a map	5	6	5	5	5	5	5	5	5	5
BEDA	Some Diff-Get in/out bed	2	6	6	6	6	6	6	6	6	6
DIMEA	Some Diff-Pick up a 5p coin	5	5	5	5	6	6	6	6	6	6
SHOPA	Some Diff-Shop for grocery	9	10	10	9	10	9	9	10	9	9
BATHA	Some Diff-Bathing, shower	12	12	11	10	10	10	9	9	9	10
ARMSA	Some Diff-Rch/xtnd arms up	11	11	11	10	11	12	11	11	10	11
DRESSA	Some Diff-Dressing	13	14	13	13	13	13	12	13	13	13
WALK100A	Some Diff-Walk 100y	12	12	12	12	14	14	13	14	13	13
SITA	Some Diff-Sit for 2 hours	14	14	14	12	13	13	13	13	11	13
CLIM1A	Some Diff-Clmb 1 flt str	15	15	14	14	15	15	15	15	14	15
HOUSEWKA	Some Diff-Doing work around the house or garden	16	17	15	15	16	15	15	16	15	15
PUSHA	Some Diff-Push/pull lg obj	18	19	18	17	18	18	17	18	16	18
LIFTA	Some Diff-Lift/carry 10lbs	26	25	24	23	23	23	22	23	21	23
CHAIRA	Some Diff-Get up fr chair	26	27	25	25	25	24	24	24	23	25
CLIMSA	Some Diff-Clmb sev flt str	36	38	34	34	35	32	31	33	30	34
STOOPA	Some Diff-Stoop/Kneel/Crch	35	38	35	35	38	37	37	39	37	37

Table 4.5: Percentage prevalence of 25 ADL outputs: WN represents Wave N in ELSA data. Prevalences are given for waves 1-9 together with the Average prevalence over all waves. ADLs have been rank ordered in increasing average prevalence. 'Diff' = 'Difficulty'.

CORE DATA

SHLT	Self-report of health	Categ	Ordinal
RXHIBP	Takes meds for high blood pressure	Categ	Binary
RXDIABI	Takes insulin for diabetes	Categ	Binary
RXDIABO	Takes oral meds for diabetes	Categ	Binary
RXDIAB	Takes meds for diabetes	Categ	Binary
RXLUNG	Takes meds for lung condition	Categ	Binary
RXASTHMA	Takes meds for asthma	Categ	Binary
TRCANCR	Received treatment for cancer	Categ	Binary
RXBLDTHN	Taking medication for heart, anticoagulant	Categ	Binary
ALZHE	Ever had Alzheimer's	Categ	Binary
DEMENE	Ever had dementia	Categ	Binary
MEMRYE	Ever had memory problem	Categ	Binary
JOINTRE	Ever had any joint replacement	Categ	Binary
HIPRE	Ever had hip replacement	Categ	Binary
INDAGER	Definitive age variable collapsed at 90 plus.	Continous	Integer
CATRCTE	Ever had cataract surgery	Categ	Binary
SIGHT	Self-rated eyesight	Categ	Ordinal
DSIGHT	Self-rated distance eyesight	Categ	Ordinal
NSIGHT	Self-rated near eyesight	Categ	Ordinal
HEARING	Self-rated hearing	Categ	Ordinal
FALL	Fallen down in last 2 years	Categ	Binary
FALLNUM	Number of falls	Continous	Integer
FALLINJ	Injured from fall	Categ	Binary
FALLEQ	Uses equipment for falls	Categ	Binary
PAINFR	Frequent problems with pain	Categ	Binary
PAINLV	Usual level of pain	Categ	Ordinal
URINAI	Any urinary incontinence	Categ	Binary
VGACTX_E	Freq vigorous phys activ	Categ	Ordinal
MDACTX_E	Freq moderate phys activ	Categ	Ordinal
LTACTX_E	Freq light phys activ	Categ	Ordinal
MMALONE	Whether able to walk alone (with aid)	Categ	Nominal
MMHSS	Whether health condition prevents from walking	Categ	Nominal
MMWILL	Whether willing to do walking test	Categ	Nominal
MMSAF	Whether interviewer feels it is safe to do walking test	Categ	Nominal
MMAVSP	Whether interviewer feels suitable space available	Categ	Nominal
WALKCOMP	Willing and able to complete walking speed test	Categ	Binary
MMTRYA	Outcome of first walk	Categ	Binary
WSPEED1	Time for walking speed test 1(sec)	Continous	Decimal
MMTRYB	Outcome of second walk	Categ	Binary
WSPEED2	Time for walking speed test 2(sec)	Continous	Decimal
WALKPAIN	Had pain during walking speed test	Categ	Ordinal
WALKFLR	Floor surface walking speed test	Categ	Nominal
WALKAID	Type aid used during s's walking speed test	Categ	Nominal
SMKEVR	Have you ever smoked	Categ	Binary
HEACTA	Freq vigorous activity level	Categ	Ordinal
HEACTB	Freq moderate activity level	Categ	Ordinal
HEACTC	Freq light activity level	Categ	Ordinal
ALCOH	Alcohol consumption in last 12 months	Categ	Ordinal

NURSE DATA

SEX	Gender of participants	Categ	Binary
DOB	Date of birth of participants	Continous	Integer
BPACT30	Did activity last 30 minutes that affects BP	Categ	Nominal
SYSTO1	Blood pressure measure (systolic) 1	Continous	Integer
DIASTO1	Blood pressure measure (diastolic) 1	Continous	Integer
PULSE1	Pulse measure 1	Continous	Integer
MAP1	Mean artherial pres 1	Continous	Integer
SYSTO2	Blood pressure measure (systolic) 2	Continous	Integer
DIASTO2	Blood pressure measure (diastolic) 2	Continous	Integer
PULSE2	Pulse measure 2	Continous	Integer
MAP2	Mean artherial pres 2	Continous	Integer
SYSTO3	Blood pressure measure (systolic) 3	Continous	Integer
DIASTO3	Blood pressure measure (diastolic) 3	Continous	Integer
PULSE3	Pulse measure 3	Continous	Integer
MAP3	Mean artherial pres 3	Continous	Integer
DOMGRIP	Dominant hand for gripping	Categ	Binary
GRIPBOTH	Whether respondent is able to use both	Categ	Nominal
DGRIP1	Dominant hand grip measurement 1(kg)	Continous	Integer
NGRIP1	Non-dominant hand grip measurement 1(kg)	Continous	Integer
DGRIP2	Dominant hand grip measurement 2(kg)	Continous	Integer
NGRIP2	Non-dominant hand grip measurement 2(kg)	Continous	Integer
DGRIP3	Dominant hand grip measurement 3(kg)	Continous	Integer
NGRIP3	Non-dominant hand grip measurement 3(kg)	Continous	Integer
GRIPPOS	Position for grip strength test	Categ	Nominal
CFIB	Blood fibrinogen level (g/l)	Continous	Integer
CHOL	Blood total cholesterol level (mmol/l)	Continous	Integer
HDL	Blood HDL level (mmol/l)	Continous	Integer
TRIG	Blood triglyceride level (mmol/l)	Continous	Integer
LDL	Blood LDL level (mmol/l)	Continous	Integer
FGLU	Blood glucose level (mmol/L) - fasting samples only	Continous	Integer
RTIN	Blood ferritin level (ng/ml)	Continous	Integer
HSCRP	Blood CRP level (mg/l)	Continous	Integer
HGB	Blood haemoglobin level (g/dl)	Continous	Integer
HBA1C	Blood glycated haemoglobin level (%)	Continous	Integer
MHEIGHT	Height measurement in meters	Continous	Integer
MWEIGHT	Weight measurement in kilograms	Continous	Integer
MBMI	Measured body mass index (kg/m2)	Continous	Integer
MWAIST	Average waist measurement in centimeters	Continous	Integer
MHIP	Average hip measurement in centimeters	Continous	Integer
HTFEV	Lung function: highest satisfactory fev reading	Continous	Integer
HTFVC	Lung function: highest satisfactory fvc reading	Continous	Integer
HTPF	Lung function: highest satisfactory pf reading	Continous	Integer

Table 4.6: Used Core and Nurse Variables

	Accuracy		**Youden's J**		**AUC**	
Mean/Mode	0.91		0.14		0.50	
	PMM	**CART**	**PMM**	**CART**	**PMM**	**CART**
MICE-MAR	0.90	0.90	0.05	0.04	0.50	0.50
MICE-MNAR	0.93	0.93	0.28	0.27	0.50	0.50

Table 4.7: Performance of imputation methods for LR predictions. Performance is evaluated with accuracy, Youden's index, and AUC (higher is better). Mean/Mode is a simple reference imputation method using mode for categorical and mean for continuous missing variables. For imputation with MICE, we considered MAR and MNAR missingness mechanisms, and PMM (default) and CART methods.

Selected Features	Number of Iterations for Disease&ADL	Total Selected Features	Variable Type
WALKRA			ADL-Core Binary
DRESSA			ADL-Core Binary
BATHA			ADL-Core Binary
EATA			ADL-Core Binary
BEDA			ADL-Core Binary
TOILTA			ADL-Core Binary
MAPA			ADL-Core Binary
MEALSA			ADL-Core Binary
SHOPA			ADL-Core Binary
PHONEA			ADL-Core Binary
MEDSA			ADL-Core Binary
HOUSEWKA			ADL-Core Binary
MONEYA			ADL-Core Binary
WALK100A			ADL-Core Binary
SITA			ADL-Core Binary
CHAIRA			ADL-Core Binary
CLIMSA			ADL-Core Binary
CLIM1A			ADL-Core Binary
STOOPA			ADL-Core Binary
LIFTA			ADL-Core Binary
DIMEA			ADL-Core Binary
ARMSA			ADL-Core Binary
PUSHA			ADL-Core Binary
ARTHRE			Disease-Core Binary
ALZHE			Core Binary
DEMENE			Core Binary
MEMRYE			Core Binary
FALLEQ			Core Binary
PAINFR			Core Binary
HEACTB			Core Cont.
PAINLV			Core Cont.
GRIPBOTH			Nurse
GRIPPOS	k=2 & 9	N=33	Nurse
HEACTC			Core Cont.
SHLT			Core Cont.
INDAGER			Core Cont.
MDACTX_E			Core Cont.
SYSTO1			Nurse
PULSE1			Nurse
SYSTO2			Nurse
SYSTO3	k=3 & 15	N=41	Nurse

Table 4.8: Selected features using the filter feature selection method, for $N = 33$ and $N = 41$ total features (reading from the top) for both disease and ADL predictions. The variable types are also indicated. The number of iterations (k) indicates how many variables were chosen for each output; N represents the total number of unique variables for all outputs.

	Diseases 1 Layer		ADLs 1 Layer	
	Damage	Repair	Damage	Repair
AUC	0.92		0.90	
R^2	0.99	0.98	0.94	0.89

	Full Features without Nurse			
AUC	0.92		0.90	
R^2	0.98	0.99	0.98	0.94

	2 Layers		2 Layers	
AUC	0.92		0.90	
R^2	0.95	0.95	0.90	0.42

	AutoKeras		AutoKeras	
AUC	0.65		0.50	
R^2	0.87	0.86	-0.85	0.82

Table 4.9: Model selection. Area under the ROC curve (AUC) is the evaluation metric for binary health state predictions, while R^2 is the metric for predicted and observed (real) values of average damage and repair probabilities. The full feature set of 134 features are used for the 1 layer, 2 layer, and AutoKeras models. In terms of AUC the 1 and 2 layer models are comparable, but the AutoKeras model does not perform as well. Negative R^2 shows a worse fit than the horizontal line defined by the mean of the data points. In terms of the R^2 the one layer model is best, and the performance is only slightly degraded without the nurse data.

AUC for Diseases	W23	W45	W67	W89
DNN	0.91	0.93	0.93	0.90
RF	0.89	0.91	0.90	0.86
AUC for ADLs				
DNN	0.90	0.90	0.91	0.90
RF	0.87	0.87	0.88	0.86

Table 4.10: Comparison of random forest (RF) and the best DNN model for different waves, as indicated. Performance is assessed by the average AUC of disease and ADL states. The full set of 134 features were used. We see that our DNN model reliably outperforms RF.

DISEASES	DNN	LR	Trivial
AUC	0.92	0.79	0.80

ADLs	DNN	LR	Trivial
AUC	0.90	0.69	0.72

Table 4.11: Comparison of the prediction performance of logistic regression (LR) and the best DNN model for 19 disease states and 25 ADLs. DNN is better than LR for both diseases and ADLs. The trivial model assumes that the target variables do not change from the previous wave. We used the full 134 features.

DISEASES	DNN	LR	Trivial
AUC	0.83	0.60	0.58

ADLs	DNN	LR	Trivial
AUC	0.89	0.66	0.76

Table 4.12: Comparison of the prediction performance of logistic regression (LR) and the best DNN model for predicting transitions of 19 disease states and 25 ADLs. DNN is better than LR for both diseases and ADLs. The trivial model assumes that the target variables all change to their opposite values from the previous wave values. We used the full 134 features.

Brier Score	**DISEASES**	**ADLs**
Wave 2 to 3	0.020	0.032
Wave 4 to 5	0.014	0.034
Wave 6 to 7	0.012	0.034
Wave 8 to 9	0.016	0.032

Table 4.13: Calibration scores: Brier score takes values between 0 and 1, which represent the complete calibration and complete non-calibration. We average Brier scores over all diseases and ADLs and have well calibration scores for all diseases and ADLs for all wave transitions. See Tables 4.14 and 4.15 for individual disease and ADL Brier scores.

Brier	**Wave 2 to 3**	**Wave 4 to 5**	**Wave 6 to 7**	**Wave 8 to 9**
HIBPE	0.029	0.026	0.026	0.040
DIABE	0.037	0.026	0.021	0.025
CANCRE	0.026	0.022	0.020	0.022
LUNGE	0.019	0.015	0.016	0.016
HEARTE	0.006	0.005	0.007	0.007
STROKE	0.016	0.018	0.022	0.025
PSYCHE	0.004	0.003	0.006	0.002
ARTHRE	0.020	0.021	0.017	0.022
ASTHMAE	0.013	0.015	0.016	0.023
HCHOLE	0.046	0.034	0.039	0.036
CATRACTE	0.016	0.010	0.009	0.008
PARKINE	0.018	0.015	0.019	0.019
HIPE	0.012	0.005	0.011	0.010
ANGINE	0.008	0.015	0.016	0.017
HRTATTE	0.004	0.005	0.008	0.009
CONHRTFE	0.022	0.021	0.023	0.021
HRTMRE	0.016	0.017	0.015	0.027
HRTRHME	0.039	0.031	0.028	0.029
OSTEOE	0.021	0.014	0.016	0.012

Table 4.14: For individual disease calibration scores: All individual diseases over all waves show a well calibration since they all have Brier scores close to 0.

Brier	Wave 2 to 3	Wave 4 to 5	Wave 6 to 7	Wave 8 to 9
WALKRA	0.033	0.032	0.032	0.034
DRESSA	0.023	0.027	0.023	0.026
BATHA	0.021	0.030	0.025	0.029
EATA	0.064	0.067	0.065	0.065
BEDA	0.060	0.062	0.056	0.066
TOILTA	0.012	0.014	0.015	0.013
MAPA	0.008	0.008	0.009	0.009
DANGERA	0.008	0.011	0.010	0.011
MEALSA	0.040	0.037	0.038	0.037
SHOPA	0.035	0.036	0.039	0.036
PHONEA	0.034	0.032	0.037	0.032
COMMUNA	0.022	0.025	0.030	0.022
MEDSA	0.021	0.027	0.031	0.026
HOUSEWKA	0.064	0.062	0.064	0.057
MONEYA	0.056	0.063	0.061	0.055
WALK100A	0.016	0.013	0.012	0.013
SITA	0.009	0.008	0.010	0.008
CHAIRA	0.009	0.010	0.010	0.008
CLIMSA	0.037	0.033	0.038	0.034
CLIM1A	0.038	0.033	0.038	0.034
STOOPA	0.027	0.037	0.030	0.027
LIFTA	0.024	0.028	0.024	0.022
DIMEA	0.021	0.031	0.029	0.027
ARMSA	0.064	0.068	0.061	0.061
PUSHA	0.056	0.060	0.059	0.056

Table 4.15: For individual ADL calibration scores: All individual ADLs over all waves show a well calibration since they all have Brier scores close to 0.

	HR of High-Low Risk Groups	
	Observed	**Predicted**
HIBPE	1.74	1.72
DIABE	1.31	1.38
ANGINE	1.94	1.87

Table 4.16: Observed and predicted HRs of high versus low risk groups for the bimodal distributions in Fig. 4.12.

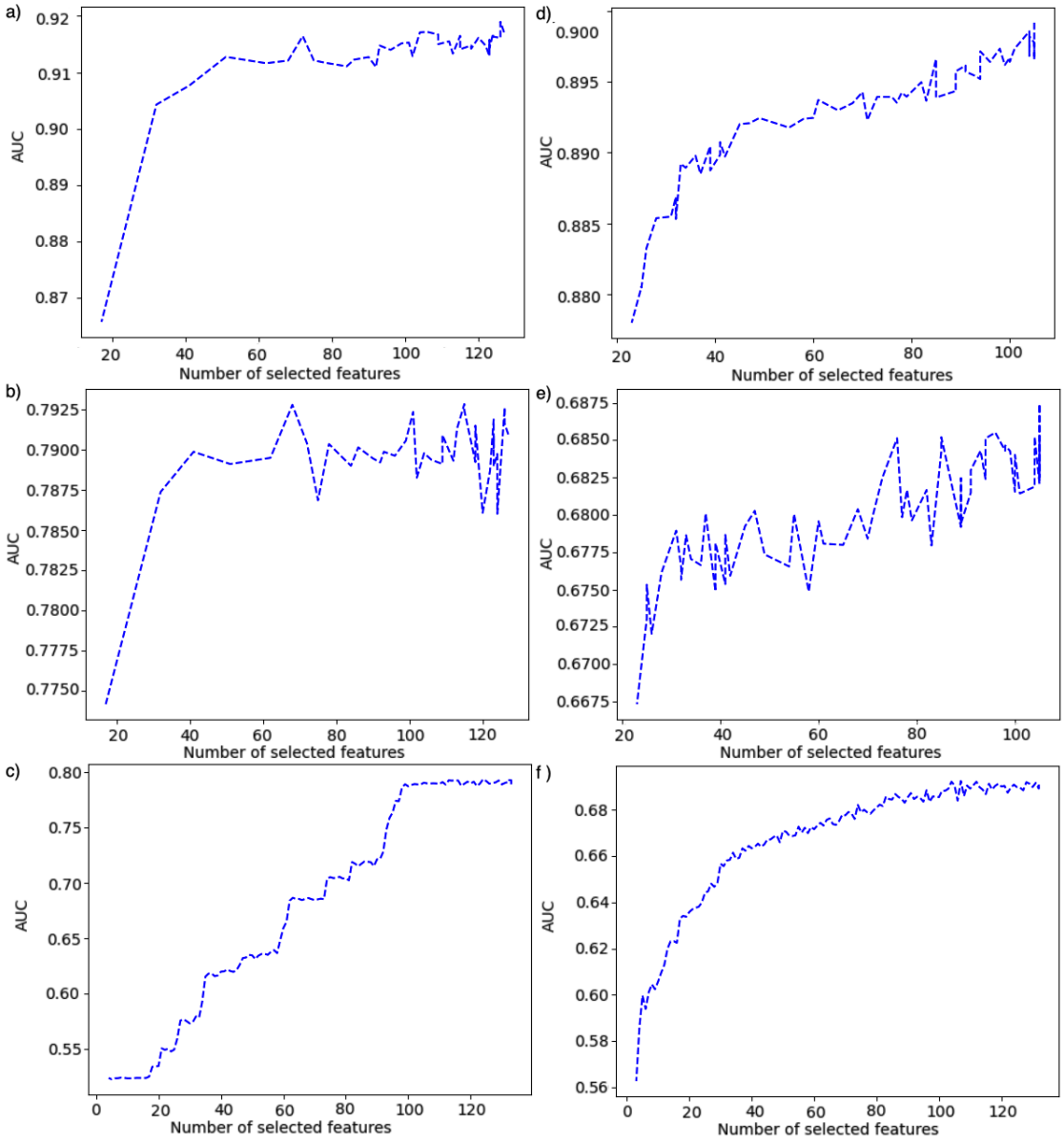

Figure 4.4: Prediction performance in terms of AUC vs #n features for diseases a) in filter DNN, b) in filter LR, c) in RFE LR methods and for ADLs d) in filter DNN, e) in filter LR, and f) in RFE LR methods. While DNN is the best model, filter method is the best feature selection method with the earliest AUC saturation for both diseases and ADLS in a) and d), respectively.

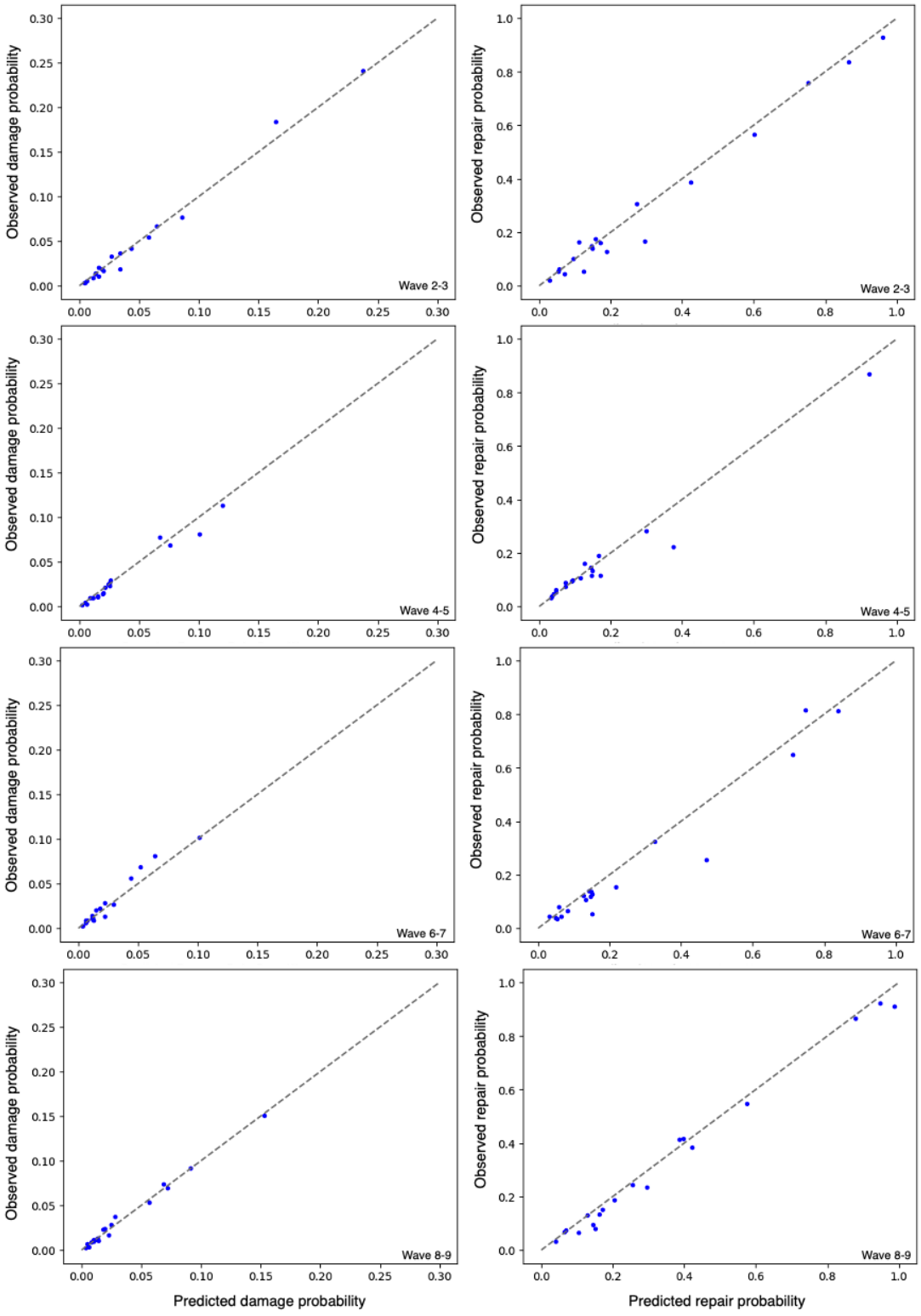

Figure 4.5: Predicted versus observed average damage and repair probabilities for Diseases. Each point represents one disease. Waves are as indicated. The average R^2 across waves are in Table 4.3 Diseases. The gray dashed line indicates perfect calibration.

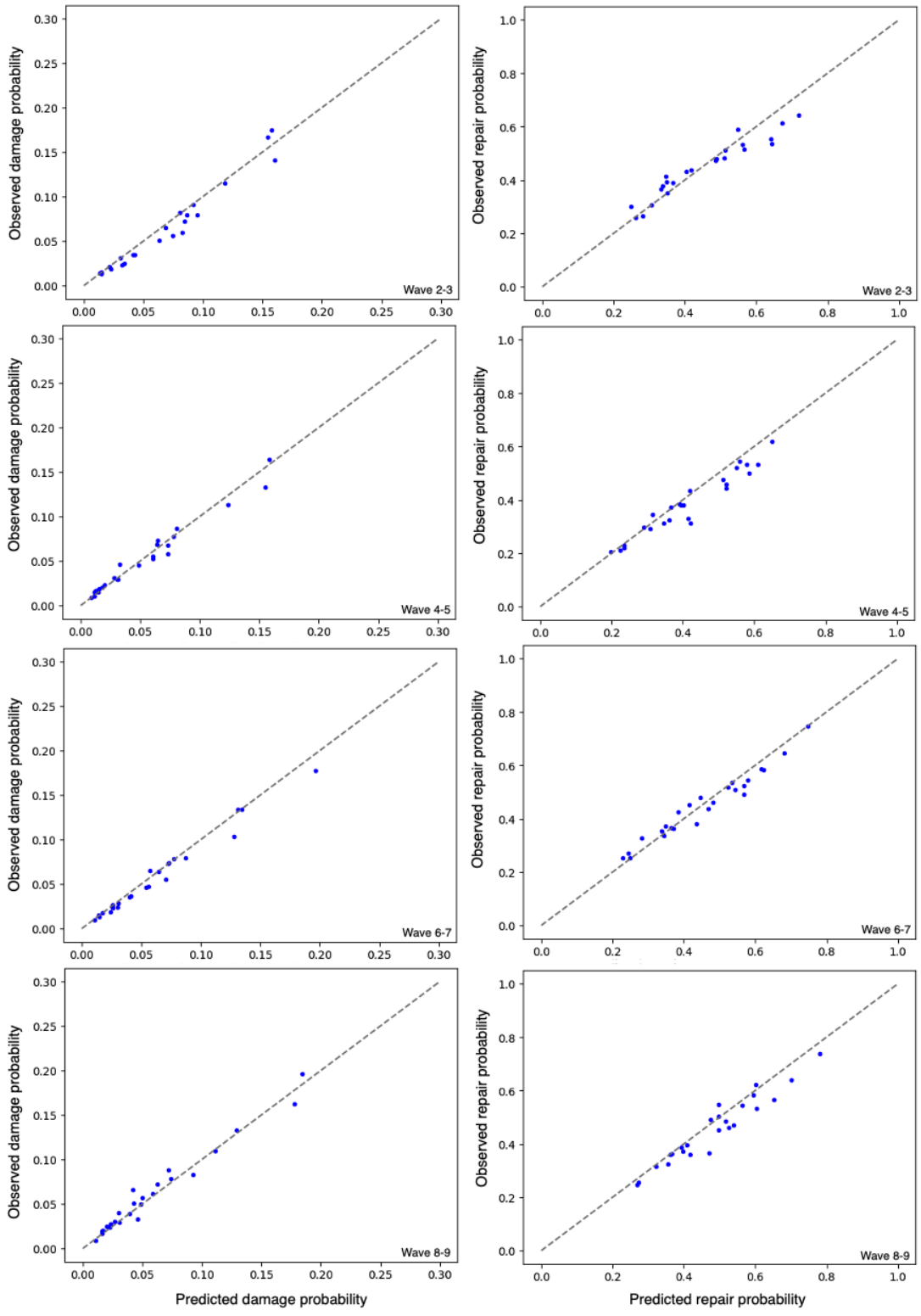

Figure 4.6: Predicted versus observed average damage and repair probabilities for ADLs. Each point represents one ADL. Waves are as indicated. The average R^2 across waves are in Table 4.3 ADLs. The gray dashed line indicates perfect calibration.

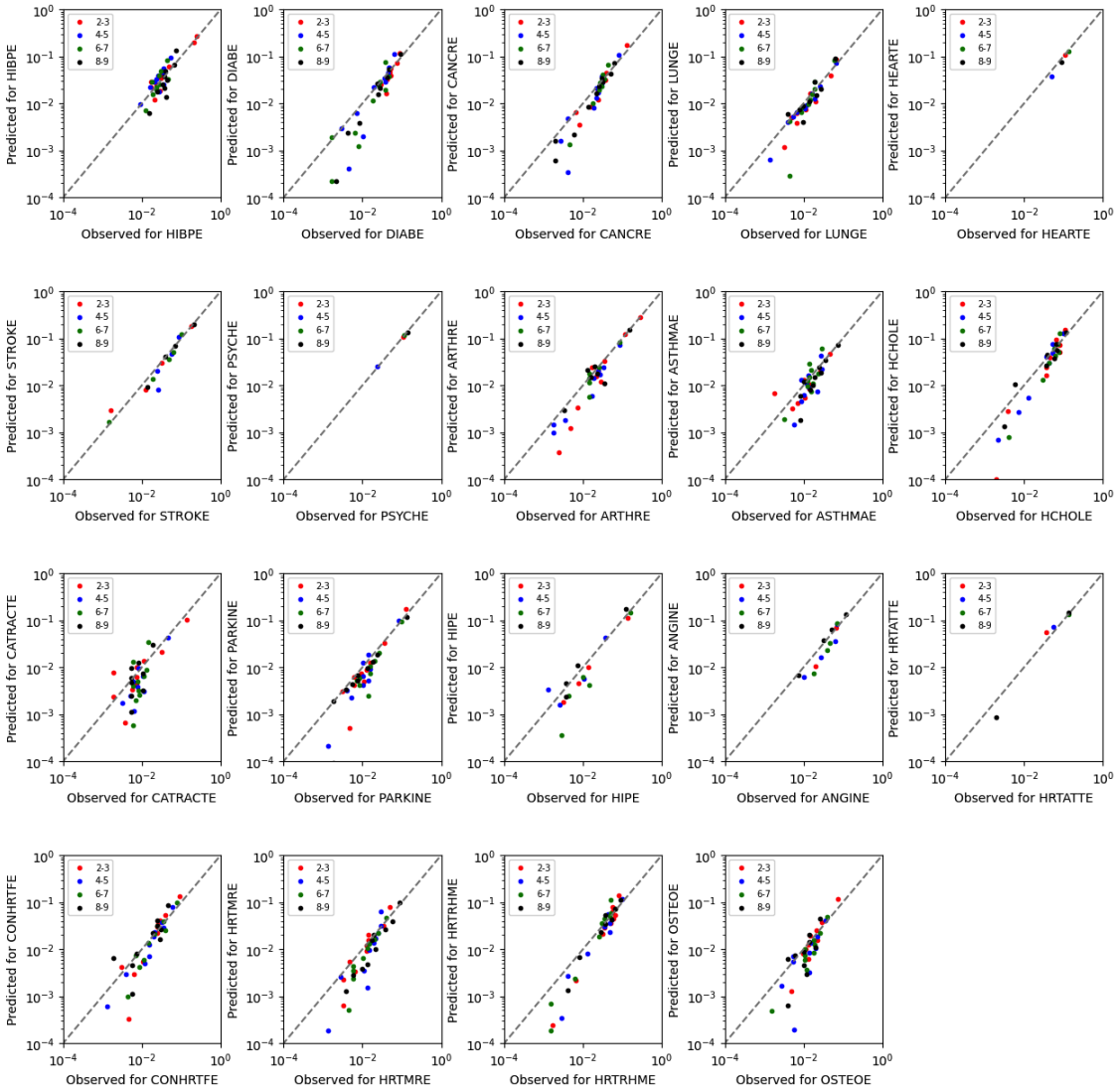

Figure 4.7: Individual disease calibration curves for damage transition probabilities. Each point is the average of one decile, for the indicated disease, and for the waves indicated by the legend. The gray dashed line indicates perfect calibration.

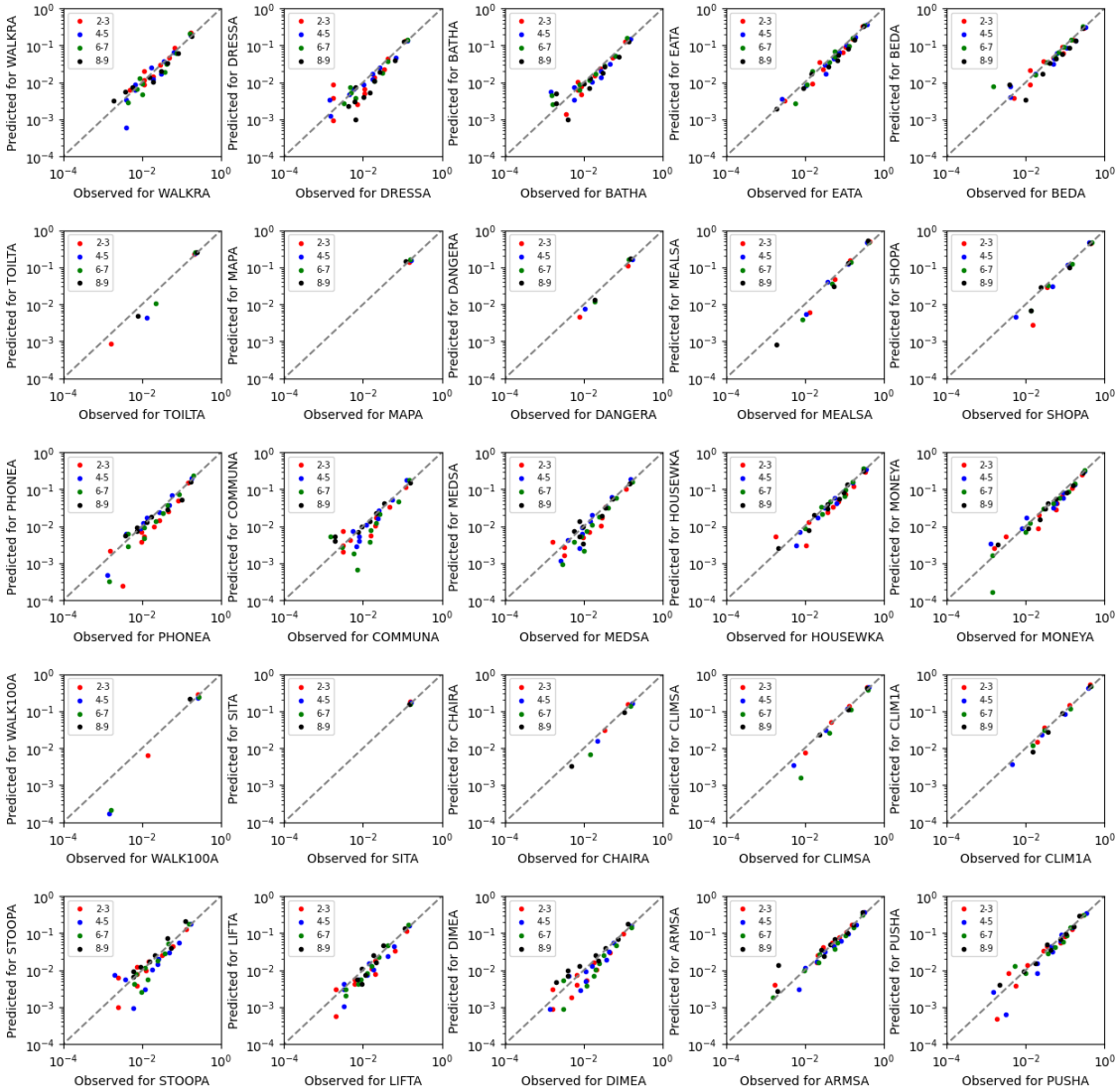

Figure 4.8: Individual ADL calibration curves for damage transition probabilities. Each point is the average of one decile, for the indicated ADL, and for the waves indicated by the legend. The gray dashed line indicates perfect calibration.

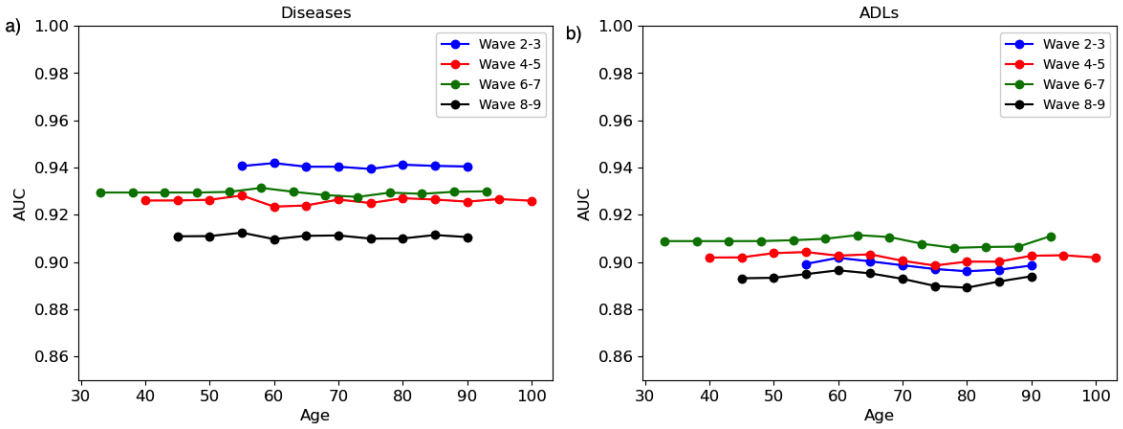

Figure 4.9: Prediction performance vs age. We show the average AUC of a) disease predictions or b) ADLs for different ages (in five year bins) as indicated, and for different waves as indicated in the legend. Prediction performance does not strongly depend on age.

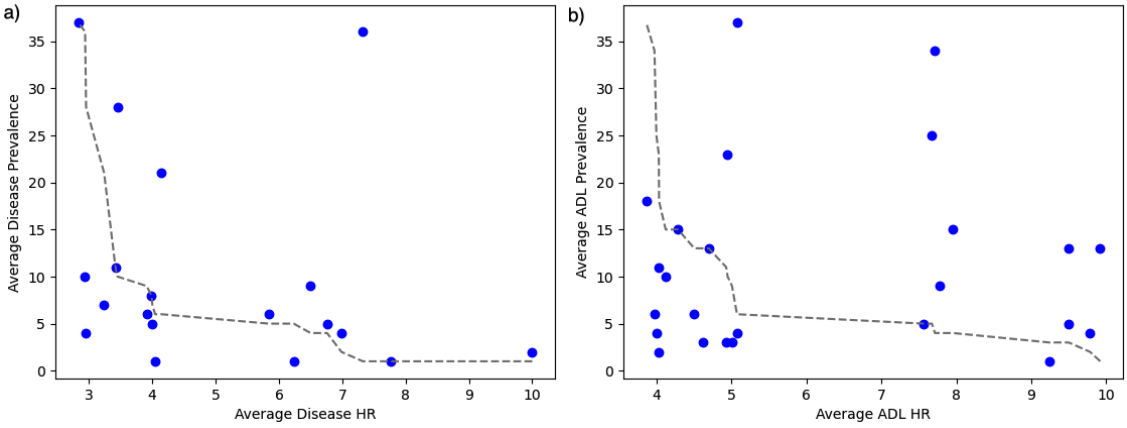

Figure 4.10: Hazard ratio versus prevalence of a) diseases and b) ADLs. The values are averaged over all wave transitions. The gray dashed curve shows the perfect inverse Spearman's rank correlation (=1.00) between prevalence and HR values. The measured Spearman's are -0.43 for diseases and 0.06 for ADLs.

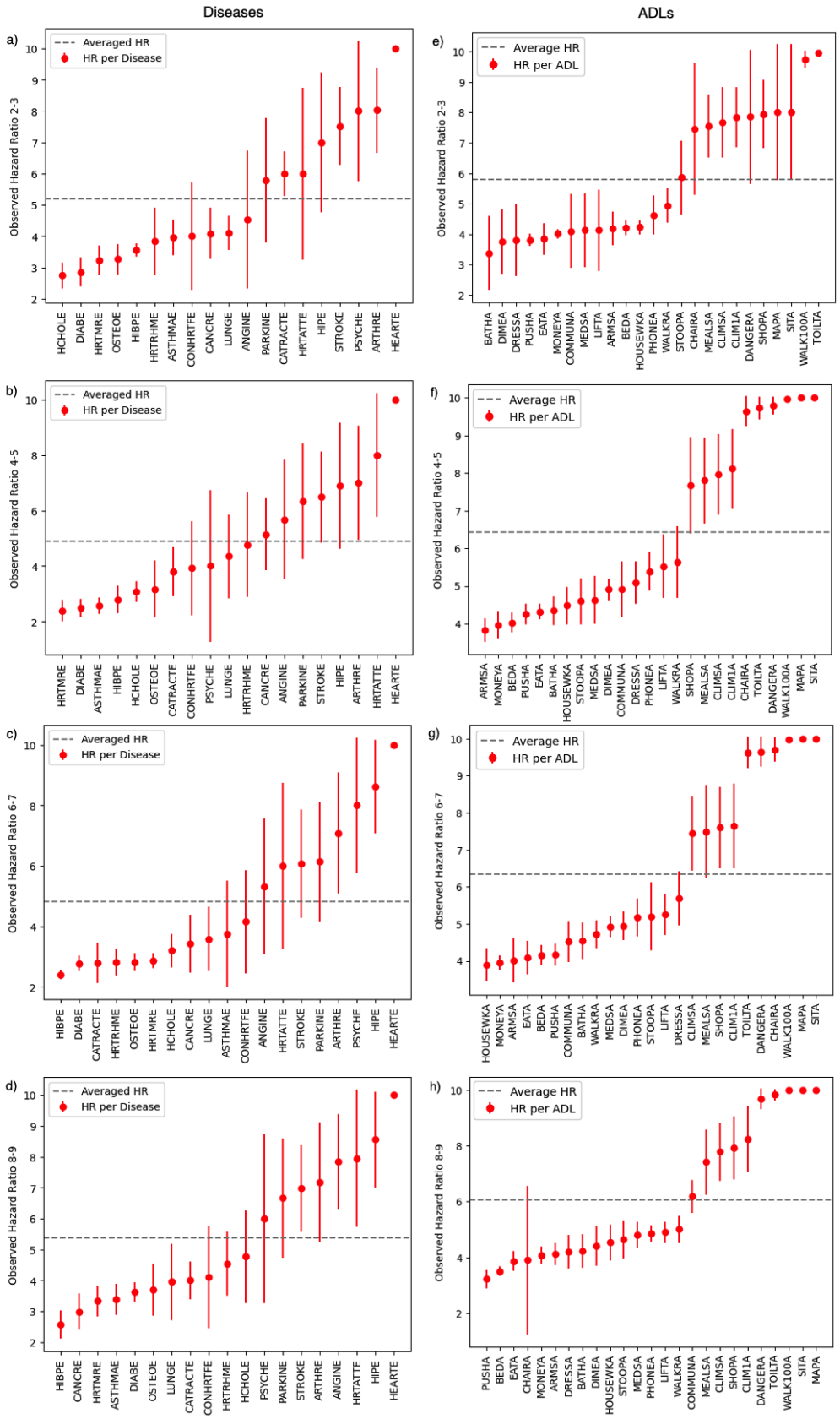

Figure 4.11: HR versus all diseases (left, a-d) and ADLs (right, e-h) for all wave transitions as indicated. HRs are rank ordered for each wave transition. The horizontal dashed line indicates the average HR over all diseases or ADLs, respectively.

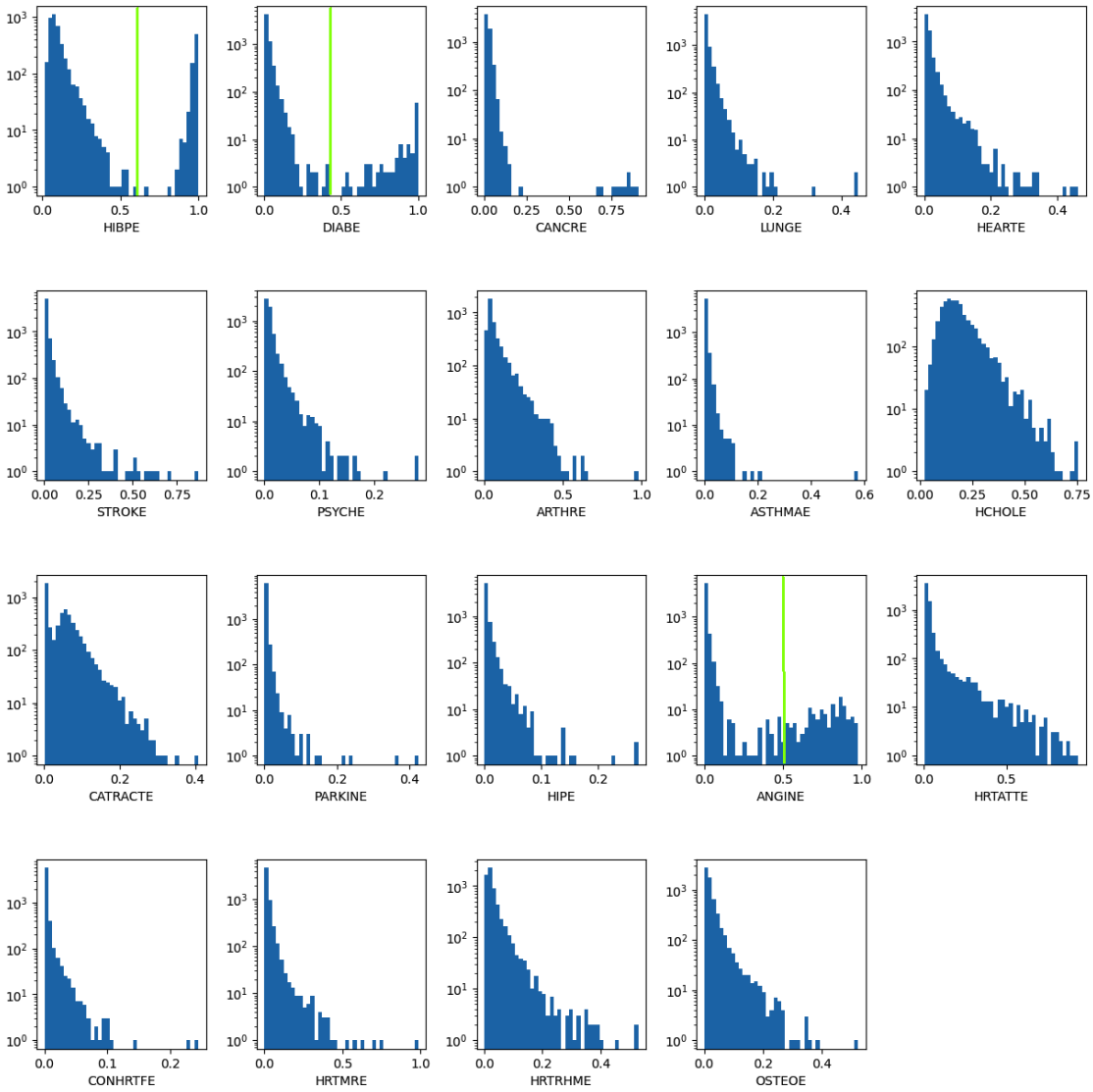

Figure 4.12: Histograms of predicted disease damage transition probabilities in Waves 2-3. Note the log-scale. All distributions are right-skewed from the mode (peak). The tail appears exponential (straight in this semilog plot) in some distributions (HIBPE, HCHOLE, and CATRACTE) but more generally power law (see Table 4.13). Some distributions appear bimodal (HIBPE, DIABE, and ANGINE) with vertical green lines indicating the cutpoints used in Table 4.16.

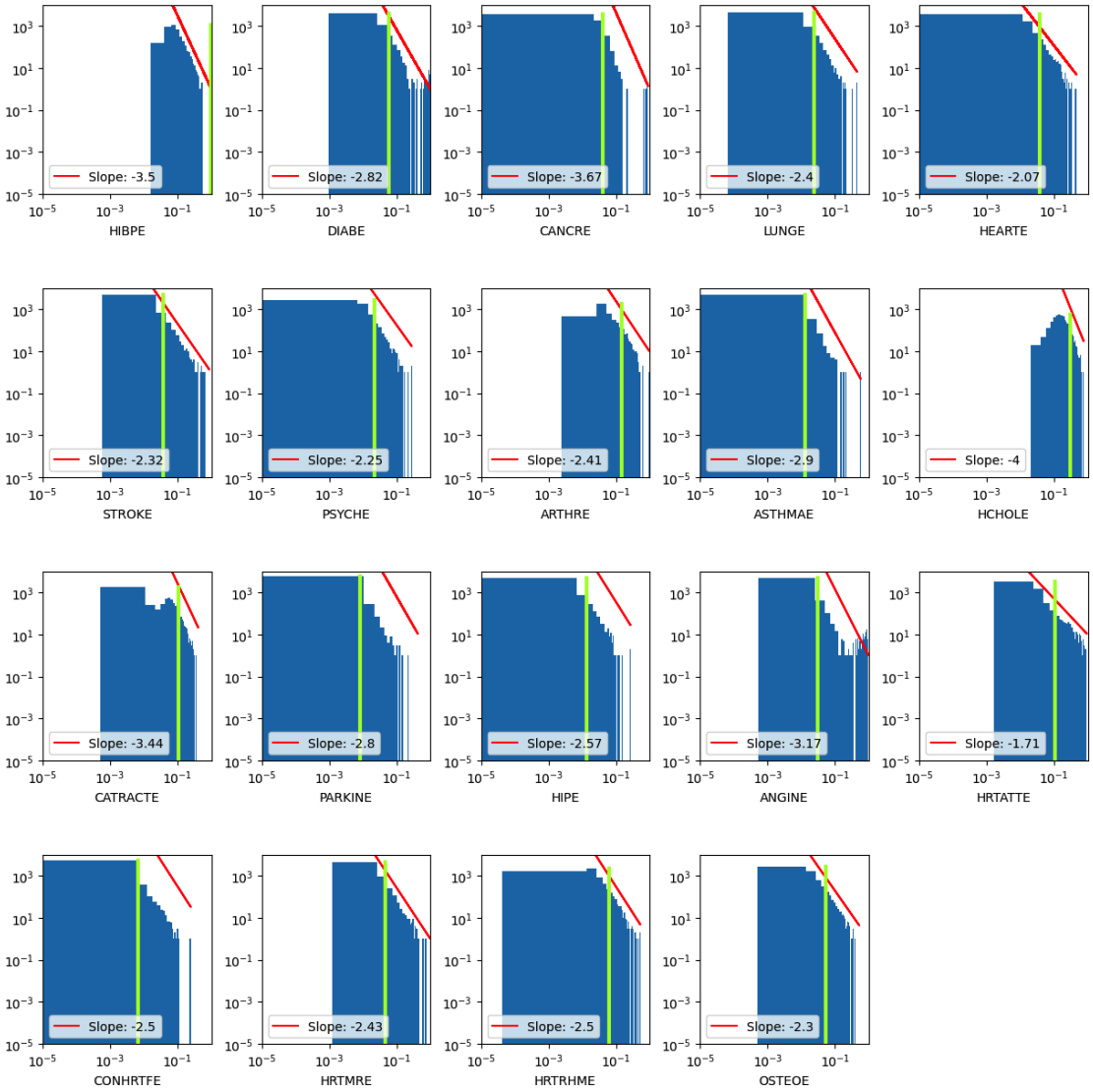

Figure 4.13: Histograms of predicted disease damage transition probabilities in Waves 2-3. Note the log scale for both axes. Green vertical lines indicate the highest decile of the transition probabilities. The red lines indicate power-law fits by eye, with indicated exponents.

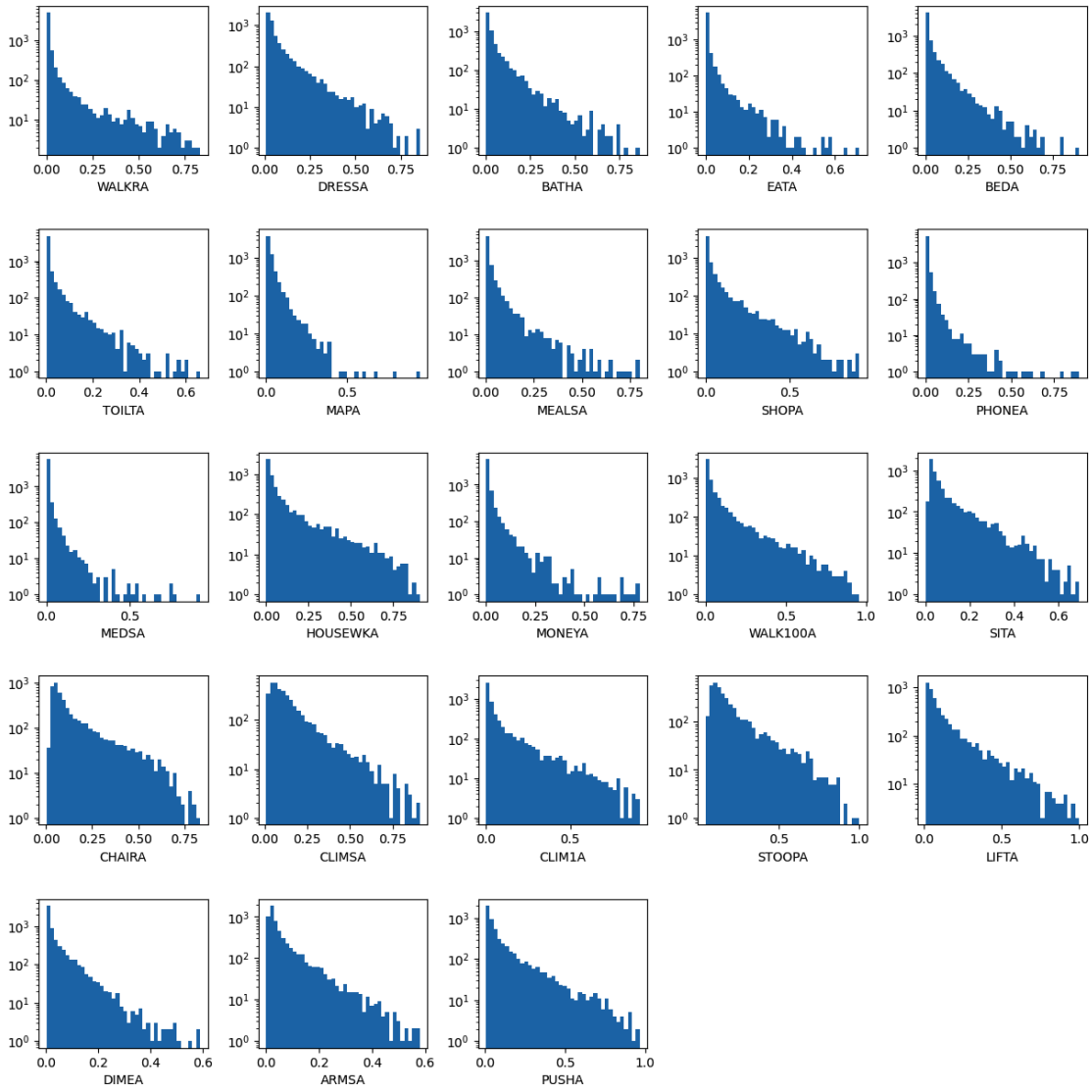

Figure 4.14: Histograms of predicted ADL damage transition probabilities in Waves 2-3. Note the log-scale. All distributions are right-skewed from the mode (peak). No distributions appear bimodal. All tails appear concave up (i.e. non-exponential).

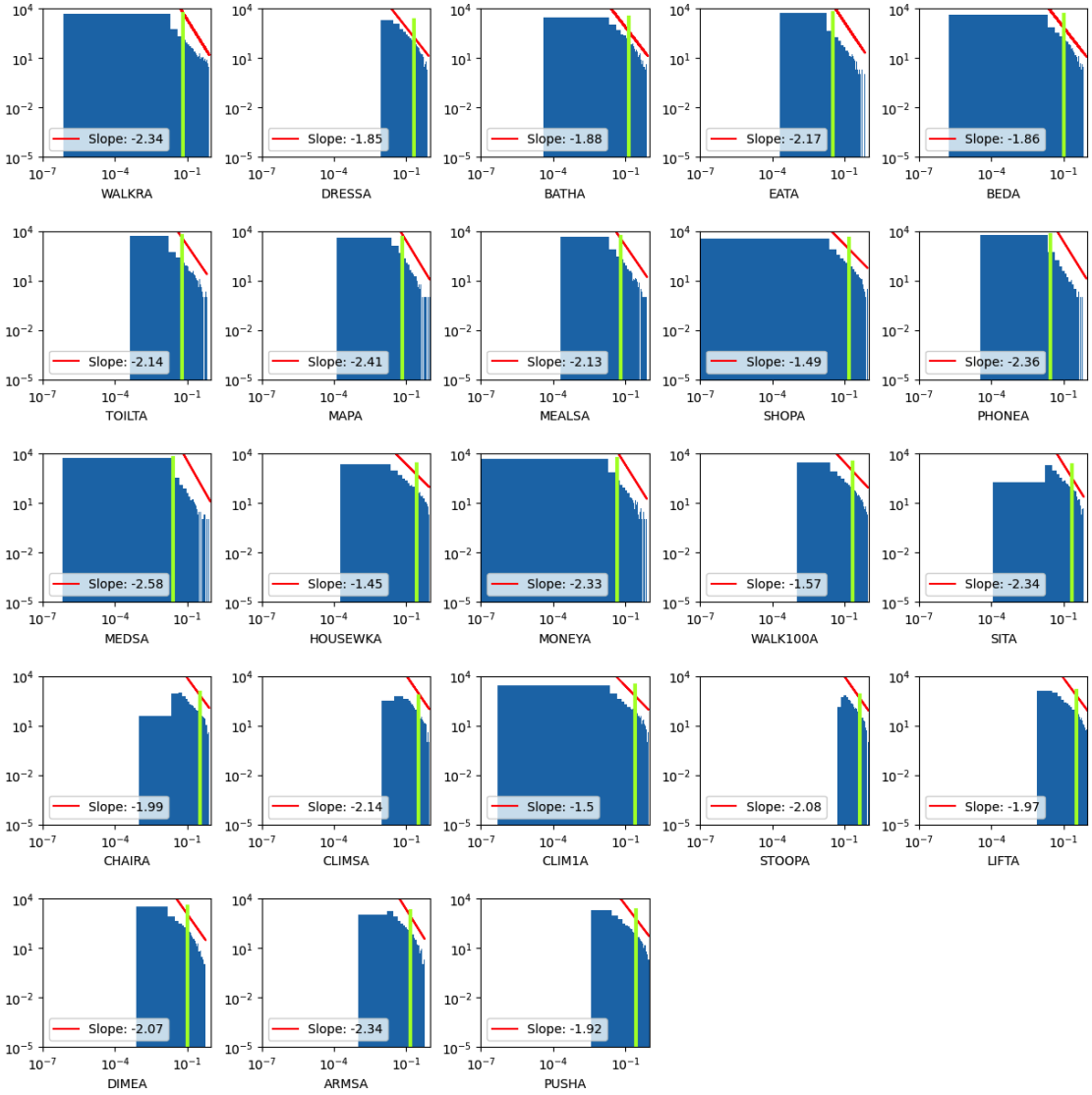

Figure 4.15: Histograms of predicted ADL damage transition probabilities in Waves 2-3. Note the log scale for both axes. Green vertical lines indicate the highest decile of the transition probabilities. The red lines indicate power-law fits by eye, with indicated exponents.

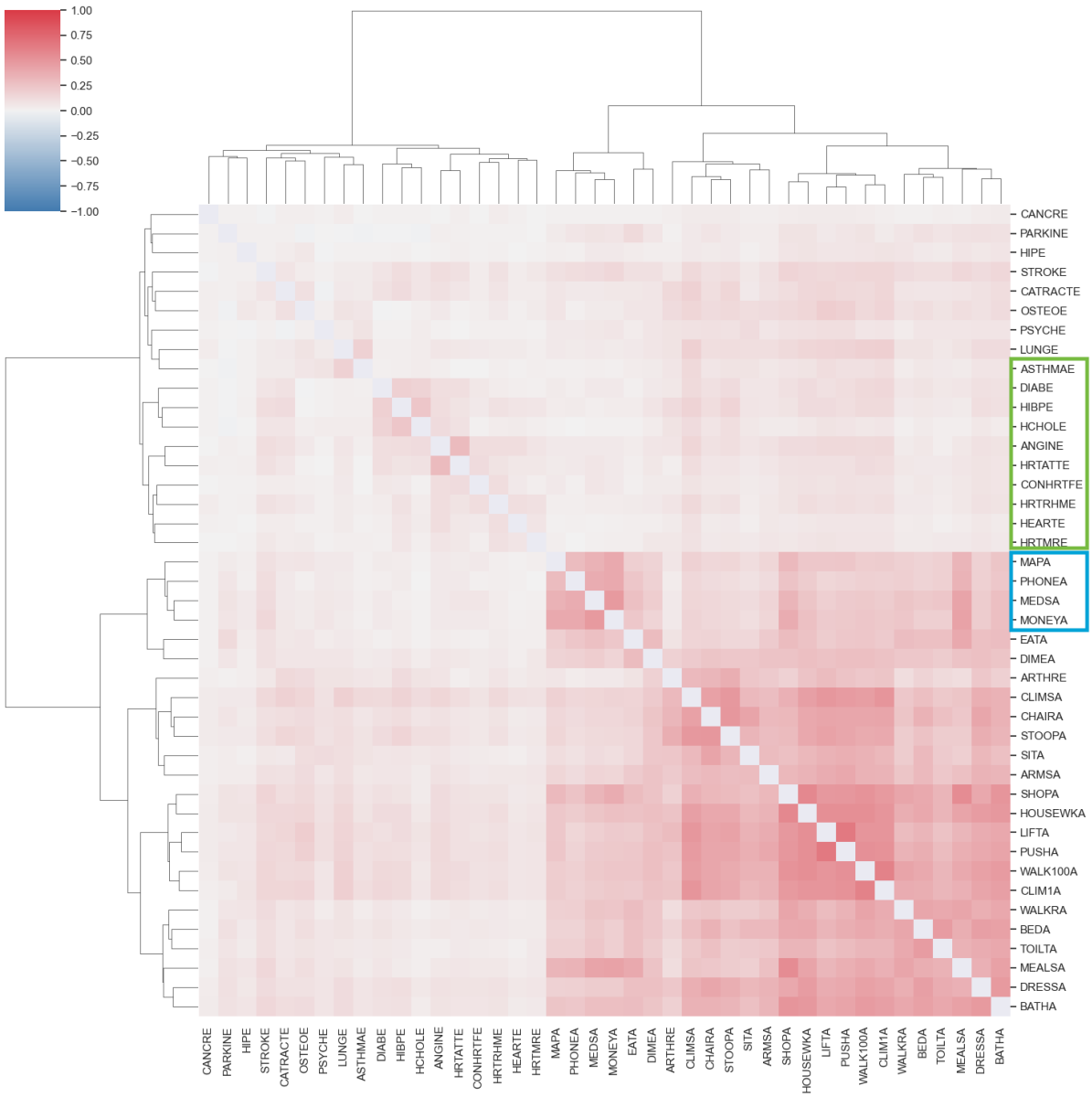

Figure 4.16: Correlations between observed health states. The hierarchical clustering is also indicated. Note that there is a strong correlation between ADL variables, and a moderate correlation between ADL and disease variables. There is a smaller correlation between disease variables. All of the stronger correlations are positive. We do not show diagonal correlations (i.e. variances).

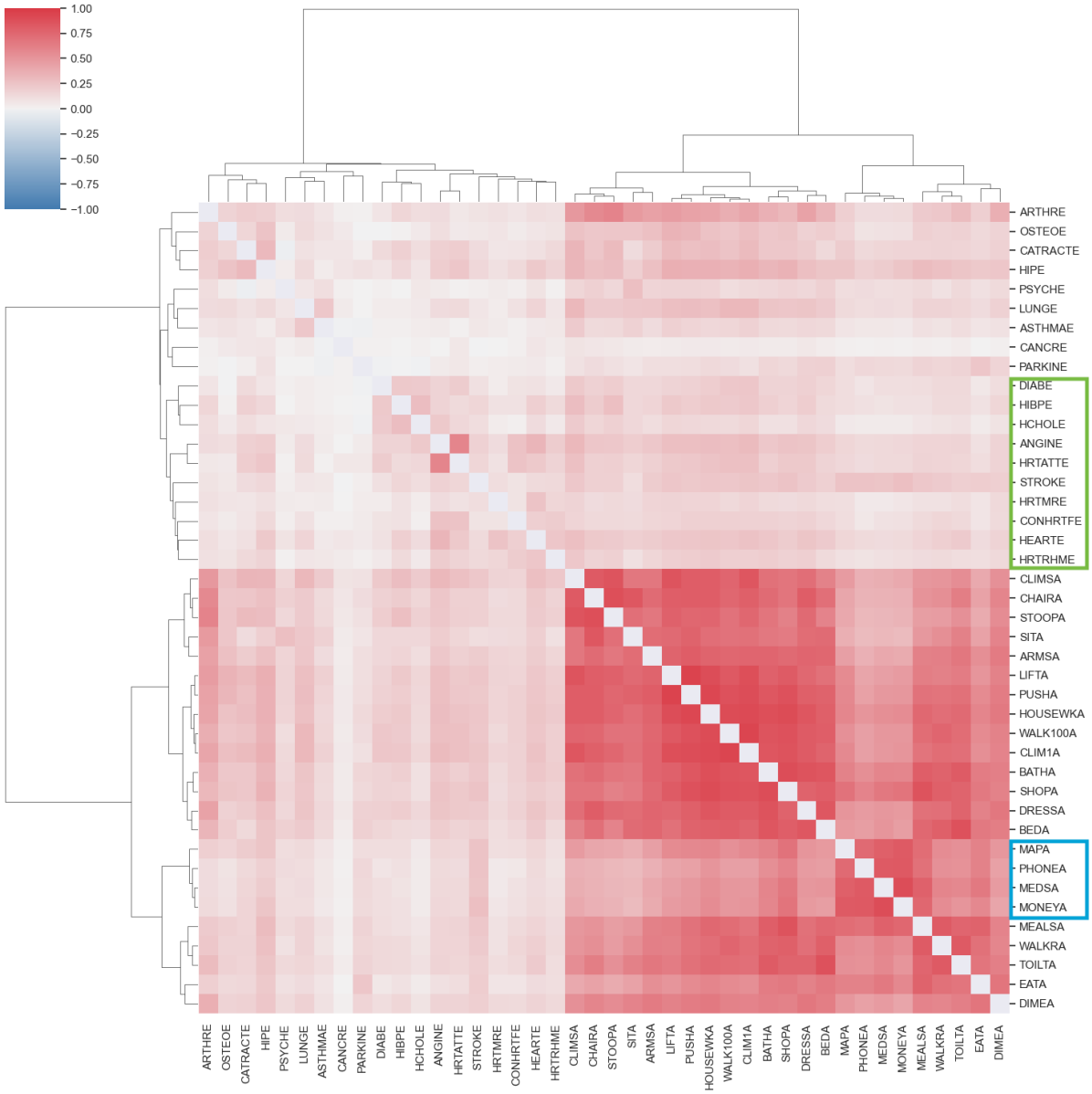

Figure 4.17: Correlations between predicted health states. The heirarchical clustering and block structure is similar to the observed case in Fig. 4.16, however the correlations are significantly stronger. Green and blue boxes are additionally put here different than the original paper, and show almost same disease and ADLs in observed and predicted cases in Tables 4.16 and 4.17, respectively.

4.4 Perspective

So far, we developed a predictive model to obtain the binary health states such as of 19 diseases and 25 ADLs between successive waves of the ELSA dataset. Then, we obtained predictive average damage and repair transition probabilities, and individual damage transition probabilities for those binary health states.

Although we obtained a very good predictive model for the binary health states and damage and repair probabilities, there are some open questions in our study. We could not obtain well calibrated correlations between observed and predicted health states, and damage transition probabilities of those binary health states, which will be discussed in next section. We are not sure if the underlying reason is the lack of a good predictive model for states or the necessity of a separate train-test process by the correlation coefficients. In addition, our model is not suitable for transfer learning by which a model trained for a previous wave transition can predict more later waves well. The lack of transfer learning may result from the variability of individuals in each wave which leads to a quasi-longitudinal data with a small test set around 2000 individuals.

As future studies, our model can be applied on different longitudinal datasets. One can try to use same steps done here and predict same variables to see if our model can generalize and give same good prediction qualities. Also, one can investigate the shortcomings of our model's on the new datasets. Poor-calibration of correlations and lack of transfer learning for our model may be overcome on different longitudinal datasets.

In the next chapter, we will give the results of some additional analysis on the shortcoming and open questions for our study. Those are poorly calibrated individual repair transition probabilities, random behaviour of male-female hazard ratios over different health states, poorly calibrated correlations between observed and predicted health states and between observed and predicted transition probabilities, and lastly the relation between HRs, prevalence and exponents for transition probability distributions.

Chapter 5

Additional Results and Analysis

5.1 Repair Transition Probabilities

We obtained repair transition probabilities for all diseases and ADLs in all wave transitions. We analyzed the calibration of repair probabilities in Tables 5.1 and 5.2 with individual disease and ADL Brier scores, and Figs. 5.1 and 5.2 with individual calibration curves for diseases and for ADLs, respectively. Although we see see similar Brier scores for damage and repair transition probabilities, results of repair cases were not as well-calibrated as the damage transition probabilities since very small number of points in scatter plots. Even some of the disease repairs, e.g. HIPE, CONHRTFE have no repair points in the deciles. These would be general reasons for this. First, the repair cases did not occur as many-times as in damage occurrence. This led us to work with a small training set. Second, repair may sometimes occur due to measurement error (e.g. disappearing symptoms that do not bother the patients, or recording error).

Brier	Wave 2 to 3	Wave 4 to 5	Wave 6 to 7	Wave 8 to 9
HIBPE	0.017	0.012	0.014	0.005
DIABE	0.015	0.018	0.012	0.027
CANCRE	0.004	0.006	0.016	0.012
LUNGE	0.020	0.012	0.008	0.017
HEARTE	0.013	0.001	0.010	0.048
STROKE	0.058	0.024	0.024	0.040
PSYCHE	0.007	0.008	0.003	0.011
ARTHRE	0.018	0.011	0.015	0.018
ASTHMAE	0.009	0.006	0.008	0.013
HCHOLE	0.014	0.017	0.016	0.022
CATRACTE	0.006	0.007	0.003	0.003
PARKINE	0.002	0.050	0.001	0.001
HIPE	0.000	0.001	0.000	0.001
ANGINE	0.020	0.054	0.037	0.026
HRTATTE	0.025	0.010	0.009	0.012
CONHRTFE	0.005	0.000	0.002	0.004
HRTMRE	0.014	0.010	0.005	0.007
HRTRHME	0.014	0.008	0.020	0.015
OSTEOE	0.030	0.020	0.020	0.024

Table 5.1: Individual disease calibration scores of repair transition probabilities. Brier score takes values between 0 and 1, which represent the complete calibration and complete non-calibration. Although all individual ADLs over all waves have Brier scores close to 0, this is not a well calibration many of them have no points in the scatter plots.

Brier	Wave 2 to 3	Wave 4 to 5	Wave 6 to 7	Wave 8 to 9
WALKRA	0.030	0.041	0.035	0.047
DRESSA	0.008	0.012	0.006	0.009
BATHA	0.016	0.014	0.018	0.023
EATA	0.073	0.064	0.063	0.061
BEDA	0.036	0.056	0.049	0.055
TOILTA	0.037	0.019	0.036	0.049
MAPA	0.032	0.031	0.019	0.035
DANGERA		0.012	0.026	0.025
MEALSA	0.053	0.065	0.058	0.052
SHOPA	0.061	0.065	0.065	0.070
PHONEA	0.018	0.031	0.035	0.016
COMMUNA		0.006	0.020	0.014
MEDSA	0.033	0.019	0.014	0.053
HOUSEWKA	0.047	0.030	0.046	0.040
MONEYA	0.038	0.046	0.058	0.043
WALK100A	0.044	0.032	0.035	0.044
SITA	0.028	0.026	0.026	0.034
CHAIRA	0.024	0.019	0.020	0.039
CLIMSA	0.073	0.059	0.058	0.071
CLIM1A	0.060	0.049	0.054	0.062
STOOPA	0.021	0.014	0.013	0.014
LIFTA	0.018	0.019	0.015	0.015
DIMEA	0.015	0.018	0.006	0.012
ARMSA	0.051	0.044	0.038	0.046
PUSHA	0.065	0.058	0.058	0.067

Table 5.2: Individual ADL calibration scores of repair transition probabilities. Brier score takes values between 0 and 1, which represent the complete calibration and complete non-calibration. Individual ADLs over all waves show a better calibration than disease case since there are not zero-point scatter plots.

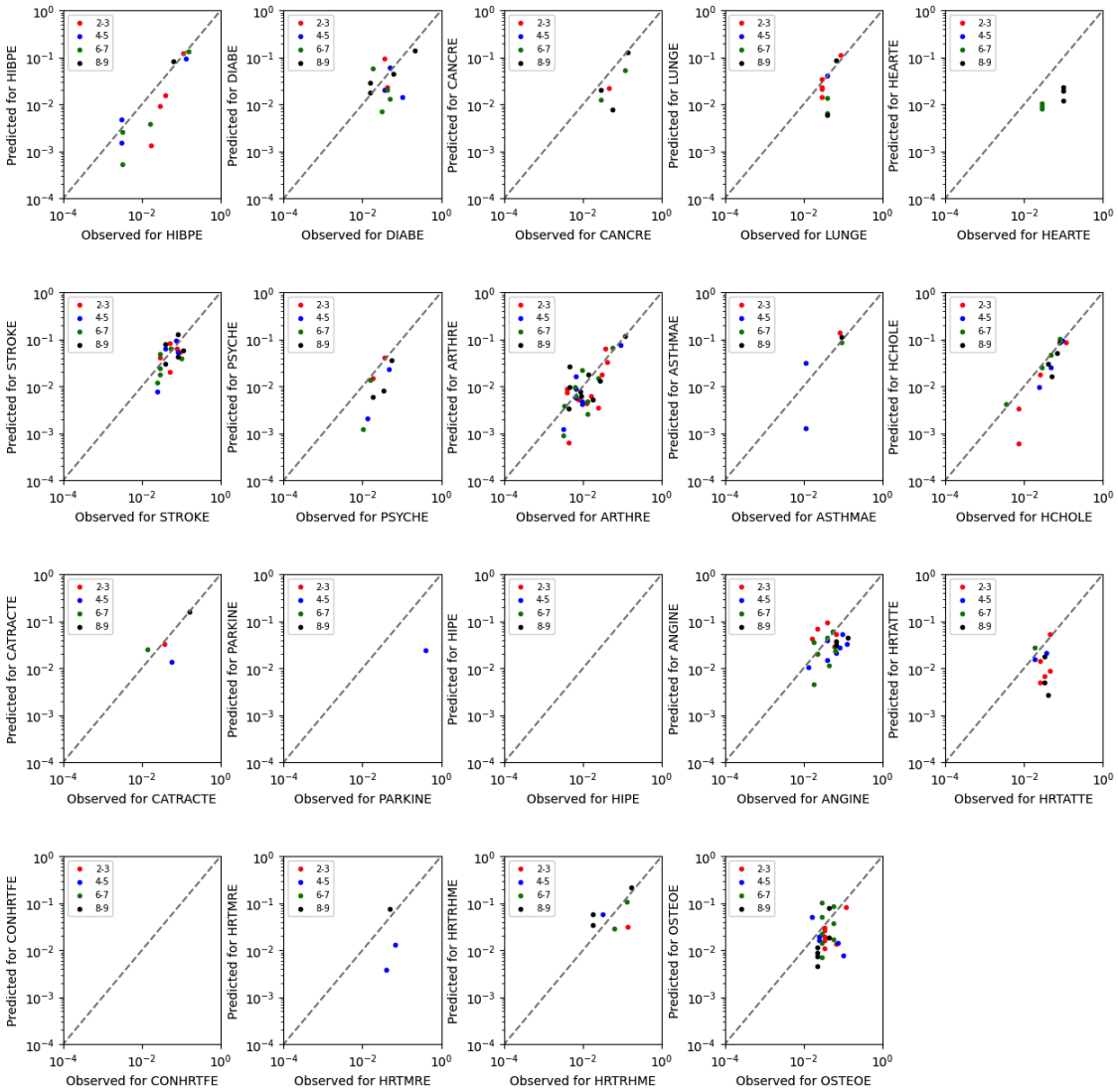

Figure 5.1: Individual disease calibration curves for repair transition probabilities. Each point is the average of one decile, for the indicated disease, and for the waves indicated by the legend. The gray dashed line indicates perfect calibration. Some diseases have very few or no points in the scatter plots such as CATRACTE, PARKINE, HIPE and CONHRTFE because there are not enough or any repaired individuals in the all deciles.

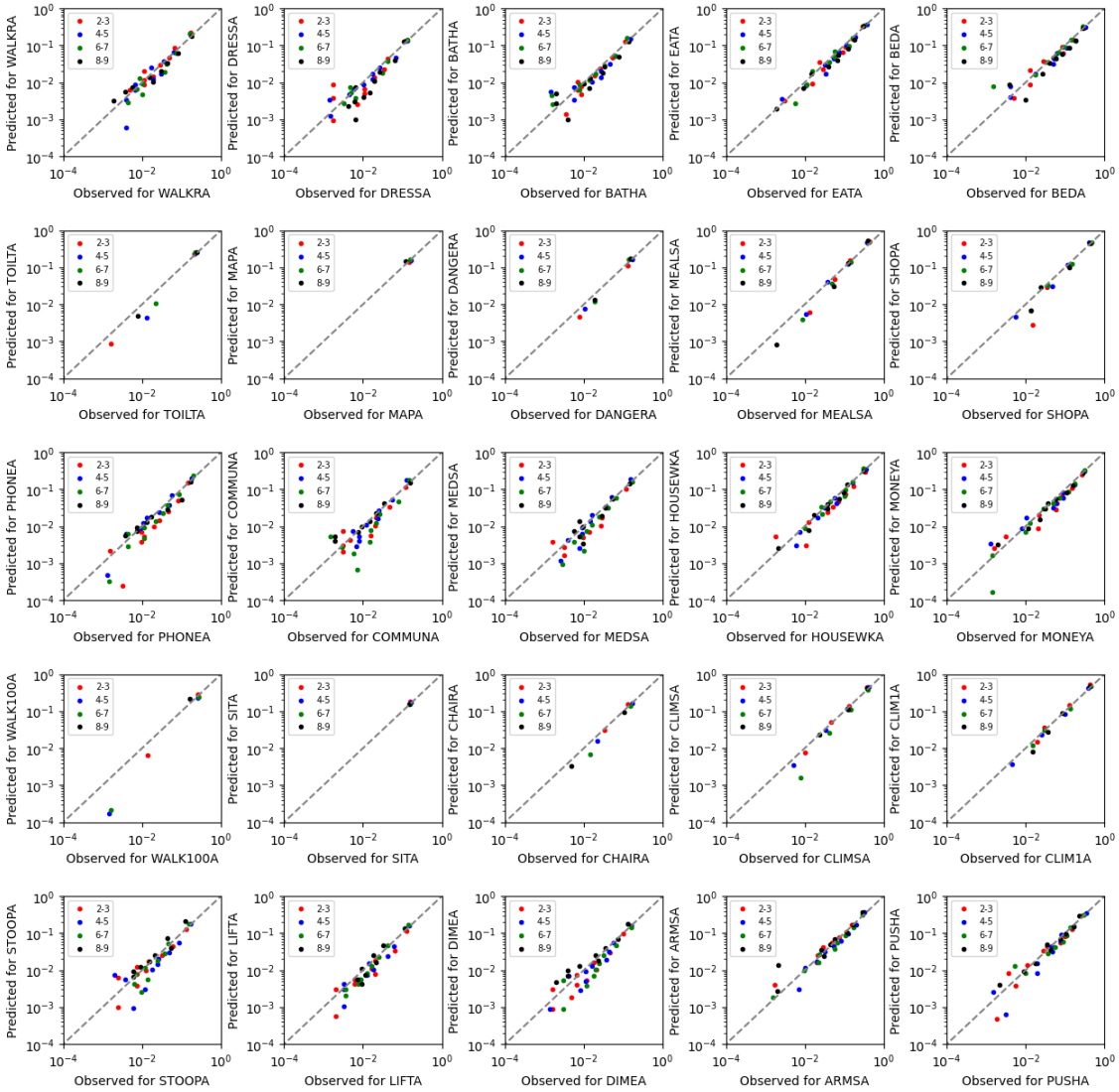

Figure 5.2: Individual ADL calibration curves for repair transition probabilities. Each point is the average of one decile, for the indicated ADL, and for the waves indicated by the legend. The gray dashed line indicates perfect calibration.

5.2 Male-Female Hazard Ratios

We repeated our HR calculations for male and female individuals separately. However, we could not get a consistent picture for male and female HRs over the diseases and ADLs in Fig. 5.3 showing the averaged HRs over all wave transitions and in Fig. 5.4 for HRs over individual wave transitions. While males have higher HRs for 10 diseases, females have 8 diseases; and while males have 14 higher HRs for ADLs, females have 9 ADLs in Fig 5.3. Moreover, we can say that a gender has sometimes higher HR for some diseases or ADLs in all waves in Fig. 5.4. For diseases, while the males have higher HR for the PARKINE, females have higher for the CONHRTFE (see Supplementary Table S1 for acronyms). For ADLs, while the males have higher HRs for DRESSA, EATA, BEDA, CHAIRA, MEALSA, CLIMSA, CLIM1A, SHOPA, females have higher HR for MAPA (see Supplementary Table S2 for acronyms).

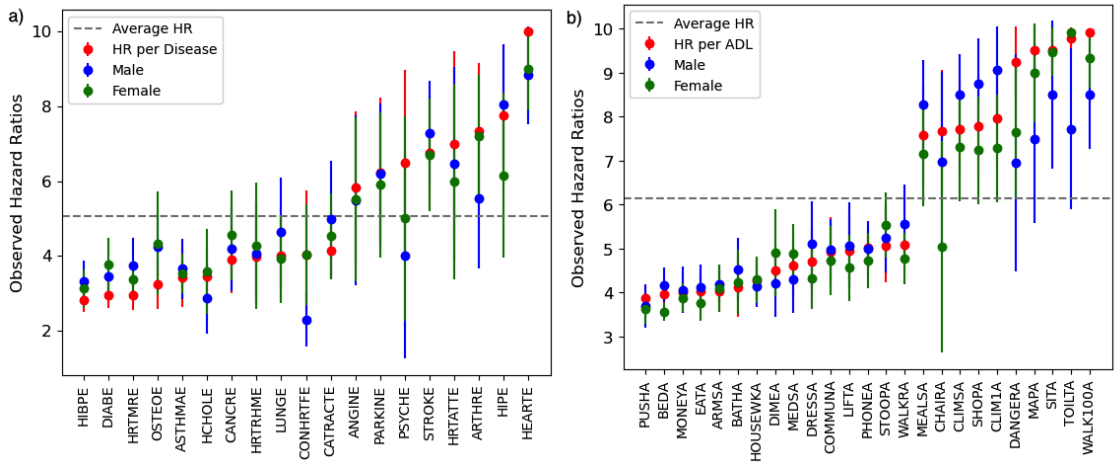

Figure 5.3: **Hazard ratios with males and females.** Observed hazard ratios (HR) between the highest decile of transition probability to the median for a) Diseases and b) ADLs, averaged over all wave transitions. For individual waves see Fig. 5.4. Error bars are standard errors from five-fold cross-validation. The dashed line indicates the average HR.

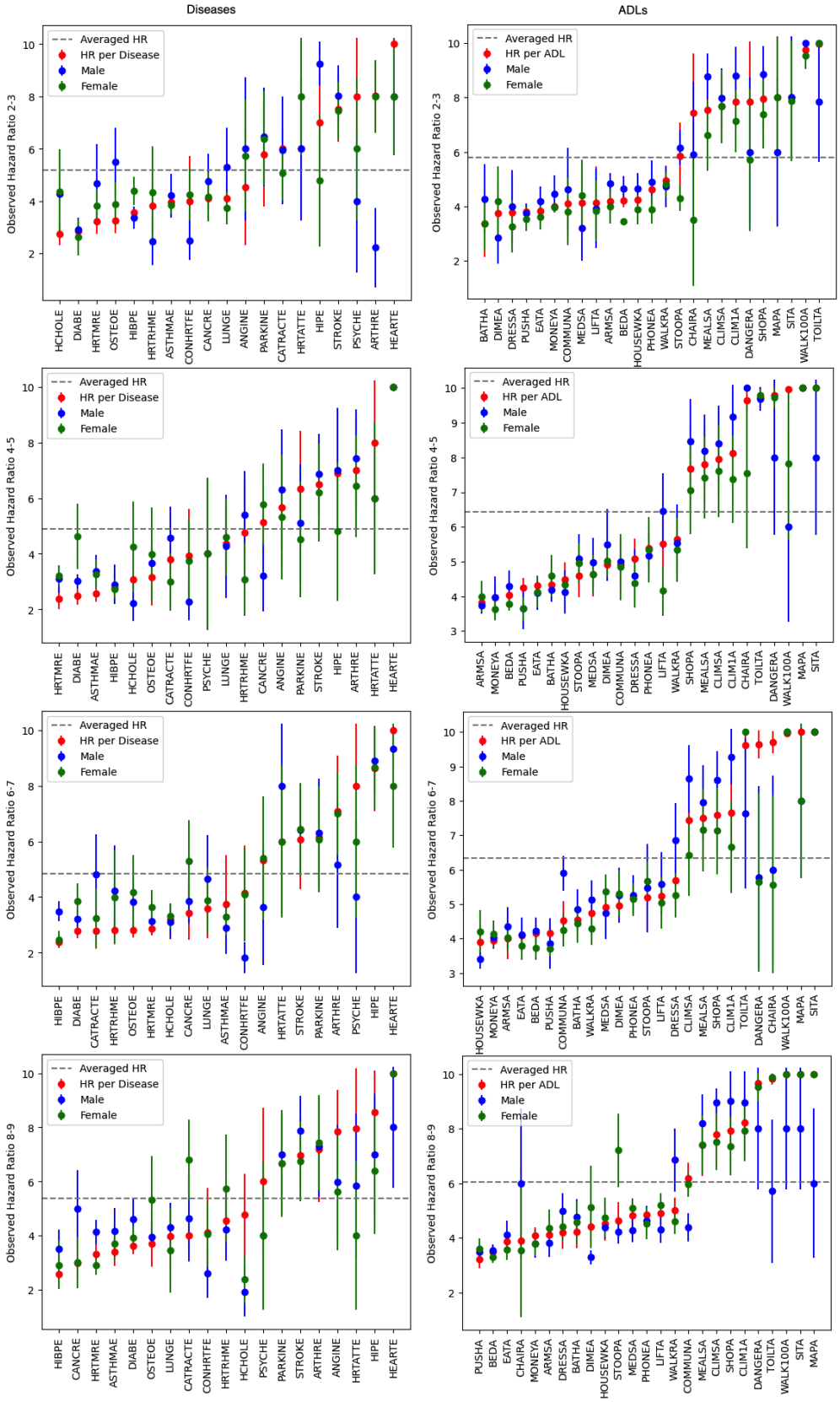

Figure 5.4: HR versus all diseases and ADLs for all, male and females and wave transitions as indicated. HRs are rank ordered for each wave transition. The horizontal dashed line indicates the average HR over all diseases or ADLs, respectively.

5.3 Correlations of Transition Probabilities

Correlations between observed damage transition probabilities and between predicted damage transition probabilities are given in Figs. 5.5 and 5.6 with the hierarchical clustering indicated. We used Pearson correlations, as we did in Figs. 4.16 and 4.17 for health states. We obtained similar behavior for the correlations of damage transition probabilities as with the health state ones –that there is a strong correlation between ADL variables, and a moderate correlation between ADL and disease variables in both observed and predicted cases. There is also a smaller correlation between disease variables. All of the stronger correlations are positive. We do not show diagonal correlations (i.e. variances). For the predicted case, the correlations, the hierarchical clustering and the block structure are similar to the observed case, however the correlations are significantly stronger. The calibration of correlations are studied in next section.

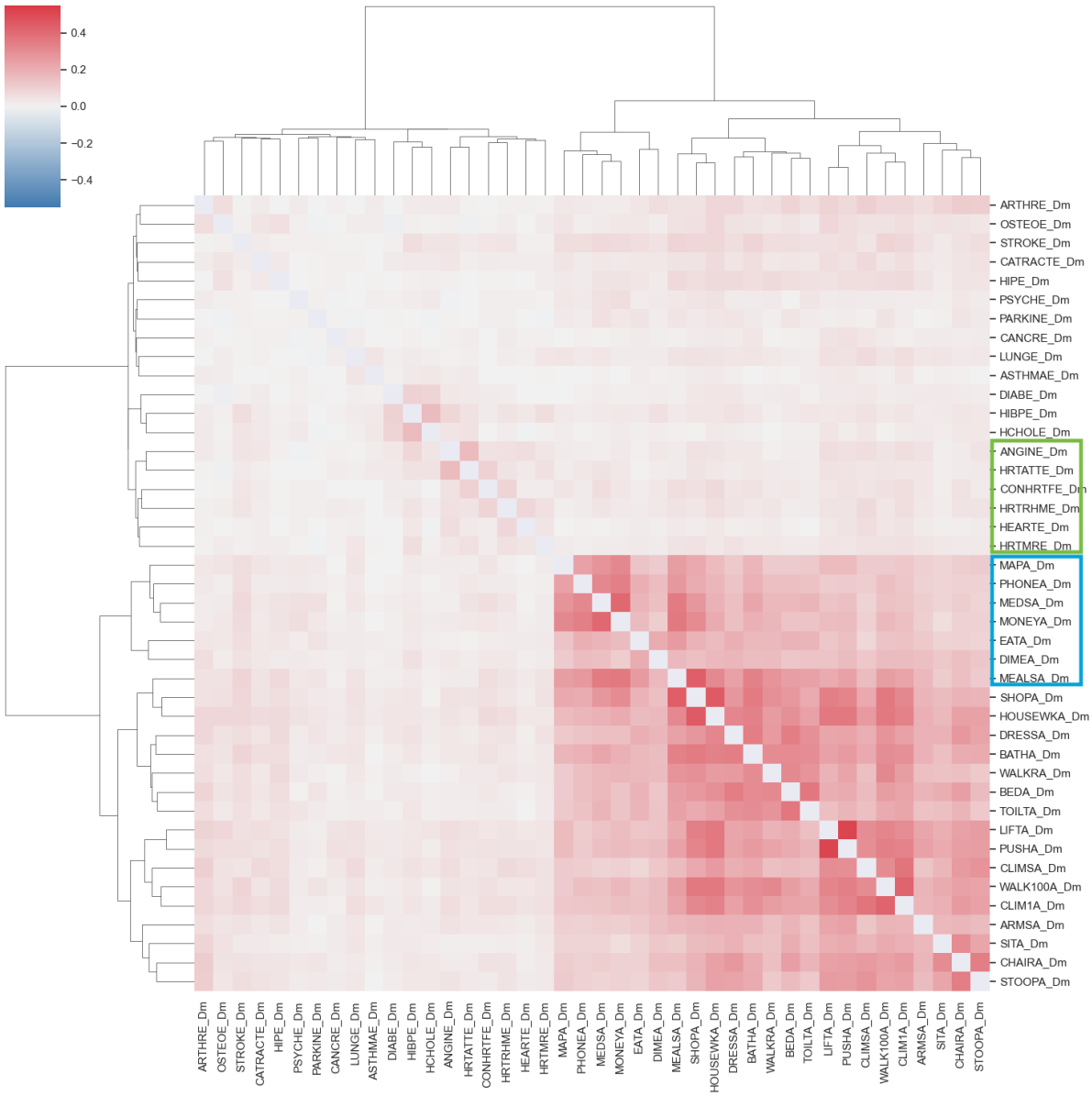

Figure 5.5: Correlations between observed damage transition probabilities. The hierarchical clustering is also indicated. Note that there is a strong correlation between ADL variables, and a moderate correlation between ADL and disease variables. There is a smaller correlation between disease variables. All of the stronger correlations are positive. We do not show diagonal correlations (i.e. variances).

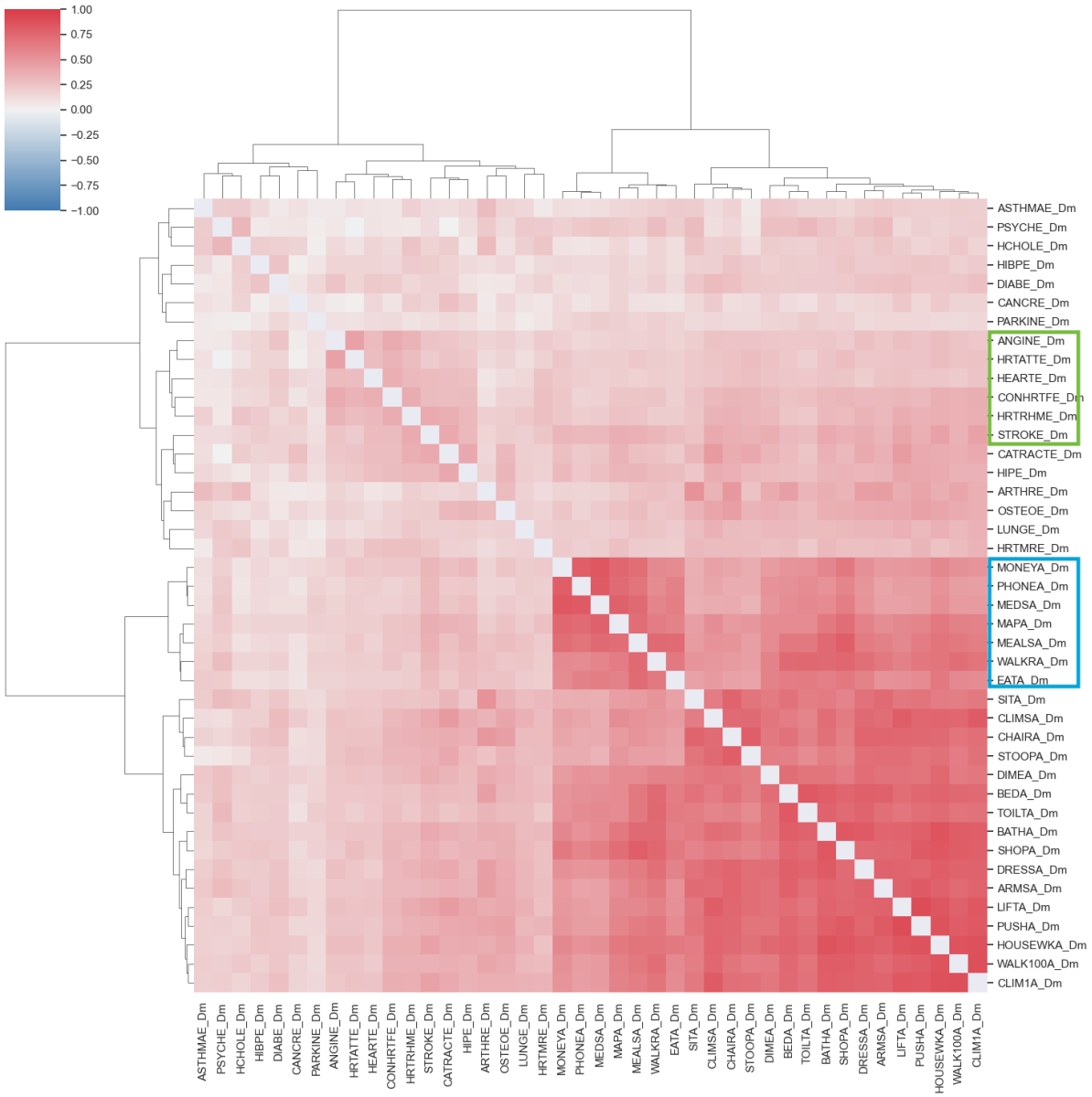

Figure 5.6: Correlations between predicted damage transition probabilities. The hierarchical clustering and block structure are similar to the observed case in Fig. 5.5, however the correlations are significantly stronger. Green and blue boxes show almost same disease and ADLs in observed and predicted cases in Tables 5.5 and 5.6, respectively.

5.4 Calibration of Correlations for Health States and Damage Transition Probabilities

To understand the significance of stronger predicted correlation coefficients than the observed, we perform a simulation study to generate a population with binary health states variables from the predicted probabilities of health outputs. The stochastic simulation algorithm used assigns a value 1 for the randomly generated health state if it is less than the predicted probability, and 0 if it is greater.

$$Y_{\text{sim}} = \begin{cases} 1, & \text{if} \quad \text{rand}(0,1) \leq Y_{\text{pred}} \\ 0, & \text{if} \quad \text{rand}(0,1) > Y_{\text{pred}} \end{cases} \tag{5.1}$$

We finally obtain a simulated population with binary health states of 19 diseases and 25 ADLs. We repeat the simulation 1 time, 100 times and 100 times with averaged health states over each simulation, such that

$$Y_{\text{sim}}(100): \quad \text{has} \quad \text{shape}(100, \text{N}, \text{m}), \tag{5.2}$$

where N is population size and m is the number of health states to be predicted. We average simulated health states over 100 simulation for each individual and predicted health states and obtain

$$Y_{\text{sim}}(\text{averaged}): \quad \text{has} \quad \text{shape}(\text{N}, \text{m}) \tag{5.3}$$

From those simulated health states, we obtain simulated correlation maps, and then we compare these coefficients with the observed, predicted, and simulated correlation coefficients by some scatter plots in Figs. 5.7–5.9.

We see 1 time and 100 times non-averaged simulations have almost same behaviours in Figs 5.7 and 5.8. For the 100 times averaged simulations in 5.9, predicted correlation coefficients are still stronger than observed correlations, and the simulated versus observed ones are compatible with the predicted ones. Lastly, the consistency between predicted and simulated ones are found when the health states are averaged before correlations are taken. We conclude that using predictive approach for future health states does not properly predict correlations but rather approximates correlations with the averaged health states according to the predicted probabilities.

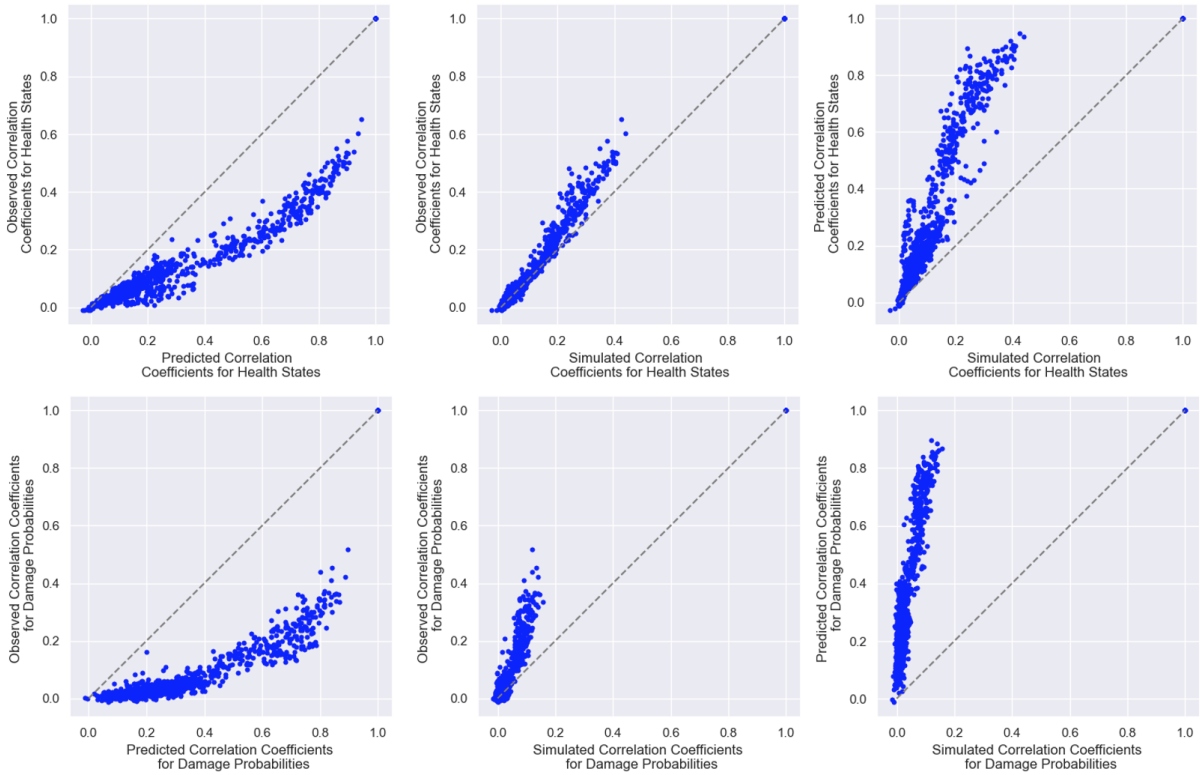

Figure 5.7: Calibration of observed, predicted and simulated health states and damage transition probability correlations. We only simulated 1 population.

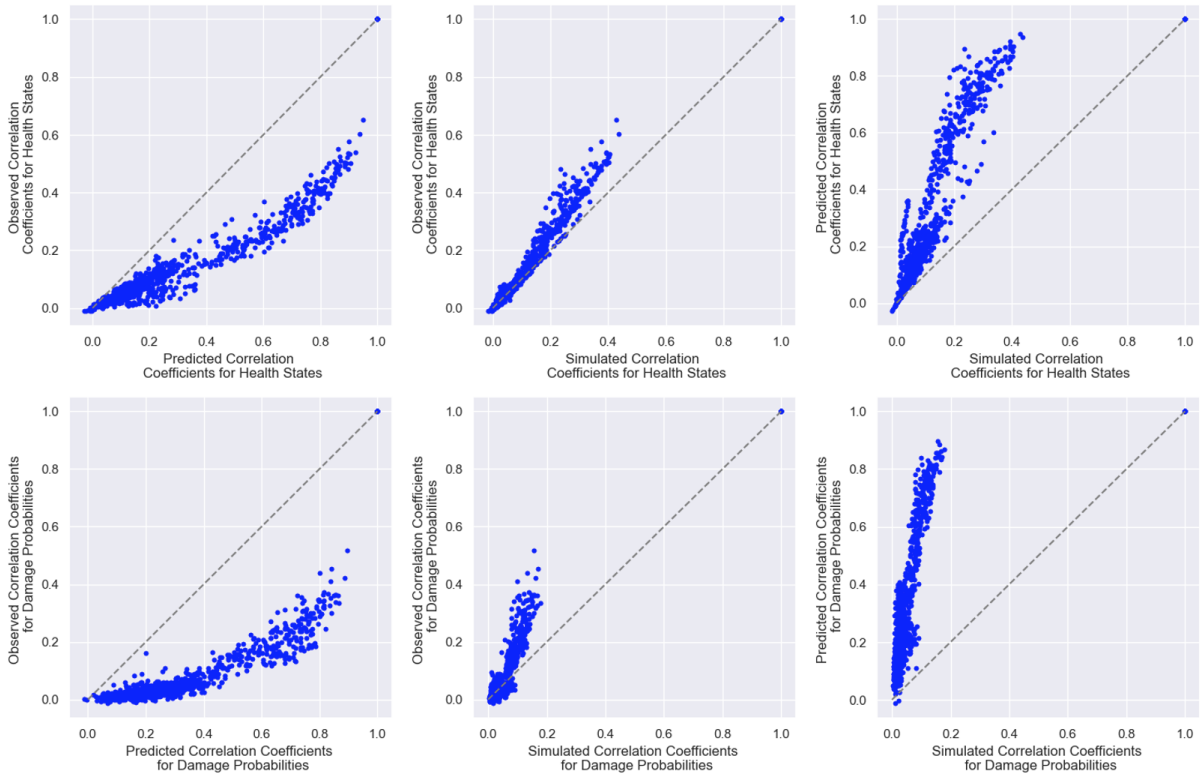

Figure 5.8: Calibration of observed, predicted and simulated health states and damage transition probability correlations. We simulated 100 populations and did not average the health states. Result is almost same as in Fig. 5.7.

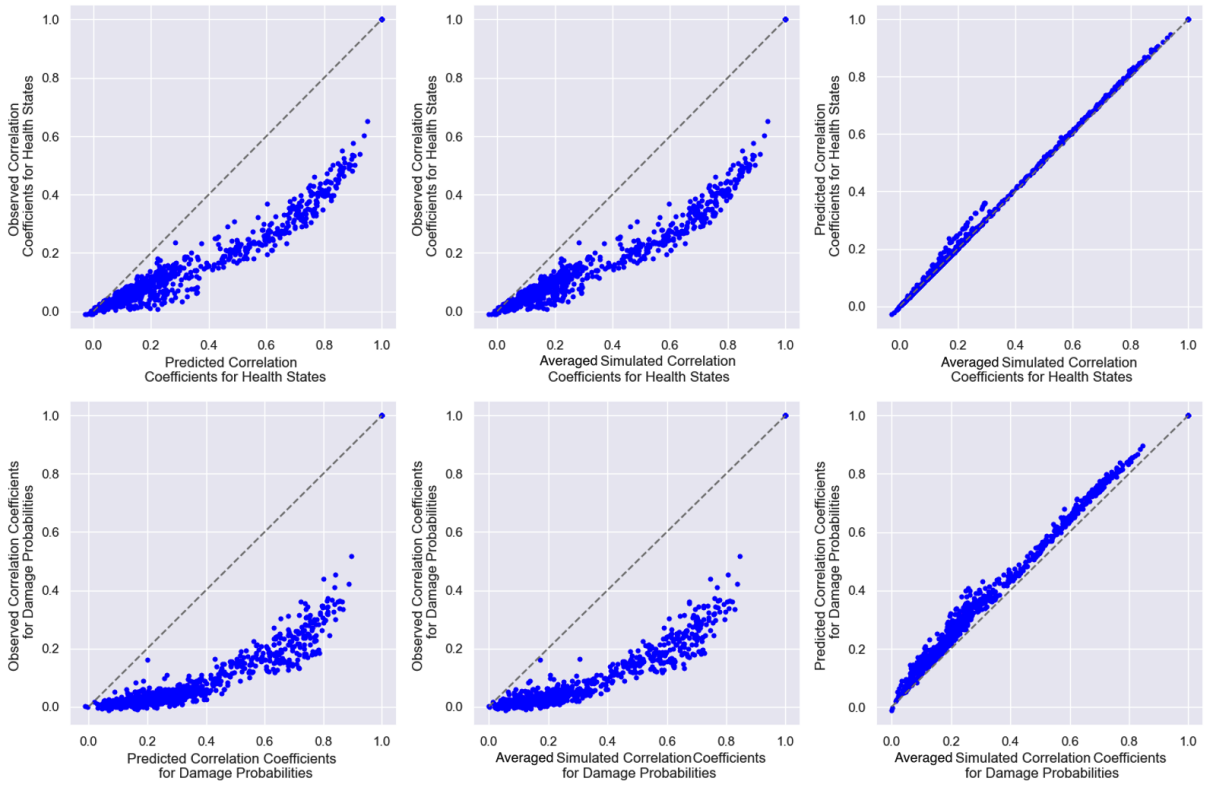

Figure 5.9: Calibration of observed, predicted and simulated health states and damage transition probability correlations. We simulated 100 populations and averaged the simulated health states. Simulated and predicted results are almost same.

5.5 Relations between Prevalences, Exponents of Transition Probabilities and HRs

We also analyze the exponent (or negative slope) of the power law tail of the histograms of disease and ADL damage transition probabilities in Figs. 4.13 and 4.15. For the HRs in Figs. 5.10 and 5.11, we find the relation between exponents and HRs in Fig. 5.12 being similar to the relation in Fig. 4.10 where high HRs correspond to low exponents. We also see the relation between exponents and prevalence such that the high exponents and high prevalence's correspond to each other. Here, the exponents and HRs are predictive quantities while the prevalence are observed quantities. Therefore, we can infer the predictive model is in accordance with the observed data in terms of damage probabilities and prevalence of the diseases and ADLs.

From the rank ordered damage transition probabilities, we obtain the HRs of each health variable by dividing the maximum probability to the mean probability in the deciles for each wave transition. We find a consistent HR picture for health outputs between different visits which the top several high risk diseases and ADLs in all four wave transitions are in accordance as in Supplementary Fig. 4.11. We also find a consistent relation between corresponding HRs and the shape of the damage transition probability distribution being a long-tailed and right-skewed. If the tail is longer toward the probability 1 with a high 0-probability frequency or with a center of mass closer to 0, it has a higher HR for that health variable, as in Supplementary Figs. 4.12–4.15. Because we obtain a positive correlation between the exponent of power law tail of the damage transition probability distributions and prevalence of same disease and ADLs in Fig. 5.12 a and d, we also infer that the relation between exponents and HRs, and prevalence and HRs are inversely correlated in Fig. 5.12 b and e, and c and f. This result shows our predictive results of exponents and HRs are consistent with the observed quantity of prevalence.

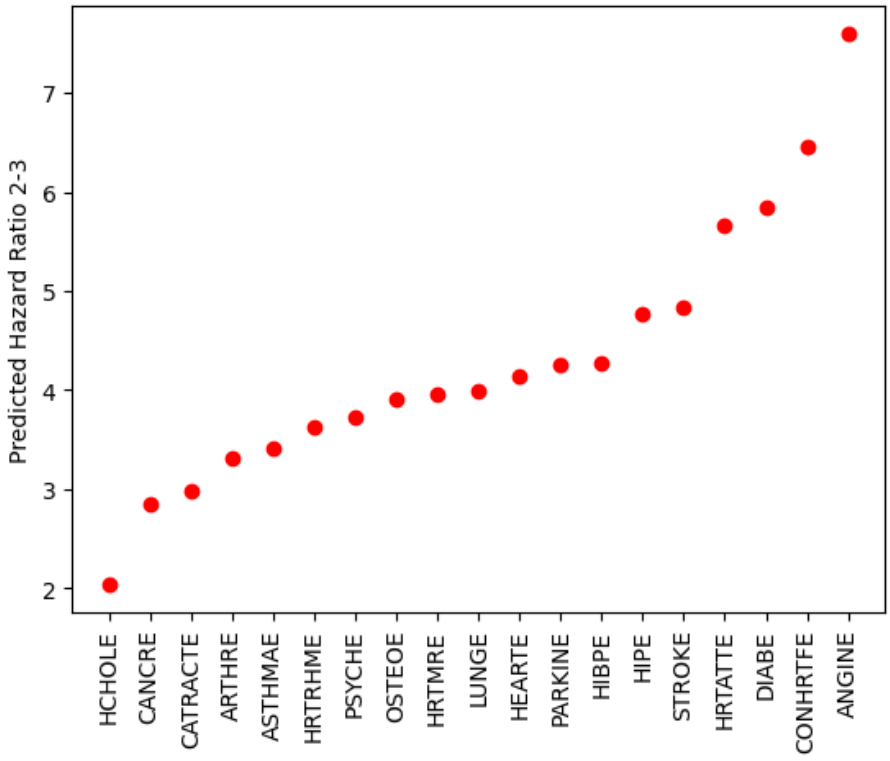

Figure 5.10: Hazard ratios for diseases in wave transition from 2 to 3.

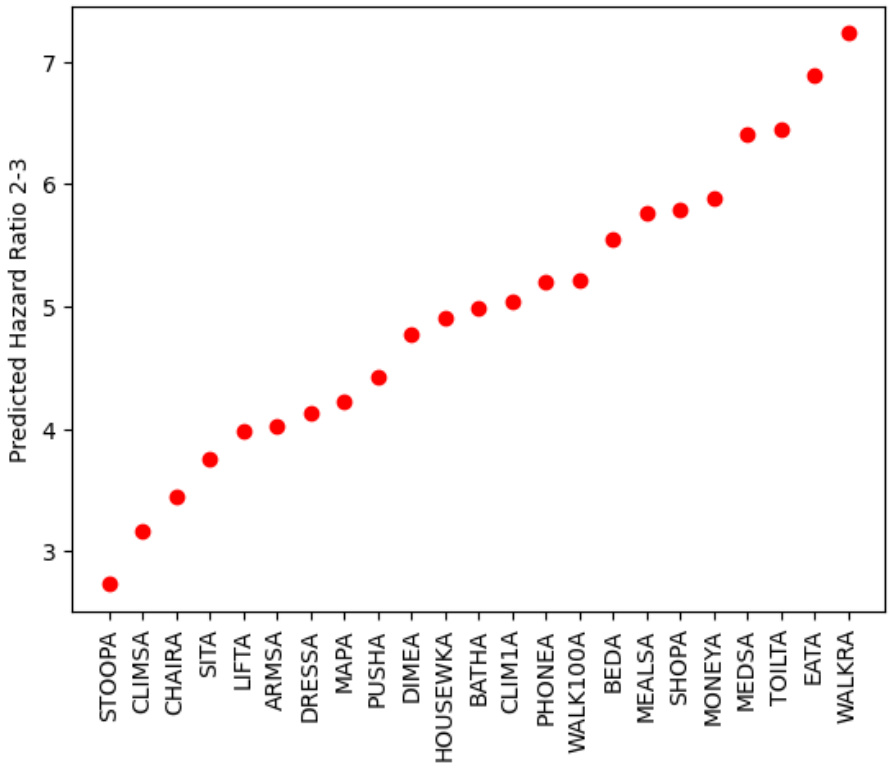

Figure 5.11: Hazard ratios for ADLs in wave transition from 2 to 3.

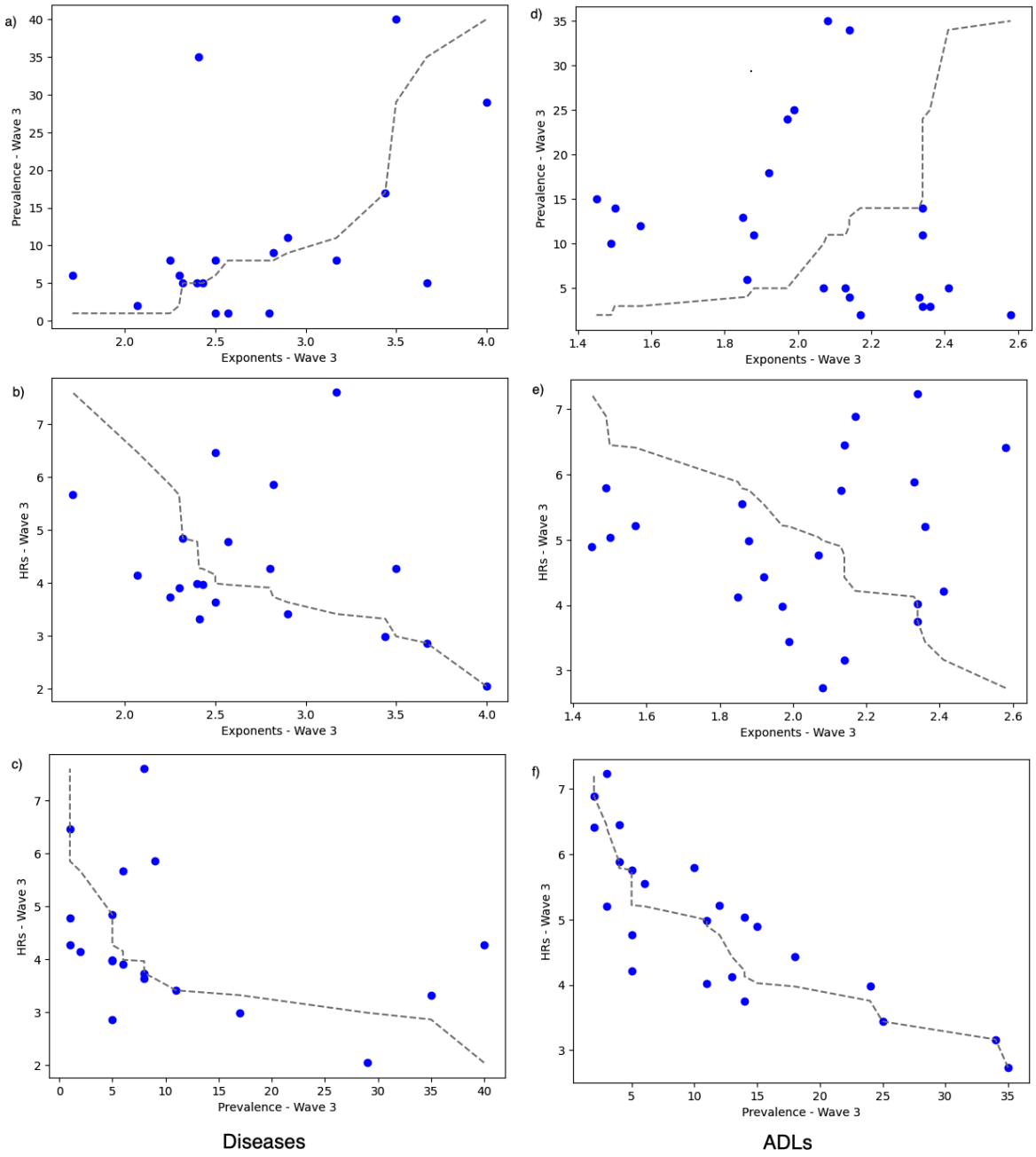

Figure 5.12: Relations between hazard ratios, exponents in damage transition probability distributions, and prevalences. Here, the dashed lines represent the perfect correlation. a) and d) There is an indication for the positive correlations between exponents and prevalences of the same health states. b) and e) We infer an inverse correlation between hazard ratios and exponents of damage distributions. c) and f) We also see similar inverse correlations between hazard ratios and prevalences of the health states.

Chapter 6

Conclusion and Discussion

In this study, we predict future binary health states of 19 diseases and 25 ADLs in the subsequent wave of the ELSA dataset, then obtain the damage and repair probabilities for the considered health states. From several candidate models, a simple deep-neural network (DNN) with 1-hidden layer predicted future states best both with the full 134 features and a selected set of 33 features (see Fig. 4.4a,d), compared to other more complex DNN models, a random forest (RF) model, or a logistic regression (LR) model. Our findings did not agree with the claim that LR is superior at clinical prediction than machine learning models [24] because our best DNN always showed better performance than LR (see Tables 4.11, 4.12, and Fig. 4.4).

We used our best DNN model to obtain health states, average damage and repair probabilities, and individual damage transition probabilities for each health state. We obtained good AUC scores (approximately 0.90) for the binary health state predictions, and excellent $R^2 \geq 0.92$ for average damage and repair probabilities and very good calibrated individual damage transition probabilities with an average Brier score of ≤ 0.035 (see Fig. 4.2 and Table 4.13). For the individual repair transition probabilities, we could not get a well-calibrated results as in Figs 5.1 and 5.2. This means that predicted damage transition probabilities corresponded to observed ones when compared in rank-ordered deciles of the predicted damage probabilities, while this does not hold for the repair transition case. We used well-calibrated damage transition probabilities to determine hazard-ratios (HRs) from the maximum decile to the mean decile ratio where we get the HR values between 3 and 10. We also analyzed the HRs for male and female individuals separately to see if there is a significant risk for some health states more than one gender over another. We could obtain some differences of one gender over another. While the males have higher disease HR for the PARKINE, females have higher disease HRs for the CONHRTFE. While the males have higher ADL HRs for DRESSA, EATA, BEDA, CHAIRA, MEALSA,

CLIMSA, CLIM1A, SHOPA, females have higher ADL HR for MAPA.

We investigated the correlations between health states in Figs. 4.16 and 4.17 and damage transition probabilities and in Figs. 5.5 and 5.6. We found similar behaviours for health states and damage transition probabilities in terms of very strong correlations between observed values than the predicted values. Notably, while correlations among ADL variables are the strongest – the correlations between ADL and disease states are comparable to those between different disease states. We could infer that our DNN model has learned from this correlation structure to better predict health states – which is an advantage of developing a model that predicts all health states (disease and ADL) at once. We also performed a simulation study to generate a population with binary health states variables from the predicted probabilities of health outputs to understand the significance of stronger predicted correlation coefficients. From those simulated health states, we obtain that predicted correlation coefficients are stronger than observed correlations, and the simulated versus observed ones are compatible with the predicted ones, and the linear consistency between predicted and simulated values for the 100 times averaged simulations in 5.9, which implies that using predictive approach for future health states enables us to see stronger correlations between health states and damage probabilities more easily than the observed ones because of the high noise in the observed data.

The analysis of the exponent (or negative slope) of the power law tail of the histograms of disease and ADL damage transition probabilities in Figs. 4.13 and 4.15 and the HRs in Figs. 5.10 and 5.11 give a relation as in Fig. 5.12 that is similar to the relation in Fig. 4.10, where high HRs correspond to low exponents or prevalences. The exponents and HRs are indirectly obtained predictive quantities while the prevalence are observed quantities. In Fig. 4.10a), there is an outlier for "Arthris" but Fig. 4.10b) is generally does not show inverse correlation. From the consistent inversely proportional relation between HRs and exponents, or HRs and prevalences, we infer the predictive model is in accordance with the observed data in terms of damage probabilities and prevalence of the diseases and ADLs.

The limitations of our approach that we mentioned so far are poorly calibrated repair transitions, correlations between health states and correlations between damage transition probabilities. The reason why our model cannot predict individual

repair probabilities accurately when it can predict average repair probabilities for each health state can be related to 1) the low number of populations due to the low number of repaired individuals in each wave and 2) the missing underlying symptoms of the considered health state while the deficit of the health state is still present there. To be able to get better calibrated correlations for health states and damage transition probabilities, the predictive model can be trained by the observed correlation coefficients themselves. Also obtaining worse prediction qualities when applying the transfer learning can be seen another shortcoming of our predictive model due to the significant number of outgoing and incoming individuals for each wave. However, we have not validated our predictive model with other datasets. It's worth highlighting that we made predictions based on the data for both disease and ADL states. However, we couldn't investigate the relationship between these observed health states and the actual underlying diseases or health conditions.

A predictive model is crucial in health sciences for onset of disease, damage and repair. According to the nature of the collected data, one can predict different quantities and different data types. We can predict the most suitable quantity and try to find the correct relations between predicted and desired variables. For instance, we could predict binary health states with a high prediction quality but not the binary transition probabilities and continuous damage rates directly. However, from the well predicted binary health states, we could obtain those continuous average damage and repair probabilities and binary damage transition probabilities indirectly with a high prediction quality. This may not be the case every time for every dataset. Therefore, finding the best predictive model and best features to obtain the target variables considered becomes a difficult task in predictive health science research. For example, individual repair transition probabilities are difficult to predict in ELSA dataset considered here. One can probably develop different models or their combinations for predicting those repair cases.

Bibliography

[1] Alan A Cohen, Luigi Ferrucci, Tamàs Fülöp, Dominique Gravel, Nan Hao, Andres Kriete, Morgan E Levine, Lewis A Lipsitz, Marcel GM Olde Rikkert, Andrew Rutenberg, et al. A complex systems approach to aging biology. *Nature Aging*, 2(7):580–591, 2022. doi: 10.1038/s43587-022-00252-6.

[2] Carlos López-Otín, Maria A Blasco, Linda Partridge, Manuel Serrano, and Guido Kroemer. The hallmarks of aging. *Cell*, 153(6):1194–1217, 2013. doi: 10.1016/j.cell.2013.05.039.

[3] Teresa Niccoli and Linda Partridge. Ageing as a risk factor for disease. *Current Biology*, 22(17):R741–R752, 2012. ISSN 0960-9822. doi: https://doi.org/10.1016/j.cub.2012.07.024.

[4] Zhe Li, Zhenkun Zhang, Yikun Ren, Yingying Wang, Jiarui Fang, Han Yue, Shanshan Ma, and Fangxia Guan. Aging and age-related diseases: from mechanisms to therapeutic strategies. *Biogerontology*, 22(2):165–187, April 2021.

[5] United nations. who world population prospects 2019 — population division, 2019. URL `https://www.un.org/development/desa/pd/news/world-population-prospects-2019-0`.

[6] Wolfgang Lutz, Warren Sanderson, and Sergei Scherbov. The coming acceleration of global population ageing. *Nature*, 451(7179):716–719, 2008. doi: 10.1038/nature06516.

[7] Xunjie Cheng, Yang Yang, David C. Schwebel, Zuyun Liu, Li Li, Peixia Cheng, Peishan Ning, and Guoqing Hu. Population ageing and mortality during 1990–2017: A global decomposition analysis. *PLOS Medicine*, 17(6), 2020. doi: 10.1371/journal.pmed.1003138.

[8] Caio Eduardo Ribeiro, Luis Henrique S Brito, Cristiane Neri Nobre, Alex A Freitas, and Luis Enrique Zárate. A revision and analysis of the comprehensiveness of the main longitudinal studies of human aging for data mining research. *Wiley Interdiscip. Rev. Data Min. Knowl. Discov.*, 7(3):e1202, May 2017. doi: 10.1002/widm.1202.

[9] Caio Eduardo Ribeiro. New longitudinal classification approaches and applications to age-related disease data, 2022.

[10] J. Banks, G. David Batty, J. Breedvelt, K. Coughlin, R. Crawford, M. Marmot, J. Nazroo, Z. Oldfield, N. Steel, A. Steptoe, M. Wood, and P. Zaninotto. *English Longitudinal Study of Ageing: Waves 0-9, 1998-2019.* UK Data Service, UK, 37 edition, 2021. doi: http://doi.org/10.5255/UKDA-SN-5050-25.

[11] Spencer Farrell, Arnold Mitnitski, Kenneth Rockwood, and Andrew D. Rutenberg. Interpretable machine learning for high-dimensional trajectories of aging health. *PLOS Computational Biology*, 18(1):e1009746, 2022. doi: 10.1371/journal.pcbi.1009746.

[12] Spencer Farrell, Alice E Kane, Elise Bisset, Susan E Howlett, and Andrew D Rutenberg. Measurements of damage and repair of binary health attributes in aging mice and humans reveal that robustness and resilience decrease with age, operate over broad timescales, and are affected differently by interventions. *eLife*, 11:e77632, 2022. doi: 10.7554/elife.77632.

[13] Kent Spackman. A program for machine learning of counting criteria: Empirical induction of logic-based classification rules. *Computer Methods and Programs in Biomedicine*, 21(3):221–226, 1985. doi: 10.1016/0169-2607(85)90007-0.

[14] T. Chard. Self-learning for a. bayesian knowledge base: How long does it take for the machine to educate itself? *Methods of Information in Medicine*, 26(04): 185–188, 1987. doi: 10.1055/s-0038-1635507.

[15] Alvin Rajkomar, Jeffrey Dean, and Isaac Kohane. Machine learning in medicine. *New England Journal of Medicine*, 380(14):1347–1358, 2019. doi: 10.1056/nejmra1814259.

[16] Anastasiya Belyaeva, Justin Cosentino, Farhad Hormozdiari, Cory Y. McLean, and Nicholas A. Furlotte. Multimodal LLMs for health grounded in individual-specific data, 2023.

[17] Julián N. Acosta, Guido J. Falcone, Pranav Rajpurkar, and Eric J. Topol. Multimodal biomedical AI. *Nature Medicine*, 28(9):1773–1784, 2022. doi: 10.1038/s41591-022-01981-2.

[18] K. Arumugam, Mohd Naved, Priyanka P. Shinde, Orlando Leiva-Chauca, Antonio Huaman-Osorio, and Tatiana Gonzales-Yanac. Multiple disease prediction using machine learning algorithms. *Materials Today: Proceedings*, 80: 3682–3685, 2023. doi: 10.1016/j.matpr.2021.07.361.

[19] Fabio Fabris, João Pedro Magalhães, and Alex A. Freitas. A review of supervised machine learning applied to ageing research. *Biogerontology*, 18(2):171–188, 2017. doi: 10.1007/s10522-017-9683-y.

[20] Tommy Hielscher, Myra Spiliopoulou, Henry Völzke, and Jens-Peter Kühn. Mining longitudinal epidemiological data to understand a reversible disorder. In *Advances in Intelligent Data Analysis XIII*, volume 8819, pages 120–130, 2014. doi: 10.1007/978-3-319-12571-8_11.

[21] Thore Buergel, Jakob Steinfeldt, Greg Ruyoga, and et al. Metabolomic profiles predict individual multidisease outcomes. *Nature Medicine*, 28(11):2309–2320, Sep 2022. doi: 10.1038/s41591-022-01980-3.

[22] Livia Archibugi, Gianmarco Ciarfaglia, Karina Cárdenas-Jaén, Goran Poropat, Taija Korpela, Patrick Maisonneuve, José R Aparicio, Juan Antonio Casellas, Paolo Giorgio Arcidiacono, Alberto Mariani, et al. Machine learning for the prediction of post-ERCP pancreatitis risk: A proof-of-concept study. *Digestive and Liver Disease*, 55(3):387–393, 2023. doi: https://doi.org/10.1016/j.dld.2022.10.005.

[23] Nhung Nghiem, June Atkinson, Binh P Nguyen, An Tran-Duy, and Nick Wilson. Predicting high health-cost users among people with cardiovascular disease using machine learning and nationwide linked social administrative datasets. *Health Economics Review*, 13(1):9, 2023. doi: https://doi.org/10.1186/s13561-023-00422-1.

[24] Evangelia Christodoulou, Jie Ma, Gary S. Collins, Ewout W. Steyerberg, Jan Y. Verbakel, and Ben Van Calster. A systematic review shows no performance benefit of machine learning over logistic regression for clinical prediction models. *Journal of Clinical Epidemiology*, 110:12–22, 2019. doi: 10.1016/j.jclinepi.2019.02.004.

[25] TingTing Wu, YueQing Wei, JingBing Wu, BiLan Yi, and Hong Li. Logistic regression technique is comparable to complex machine learning algorithms in predicting cognitive impairment related to post intensive care syndrome. *Scientific Reports*, 13(1):2485, 2023. doi: https://doi.org/10.1038/s41598-023-28421-6.

[26] Damian C Stanziano, Michael Whitehurst, Patricia Graham, and Bernard A Roos. A review of selected longitudinal studies on aging: past findings and future directions. *J. Am. Geriatr. Soc.*, 58 Suppl 2(Suppl 2):S292–7, October 2010.

[27] Angelika Kaiser. A review of longitudinal datasets on ageing. *J. Popul. Ageing*, 6(1):5–27, June 2013.

[28] Gateway to global aging data. https://g2aging.org/? Accessed: 2023-10-23.

[29] Peter W. Wilson, Ralph B. D'Agostino, Daniel Levy, Albert M. Belanger, Halit Silbershatz, and William B. Kannel. Prediction of coronary heart disease using risk factor categories. *Circulation*, 97(18):1837–1847, 1998. doi: 10.1161/01.cir.97.18.1837.

[30] Nikhil Bhagwat, Joseph D. Viviano, Aristotle N. Voineskos, and M. Mallar Chakravarty. Modeling and prediction of clinical symptom trajectories in alzheimer's disease using longitudinal data. *PLOS Computational Biology*, 14 (9), 2018. doi: 10.1371/journal.pcbi.1006376.

[31] Lois M. Verbrugge, Dustin C. Brown, and Anna Zajacova. Disability Rises Gradually for a Cohort of Older Americans. *The Journals of Gerontology: Series B*, 72(1):151–161, 03 2016. ISSN 1079-5014. doi: 10.1093/geronb/gbw002. URL https://doi.org/10.1093/geronb/gbw002.

[32] Michael Marmot, James Banks, Richard Blundell, Carli Lessof, and James Nazroo. Health, wealth and lifestyles of the older population in england. *London: Institute of Fiscal Studies*, 2003.

[33] Andrew Steptoe, Elizabeth Breeze, James Banks, and James Nazroo. Cohort profile: the english longitudinal study of ageing. *International journal of epidemiology*, 42(6):1640–1648, 2013.

[34] Nikos Fazakis, Elias Dritsas, Otilia Kocsis, Nikos Fakotakis, and Konstantinos Moustakas. Long-term cholesterol risk prediction using machine learning techniques in elsa database. *Proceedings of the 13th International Joint Conference on Computational Intelligence*, 2021. doi: 10.5220/0010727200003063.

[35] Elias Dritsas, Nikos Fazakis, Otilia Kocsis, Nikos Fakotakis, and Konstantinos Moustakas. Long-term hypertension risk prediction with ml techniques in elsa database. In *Learning and Intelligent Optimization: 15th International Conference, LION 15, Athens, Greece, June 20–25, 2021, Revised Selected Papers 15*, pages 113–120. Springer, 2021.

[36] Daniel Stamate, Henry Musto, Olesya Ajnakina, and Daniel Stahl. Predicting risk of dementia with survival machine learning and statistical methods: Results on the english longitudinal study of ageing cohort. In *IFIP International Conference on Artificial Intelligence Applications and Innovations*, pages 436–447. Springer, 2022.

[37] Yann LeCun, Yoshua Bengio, and Geoffrey Hinton. Deep learning. *Nature*, 521 (7553):436–444, May 2015.

[38] M Shamim Hossain, Ghulam Muhammad, and Nadra Guizani. Explainable AI and mass surveillance System-Based healthcare framework to combat COVID-I9 like pandemics. *IEEE Netw.*, 34(4):126–132, 2020. ISSN 0890-8044, 1558-156X. doi: 10.1109/MNET.011.2000458.

[39] Mohammad Shorfuzzaman and M Shamim Hossain. MetaCOVID: A siamese neural network framework with contrastive loss for n-shot diagnosis of COVID-19 patients. *Pattern Recognit.*, 113:107700, May 2021. ISSN 0031-3203. doi: 10.1016/j.patcog.2020.107700.

[40] Jobeda Jamal Khanam and Simon Y Foo. A comparison of machine learning algorithms for diabetes prediction. *ICT Express*, 7(4):432–439, December 2021. ISSN 2405-9595. doi: 10.1016/j.icte.2021.02.004.

[41] S Soundarya, M S Sruthi, S Sathya Bama, S Kiruthika, and J Dhiyaneswaran. Early detection of alzheimer disease using gadolinium material. *Materials Today: Proceedings*, 45:1094–1101, January 2021. ISSN 2214-7853. doi: 10.1016/j.matpr.2020.03.189.

[42] Syed Nawaz Pasha, Dadi Ramesh, Sallauddin Mohmmad, A Harshavardhan, and Shabana. Cardiovascular disease prediction using deep learning techniques. *IOP Conf. Ser.: Mater. Sci. Eng.*, 981(2):022006, December 2020. ISSN 1757-899X. doi: 10.1088/1757-899X/981/2/022006.

[43] Pranav Rajpurkar, Jeremy Irvin, Kaylie Zhu, Brandon Yang, Hershel Mehta, Tony Duan, Daisy Ding, Aarti Bagul, Curtis Langlotz, Katie Shpanskaya, et al. Chexnet: Radiologist-level pneumonia detection on chest x-rays with deep learning. *arXiv preprint arXiv:1711.05225*, 2017.

[44] Edward Choi, Mohammad Taha Bahadori, Andy Schuetz, Walter F Stewart, and Jimeng Sun. Doctor ai: Predicting clinical events via recurrent neural networks. In *Machine learning for healthcare conference*, pages 301–318. PMLR, 2016.

[45] Kipp W Johnson, Jessica Torres Soto, Benjamin S Glicksberg, Khader Shameer, Riccardo Miotto, Mohsin Ali, Euan Ashley, and Joel T Dudley. Artificial intelligence in cardiology. *Journal of the American College of Cardiology*, 71(23): 2668–2679, 2018.

[46] Zhengping Che, Sanjay Purushotham, Kyunghyun Cho, David Sontag, and Yan Liu. Recurrent neural networks for multivariate time series with missing values. *Scientific reports*, 8(1):6085, 2018.

[47] Wonho Yang, M Benbouchta, and R Yantorno. Performance of the modified bark spectral distortion as an objective speech quality measure. In *Proceedings of the 1998 IEEE International Conference on Acoustics, Speech and Signal Processing, ICASSP '98 (Cat. No.98CH36181)*, volume 1, pages 541–544 vol.1. IEEE, 1998.

[48] Leo Breiman. Random forests. *Mach. Learn.*, 45(1):5–32, October 2001.

[49] D Richard Cutler, Thomas C Edwards, Jr, Karen H Beard, Adele Cutler, Kyle T Hess, Jacob Gibson, and Joshua J Lawler. Random forests for classification in ecology. *Ecology*, 88(11):2783–2792, November 2007.

[50] Wanyue Li, Yanan Song, Kang Chen, Jun Ying, Zhong Zheng, Shen Qiao, Ming Yang, Maonian Zhang, and Ying Zhang. Predictive model and risk analysis for diabetic retinopathy using machine learning: a retrospective cohort study in china. *Bmj Open*, 11(11):e050989, 2021.

[51] Xiangxue Wang, Andrew Janowczyk, Yu Zhou, Rajat Thawani, Pingfu Fu, Kurt Schalper, Vamsidhar Velcheti, and Anant Madabhushi. Prediction of recurrence in early stage non-small cell lung cancer using computer extracted nuclear features from digital h&e images. *Scientific reports*, 7(1):13543, 2017.

[52] Jussi Pirneskoski, Joonas Tamminen, Antti Kallonen, Jouni Nurmi, Markku Kuisma, Klaus T Olkkola, and Sanna Hoppu. Random forest machine learning method outperforms prehospital national early warning score for predicting one-day mortality: A retrospective study. *Resuscitation plus*, 4:100046, 2020.

[53] Alex Bottle, Paul Aylin, and Azeem Majeed. Identifying patients at high risk of emergency hospital admissions: a logistic regression analysis. *Journal of the Royal Society of Medicine*, 99(8):406–414, 2006.

[54] Joseph S Ross, Gregory K Mulvey, Brett Stauffer, Vishnu Patlolla, Susannah M Bernheim, Patricia S Keenan, and Harlan M Krumholz. Statistical models and patient predictors of readmission for heart failure: a systematic review. *Archives of internal medicine*, 168(13):1371–1386, 2008.

[55] Behrad Barghi and Nasibeh Azadeh-Fard. Predicting risk of sepsis, comparison between machine learning methods: a case study of a virginia hospital. *European Journal of Medical Research*, 27(1):213, 2022.

[56] Bita Farhadi, Jiaxue You, Dexu Zheng, Lu Liu, Sajian Wu, Jianxun Li, Zhipeng Li, Kai Wang, and Shengzhong Liu. Machine learning for fast development of advanced energy materials. *Next Materials*, 1(3):100025, 2023.

[57] Susan E Howlett and Kenneth Rockwood. New horizons in frailty: ageing and the deficit-scaling problem. *Age Ageing*, 42(4):416–423, July 2013.

[58] Qing Wang. Cancer predisposition genes: molecular mechanisms and clinical impact on personalized cancer care: examples of lynch and hboc syndromes. *Acta Pharmacologica Sinica*, 37(2):143–149, 2016.

[59] Yuanliang Xia, Jianshu Zhu, Ruohan Yang, Hengyi Wang, Yuehong Li, and Changfeng Fu. Mesenchymal stem cells in the treatment of spinal cord injury: Mechanisms, current advances and future challenges. *Frontiers in Immunology*, 14:1141601, 2023.

[60] George K Michalopoulos. Liver regeneration. *Journal of cellular physiology*, 213(2):286–300, 2007.

[61] Chara Papalouka, Maria Adamaki, Panagiota Batsaki, Panagiotis Zoumpourlis, Antonis Tsintarakis, Maria Goulielmaki, Sotirios P Fortis, Constantin N Baxevanis, and Vassilis Zoumpourlis. Dna damage response mechanisms in head and neck cancer: significant implications for therapy and survival. *International Journal of Molecular Sciences*, 24(3):2760, 2023.

[62] Benjamin J Moeller, John S Yordy, Michelle D Williams, Uma Giri, Uma Raju, David P Molkentine, Lauren A Byers, John V Heymach, Michael D Story, J Jack Lee, et al. Dna repair biomarker profiling of head and neck cancer: Ku80 expression predicts locoregional failure and death following radiotherapy. *Clinical Cancer Research*, 17(7):2035–2043, 2011.

[63] Nadine Tung, Chiara Battelli, Brian Allen, Rajesh Kaldate, Satish Bhatnagar, Karla Bowles, Kirsten Timms, Judy E Garber, Christina Herold, Leif Ellisen, et al. Frequency of mutations in individuals with breast cancer referred for brca 1 and brca 2 testing using next-generation sequencing with a 25-gene panel. *Cancer*, 121(1):25–33, 2015.

[64] David Gems and João Pedro de Magalhães. The hoverfly and the wasp: A critique of the hallmarks of aging as a paradigm. *Ageing Res. Rev.*, 70:101407, September 2021.

[65] A B Mitnitski, A J Mogilner, and K Rockwood. Accumulation of deficits as a proxy measure of aging. *ScientificWorldJournal*, 1:323–336, August 2001.

[66] Frailty in older adults: evidence for a phenotype. *J. Gerontol. A Biol. Sci. Med. Sci.*, 56(3):M146–56, March 2001.

[67] Arnold Mitnitski, Xiaowei Song, and Kenneth Rockwood. Improvement and decline in health status from late middle age: modeling age-related changes in deficit accumulation. *Exp. Gerontol.*, 42(11):1109–1115, November 2007.

[68] Gotaro Kojima, Yu Taniguchi, Steve Iliffe, Stephen Jivraj, and Kate Walters. Transitions between frailty states among community-dwelling older people: A systematic review and meta-analysis. *Ageing Res. Rev.*, 50:81–88, March 2019.

[69] Svetlana Ukraintseva, Konstantin Arbeev, Matt Duan, Igor Akushevich, Alexander Kulminski, Eric Stallard, and Anatoliy Yashin. Decline in biological resilience as key manifestation of aging: Potential mechanisms and role in health and longevity. *Mech. Ageing Dev.*, 194:111418, March 2021.

[70] James L Kirkland, Michael B Stout, and Felipe Sierra. Resilience in aging mice. *J. Gerontol. A Biol. Sci. Med. Sci.*, 71(11):1407–1414, November 2016.

[71] Evan C Hadley, George A Kuchel, Anne B Newman, and Workshop Speakers and Participants. Report: NIA workshop on measures of physiologic resiliencies in human aging. *J. Gerontol. A Biol. Sci. Med. Sci.*, 72(7):980–990, July 2017.

[72] Konstantin G Arbeev, Svetlana V Ukraintseva, Olivia Bagley, Ilya Y Zhbannikov, Alan A Cohen, Alexander M Kulminski, and Anatoliy I Yashin. "physiological dysregulation" as a promising measure of robustness and resilience in studies of aging and a new indicator of preclinical disease. *J. Gerontol. A Biol. Sci. Med. Sci.*, 74(4):462–468, March 2019.

[73] Andres Kriete. Robustness and aging–a systems-level perspective. *Biosystems.*, 112(1):37–48, April 2013.

[74] Maher M. El-Masri and Susan M. Fox-Wasylyshyn. Missing data: An introductory conceptual overview for the novice researcher. *Canadian Journal of Nursing Research*, 37(4):156–171, 2005. doi: https://scholar.uwindsor.ca/nursingpub/87/.

[75] Jürgen Schmidhuber. Deep learning in neural networks: An overview. *Neural networks*, 61:85–117, 2015.

[76] Ian Goodfellow, Yoshua Bengio, and Aaron Courville. *Deep learning*. MIT press, 2016.

[77] Michael A Nielsen. *Neural networks and deep learning*, volume 25. Determination press San Francisco, CA, USA, 2015.

[78] Yann LeCun, Yoshua Bengio, and Geoffrey Hinton. Deep learning. *nature*, 521 (7553):436–444, 2015.

[79] Leo Breiman. Random forests. *Machine learning*, 45:5–32, 2001.

[80] Andy Liaw, Matthew Wiener, et al. Classification and regression by randomforest. *R news*, 2(3):18–22, 2002.

[81] David W Hosmer Jr, Stanley Lemeshow, and Rodney X Sturdivant. *Applied logistic regression*, volume 398. John Wiley & Sons, 2013.

[82] Pankaj Mehta, Marin Bukov, Ching-Hao Wang, Alexandre G.R. Day, Clint Richardson, Charles K. Fisher, and David J. Schwab. A high-bias, low-variance introduction to machine learning for physicists. *Physics Reports*, 810:1–124, 2019. ISSN 0370-1573. doi: https://doi.org/10.1016/j.physrep.2019.03.001. A high-bias, low-variance introduction to Machine Learning for physicists.

[83] Alan C. Acock. Working with missing data. *Family Science Review*, 1(10): 76–102, 1997.

[84] Mark R. Raymond and Dennis M. Roberts. A comparison of methods for treating incomplete data in selection research. *Educational and Psychological Measurement*, 47(1):13–26, 1987. doi: 10.1177/0013164487471002.

[85] Zhongheng Zhang. Missing data imputation: focusing on single imputation. *Annals of translational medicine*, 4(1):9, 2016. doi: 10.3978/j.issn.2305-5839.2015.12.38.

[86] A Jović, K Brkić, and N Bogunović. A review of feature selection methods with applications. In *2015 38th International Convention on Information and Communication Technology, Electronics and Microelectronics (MIPRO)*, pages 1200–1205, May 2015. doi: 10.1109/mipro.2015.7160458.

[87] Isabelle Guyon, Jason Weston, Stephen Barnhill, and Vladimir Vapnik. Gene selection for cancer classification using support vector machines. *Machine Learning*, 46(1/3):389–422, 2002. doi: 10.1023/a:1012487302797.

[88] Glen Pridham, Kenneth Rockwood, and Andrew Rutenberg. Efficient representations of binarized health deficit data: the frailty index and beyond. *Geroscience*, 45(3):1687–1711, June 2023. doi: 10.1007/s11357-022-00723-z.

[89] Brian K Kennedy, Shelley L Berger, Anne Brunet, Judith Campisi, Ana Maria Cuervo, Elissa S Epel, Claudio Franceschi, Gordon J Lithgow, Richard I Morimoto, Jeffrey E Pessin, et al. Geroscience: linking aging to chronic disease. *Cell*, 159(4):709–713, 2014. doi: 10.1016/j.cell.2014.10.039.

[90] Thomas BL Kirkwood. Understanding the odd science of aging. *Cell*, 120(4): 437–447, 2005. doi: 10.1016/j.cell.2005.01.027.

[91] Spencer Farrell, Garrett Stubbings, Kenneth Rockwood, Arnold Mitnitski, and Andrew Rutenberg. The potential for complex computational models of aging. *Mechanisms of Ageing and Development*, 193:111403, 2021. ISSN 0047-6374. doi: https://doi.org/10.1016/j.mad.2020.111403.

[92] Yingxiang Huang, Wentao Li, Fima Macheret, Rodney A Gabriel, and Lucila Ohno-Machado. A tutorial on calibration measurements and calibration models for clinical prediction models. *Journal of the American Medical Informatics Association*, 27(4):621–633, 02 2020. doi: 10.1093/jamia/ocz228.

[93] Stef van Buuren. *Flexible imputation of missing data*. Chapman and Hall/CRC, 2021. doi: https://doi.org/10.1201/9780429492259.

[94] David M. W. Powers. Evaluation: from precision, recall and F-measure to ROC, informedness, markedness and correlation, 2020.

[95] Glen Pridham, Kenneth Rockwood, and Andrew Rutenberg. Strategies for handling missing data that improve frailty index estimation and predictive power: Lessons from the NHANES dataset. *GeroScience*, 44(2):897–923, 2022. doi: 10.1007/s11357-021-00489-w.

[96] Keras-Team. Keras documentation: Probabilistic losses, 2023. URL `https://keras.io/api/losses/probabilistic_losses/#binarycrossentropy-class`.

[97] Guillaume Lemaître, Fernando Nogueira, and Christos K. Aridas. Imbalanced-learn: A python toolbox to tackle the curse of imbalanced datasets in machine learning. *Journal of Machine Learning Research*, 18(17):1–5, 2017. URL `http://jmlr.org/papers/v18/16-365.html`.

[98] Alexandru Niculescu-Mizil and Rich Caruana. Predicting good probabilities with supervised learning. In *Proceedings of the 22nd international conference on Machine learning*, ICML '05, pages 625–632, New York, NY, USA, August 2005. Association for Computing Machinery. doi: 10.1145/1102351.1102430.

[99] Glenn W. Brier. Verification of forecasts expressed in terms of probability. *Monthly Weather Review*, 78(1):1–3, 1950. doi: https://doi.org/10.1175/1520-0493(1950)078⟨0001:VOFEIT⟩2.0.CO;2.

[100] Binh P Nguyen, Hung N Pham, Hop Tran, Nhung Nghiem, Quang H Nguyen, Trang TT Do, Cao Truong Tran, and Colin R Simpson. Predicting the onset of type 2 diabetes using wide and deep learning with electronic health records. *Computer methods and programs in biomedicine*, 182:105055, 2019. doi: https://doi.org/10.1016/j.cmpb.2019.105055.

[101] Lu Men, Noyan Ilk, Xinlin Tang, and Yuan Liu. Multi-disease prediction using LSTM recurrent neural networks. *Expert Systems with Applications*, 177: 114905, 2021. doi: https://doi.org/10.1016/j.eswa.2021.114905.

[102] Shelley-Ann M Girwar, Robert Jabroer, Marta Fiocco, Stephen P Sutch, Mattijs E Numans, and Marc A Bruijnzeels. A systematic review of risk stratification tools internationally used in primary care settings. *Health Science Reports*, 4(3):e329, 2021. doi: https://doi.org/10.1002/hsr2.329.

[103] Joana Mora, Miren David Iturralde, Lucía Prieto, and et al. Key aspects related to implementation of risk stratification in health care systems-the assehs study. *BMC health services research*, 17:1–8, 2017. doi: https://doi.org/10.1186/s12913-017-2275-3.

Appendix A

Supplementary Material

Tables S1-S2. Percentage prevalence of Diseases and ADLs for all waves.

	Health Output	W1	W2	W3	W4	W5	W6	W7	W8	W9	Av.
PARKINE	ever had parkinson disease	0	1	1	1	1	1	1	1	1	1
CONHRTFE	ever had congestive heart failure	1	1	1	1	1	1	1	1	1	1
HIPE	ever had hip fracture	2	1	1	1	1	1	1	1	1	1
HEARTE	ever had heart problems	0	1	2	3	3	4	2	2	6	2
HRTMRE	ever had heart murmur	5	4	5	4	4	4	4	4	3	4
HRTATTE	ever had heart attack	0	6	6	6	6	5	5	5	1	4
STROKE	ever had stroke	4	5	5	4	5	5	5	5	3	5
LUNGE	ever had lung disease	6	7	5	5	5	5	5	5	5	5
CANCRE	ever had cancer	6	7	5	5	6	6	6	7	8	6
ANGINE	ever had angina	9	8	8	7	9	6	3	3	1	6
OSTEOE	ever had osteoporosis	5	6	6	7	8	7	8	8	8	7
HRTRHME	ever had abnormal heart rhythm	6	7	8	7	8	8	9	10	8	8
PSYCHE	ever had psych problems	8	9	8	9	10	10	10	10	8	9
DIABE	ever had diabetes	7	6	9	9	11	11	12	13	12	10
ASTHMAE	ever had asthma	12	13	11	11	11	11	11	11	11	11
CATRACTE	ever had cataracts	13	17	17	17	20	22	25	30	30	21
HCHOLE	ever had high cholesterol	0	18	29	32	36	34	35	36	32	28
ARTHRE	ever had arthritis	32	36	35	34	37	37	37	40	38	36
HIBPE	ever had high blood pressure	37	32	40	38	40	38	38	39	36	37

Table S1: Percentage prevalence of 19-Disease outputs. Av. stands for average of the values in all waves, W1, W2, ..., W9

	Health Output	W1	W2	W3	W4	W5	W6	W7	W8	W9	Av.
DANGERA	Some Diff-Recognizing when in physical danger	0	0	0	2	2	2	2	2	2	1
EATA	Some Diff-Eating	2	2	2	2	2	3	2	3	3	2
MEDSA	Some Diff-Take medications	1	2	2	2	3	3	3	3	3	3
COMMUNA	Some Diff-Communication (speech, hearing, eye)	0	0	0	4	4	4	4	4	4	3
PHONEA	Some Diff-Use a telephone	2	2	3	3	3	3	3	3	3	3
BEDA	Some Diff-Get in/out bed	2	6	6	6	6	6	6	6	6	6
MONEYA	Some Diff-Managing money	2	3	4	3	4	4	4	4	4	4
WALKRA	Some Diff-Walk across room	3	4	3	3	3	4	4	4	4	4
TOILTA	Some Diff-Using the toilet	3	3	4	3	4	4	4	4	4	4
MEALSA	Some Diff-Prepare hot meal	4	5	5	5	5	5	5	6	6	5
MAPA	Some Diff-Use a map	5	6	5	5	5	5	5	5	5	5
BATHA	Some Diff-Bathing, shower	12	12	11	10	10	10	9	9	9	10
DIMEA	Some Diff-Pick up a 5p coin	5	5	5	5	6	6	6	6	6	6
SHOPA	Some Diff-Shop for grocery	9	10	10	9	10	9	9	10	9	9
ARMSA	Some Diff-Rch/xtnd arms up	11	11	11	10	11	12	11	11	10	11
DRESSA	Some Diff-Dressing	13	14	13	13	13	13	12	13	13	13
WALK100A	Some Diff-Walk 100y	12	12	12	12	14	14	13	14	13	13
SITA	Some Diff-Sit for 2 hours	14	14	14	12	13	13	13	13	11	13
CLIM1A	Some Diff-Clmb 1 flt str	15	15	14	14	15	15	15	15	14	15
HOUSEWKA	Some Diff-Doing work around the house or garden	16	17	15	15	16	15	15	16	15	15
PUSHA	Some Diff-Push/pull lg obj	18	19	18	17	18	18	17	18	16	18
LIFTA	Some Diff-Lift/carry 10lbs	26	25	24	23	23	23	22	23	21	23
CHAIRA	Some Diff-Get up fr chair	26	27	25	25	25	24	24	24	23	25
CLIMSA	Some Diff-Clmb sev flt str	36	38	34	34	35	32	31	33	30	34
STOOPA	Some Diff-Stoop/Kneel/Crch	35	38	35	35	38	37	37	39	37	37

Table S2: Percentage prevalence of 25-ADL outputs. Av. stands for average of the values in all waves, W1, W2, ..., W9