Statistical Methods for Modelling Complex Longitudinal Data with Applications in Cancer Pharmacogenetics and Ageing

Evanthia Koukouli, B.Sc., M.Sc.

Submitted for the degree of Doctor of Statistics at Lancaster university.

October, 2022

Abstract

Technological and scientific advancements have promoted data gathering across multiple disciplines emphasizing the necessity for the development of rigorous statistical methods to draw conclusions. Longitudinal data is a key tool to study temporal changes, however, with the increasing data complexity, existing methodologies are often unable to capture non-linear or non-stationary trends. Additionally, irregularly collected, non-continuous or high-dimensional data make statistical analysis even more challenging. Through this work, we develop three statistical models to analyse complex longitudinal data from two real-world databases, the Genomics of Drug Sensitivity in Cancer and the English Longitudinal Study of Ageing.

The first part of this work is motivated by the Genomics of Drug Sensitivity in Cancer project and focuses on the prediction and detection of biomarkers associated with anti-cancer drug dose-response. Here, the longitudinal data available are characterised by complete observed trajectories of drug response over multiple drug dosages which are potentially associated with high-dimensional covariates (these include expression profiles of tens of thousands of genes) in a non-stationary manner. These trends are not easily amenable to analysis by classic parametric or semi-parametric mixed models, especially if high dimensionality is present. We built a dose-varying regression model combined with a two-stage variable selection algorithm (variable screening followed by penalised regression) to identify genetic factors associated with drug response and estimate their effect over the varying dosages.

The second part of this work is motivated by the English Longitudinal Study of Ageing data set. The longitudinal data available in this study are characterised by irregularly collected and, often, incomplete trajectories and many response variables of ordinal type which measure only a small number of ageing domains (data are derived from multiple questionnaires measuring multiple aspects of older peoples' life). The

ultimate aim is to understand the ageing dynamics and study the interrelationships between factors associated with it. To do so, we first explore the theoretical foundations of ageing and the data set itself. Next, we adopt and extend the methodological framework of Dawson and Müller (2018) to estimate the quantile dynamics and derive predictions for a common surrogate of ageing, frailty, addressing the problem of incomplete individual responses over the age interval of interest. Finally, we develop a bivariate Gaussian process framework for ordinal and potentially irregularly sampled data which allows the available questionnaire responses to be modelled directly. Here, the unobserved ageing domains are assumed to be smooth functions of age. This method allows the assessment of the interrelationships between several ageing domains after adjusting for individual variation across the observed longitudinal trajectories.

Acknowledgement

Above all things, I would like to thank my supervisors Dr Juhyun Park, Prof Andrew Titman, Dr Stefanie Doebler and Dr Frank Dondelinger for guiding me through my PhD; their help and support have been invaluable for the completion of this thesis. Special thanks to Dr Juhyun Park for her encouragement and assistance over the whole course of my PhD. Thanks to Prof Andrew Titman who took over my supervision halfway to completion; he was always there when I needed him to provide his valuable advice. Many thanks to Dr Frank Dondelinger who supported me during the first stage of my PhD. I would also like to thank Dr Stefanie Doebler from the Sociology Department of Lancaster University and Dr Dennis Wang from the University of Sheffield for providing their insights without which findings interpretation would be extremely challenging. I could not miss mentioning my PhD chairs, Dr Simon Lunagomez and Prof Brian Francis for their irreplaceable comments on my work.

I would also like to thank my partner, Apostolis, for his constant love and encouragement and my friends, Anastasia and Ierotheos for always being there for me, even during the most challenging times. I am also incredibly grateful for having my dog, Roxann, which was always beside me (most of the time sleeping) accompanying me throughout my write-up stage. I am also grateful for my parents, Avgoustina and Dimitris for their love and support and my brother Panagiotis for reminding me how to be cheerful and relaxed. I could not miss mentioning my very dear friends in Greece, Eleni, Eva, Vaso, Stavroula and Ioanna who were always there to cheer me up.

Thanks to Cyrus Gaviri from the IT section of the Mathematics and Statistics department at Lancaster University who was patient and always present when I had a problem with my laptop or when I was facing trouble with the high-end computing server. Without his help, this thesis could not have been completed. Furthermore, I acknowledge Emily Chambers and the Sheffield Bioinformatics Core Facility for their

assistance in GDSC2 data pre-processing.

I could not miss mentioning the Economic and Social Research Council, to which I am extremely grateful for funding my research through the Advanced Quantitative Methods award. It helped me achieve my dream.

This thesis is dedicated to my grandmother, Evanthia.

Declaration

I declare that the work in this thesis has been done by myself, except where noted below, and has not been submitted elsewhere for the award of any other degree. The word count of this thesis (including the bibliography) is 85,515 words.

The content in Chapter 4 has been accepted for publication as Koukouli E, Wang D, Dondelinger F, Park J (2021) A regularized functional regression model enabling transcriptome-wide dosage-dependent association study of cancer drug response. PLoS Comput Biol 17(1): e1008066. `https://doi.org/10.1371/journal.pcbi.1008066`. In this chapter, I contributed the analysis and writing of all sections apart from some parts of Section 4.4 where Dr Dennis Wang and Dr Frank Dondelinger contributed to the interpretation of the biological findings.

Evanthia Koukouli

Contents

List of Figures

List of Tables

List of Abbreviations

ADLs	Activities of Daily Living
AUC	Area Under the Curve
BiLGP	Bivariate Latent Gaussian Process
BMI	Body Mass Index
CAPI	Computer-Assisted Personal Interview
CCLE	Cancer Cell Line Encyclopedia
CF	Cognitive function
CFA	Confirmatory Factor Analysis
CFI	Comparative Fit Index
CMAP	Connectivity Map
CQ	Conditional Distribution and Quantiles
CVD	Cardiovascular Disease
EFA	Exploratory Factor Analysis
ELSA	English Longitudinal Study of Ageing
EM	Expectation-Maximation
FA	Factor Analysis
FI	Frailty Index
GDSC	Genomics of Drug Sensitivity in Cancer
Greedy INIS	Greedy Iterative Non-parametric Independence Screening
gSCAD	group Smoothly Clipped Absolute Deviation
HAP	Healthy Ageing Phenotype
HSE	Health Survey for England
IADLs	Instrumental Activities of Daily Living
ICGC	International Cancer Genome Consortium
IQR	Interquartile Range
IRT	Item Response Theory
LASSO	Least Absolute Shrinkage and Selection Operator
LGCM	Latent Growth Curve Modelling
LIRT	Longitudinal Item Response Theory
LME	Linear Mixed Effects
LMIRT	Longitudinal Multidimensional Item Response Theory
MAE	Mean Absolute Error
MAR	Missing at Random
MCA	Multiple Correspondence Analysis
MEM	Mixed Effects Modelling
NLME	Non-Linear Mixed Effects
NV	Nurse Visit
PA	Physical Ability
PCA	Principal Components Analysis

PW	Psychological Well-being
QGC	Multilevel Quadratic Growth Curve
RE	Relative Error
RMISE	Root Mean Integrated Squared Error
RMSE	Root Mean Square Error
RMSEA	Root Mean Square Error of Approximation
SCAD	Smoothly Clipped Absolute Deviation
SEM	Structural Equation Modelling
SHARE	Survey of Health, Ageing and Retirement in Europe
SMISE	Sum of the Mean Integrated Squared Error
SW	Social Well-being
SWLS	Satisfaction with life Scale
TCGA	The Cancer Genome Atlas
TLI	Tucker-Lewis Index
WHO	World Health Organisation

Chapter 1

Introduction

In pharmacokinetics and pharmacodynamics, drug response data are collected over multiple doses and across multiple individual patients to understand the biochemical, physiological, and molecular effects of drugs on the human body, i.e., "what the drug does to the body"; as well as, drug absorption, distribution, metabolism, and excretion, i.e., "what the body does to the drug". In social sciences, frailty data are usually collected over time to study the mental and physical changes that old people experience over time. This repeated collection of independent subject responses and relevant covariates over multiple occasions, also known as longitudinal design, is common across many scientific fields and provides an essential tool to study change over time. Although longitudinal studies can provide a unique insight on real-world processes, development and lifespan issues which could not be studied otherwise, longitudinal data collection involves many challenges such as high cost, poor participant recruitment and selective attrition. Therefore, it is essential to develop robust statistical methods that can make good use of the available information and handle any arising problems.

Figure 1.1, illustrates an example of dose response measurements from the Genomics of Drug Sensitivity in Cancer (GDSC) data where multiple cancer cell lines were treated with five different BRAF targeted compounds to explore in vitro drug efficacy. This is a case of longitudinal data where the measurement points are not time but dose. Before analysis, typically such data are summarised into cross-sectional drug response data (e.g., IC_{50}) for each experimental unit (in this case that is each cancer cell line) to reveal drug response trends. However, for drugs with potentially severe cytotoxic effects these summary measures are inappropriate and focusing on the full dose response curve

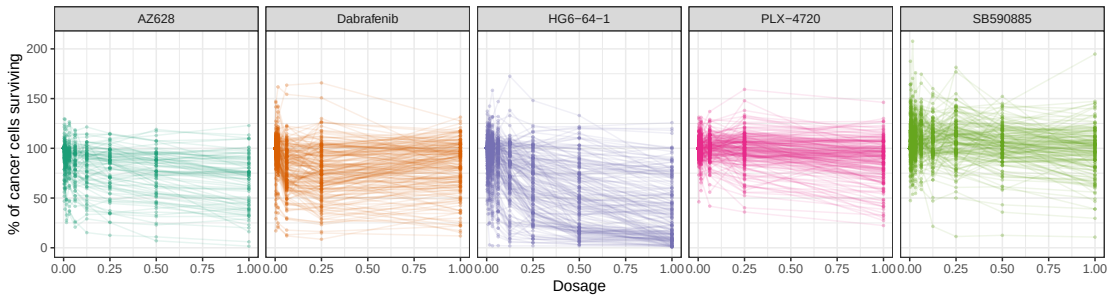


Figure 1.1: Dose response measurements after treating human cancer cell lines with five BRAF targeted compounds. Data were derived from the GDSC website (`https://www.cancerrxgene.org/`).

becomes essential. Non-linear mixed effects (NLME) models are often used to assess the dose response functional form and estimate drug efficacy over the varying drug doses whilst accounting for individual cancer cell line heterogeneity (Abbas-Aghababazadeh et al., 2019). Yet, when cancer cell line characteristics are expected to affect drug efficacy over different doses, NLME models are ineffective. That is because non-stationary trends and the dose varying effect of covariates on drug efficacy cannot be accounted for. In pharmacogenetics, more challenges are introduced since these covariates are often ultra-high dimensional; for instance, gene expression data.

As a tool to analyse longitudinal data, mixed effects models are a popular method for studying the change of biological processes over time. Figure 1.2, shows longitudinal frailty data collected through aggregating individual responses from the English Longitudinal Study of Ageing (ELSA) questionnaires. In this case, time can be either the data collection wave (Figure 1.2A) or the (true) process time which is the participants' age (Figure 1.2B). As shown in the plots, depending on the time measure chosen different statistical challenges occur. In the first case, analysis can be performed using mixed effects models. Missing data are, then, addressed through the model structure (Rogers et al., 2017) and model implementation is straightforward. This approach does not provide sufficient information on the dynamics of frailty over age and results' interpretation becomes challenging. On the other hand, the frailty data as shown in Figure 1.2B are irregularly collected with large missing data intervals occurring per each individual trajectory. Such data resembles sparse functional data which introduces further modelling challenges, for instance, increased model complexity and, sometimes, inference infeasibility (Yao et al., 2005; Kraus, 2015; Kneip and Liebl, 2020; Liebl and Rameseder, 2019). Notice that most of these methods focus on modelling the mean process trajectory without considering

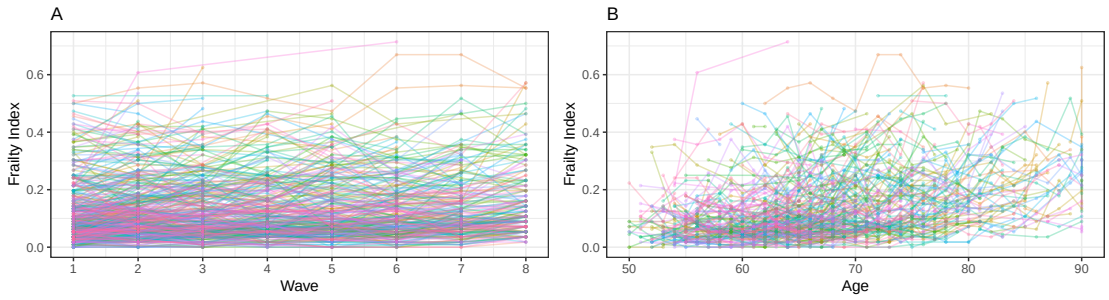


Figure 1.2: Frailty index longitudinal trajectories plotted over (data collection) wave (A) and age (B) for a random sample of 500 individuals. Frailty index data were produced using individual responses to the ELSA questionnaires.

the population dynamics and distributional changes over time.

In ageing studies, longitudinal data are usually collected using multiple choice questionnaires. Summary measures, such as the frailty index, can then be created by aggregating individual responses to groups of questions measuring specific traits. However, this method can overlook important information regarding individual trait change (Jabrayilov et al., 2016) and latent growth curve models along with item response theory models are often more appropriate (Wang and Nydick, 2020). However, the underlying assumptions of these models make it impossible to analyse irregularly collected multivariate longitudinal data where age is perceived as the time component. That is due to the large number of missing responses which makes estimation of the covariance matrix impossible. At the same time, these models often make strong assumptions about the functional form of the underlying trend of the examined process. Recent developments have extended the latent Gaussian process modelling framework to analyse intensive and irregularly collected longitudinal data (Chen and Zhang, 2020). This framework allows modelling multivariate longitudinal data with inherited flexibility into the functional form of the underlying process under study. If extended, this framework can provide the opportunity to model a single or multiple traits' change over time without making any strong assumptions for the underlying trends whilst addressing the problem of irregularly collected longitudinal data.

The aforementioned examples highlight the need for flexible modelling approaches which can improve the explanatory power while maintaining good model fit, results validity and prediction power. Non-parametric approaches promote statistical model flexibility and data driven inference. Motivated by two real-world problems, i.e., dose finding in cancer research and ageing, we develop three statistical modelling frameworks to

analyse complex longitudinal data while addressing the statistical challenges arising. Some of the tackled issues include model definition, multivariate and high-dimensional data handling, computational intensity of the developed algorithms and model implementation using highly complex data extracted from two large databases. Many of the available longitudinal data analysis tools have been rigorously systematised in Fitzmaurice et al. (2008), Diggle et al. (2002) and Hedeker and Gibbons (2006), however, the handled problems have not been extensively explored across the literature yet.

Outline of the Thesis and Contributions

In the current work, we focus on modelling complex longitudinal data collected from one experimental and one observational study. The main contributions of this work are summarised as follows. In the first study, we examine the dose varying effect of anticancer drugs to cancer cell line survival while adjusting for cancer cell line genetic characteristics. This study involves longitudinal data with non-linear trajectories over dose and dose varying covariate effects. The data come along with ultra-high dimensional gene expression data which need to be screened and whose effect need to be accounted for during inference. Statistical challenges tackled include ultra-high dimensional variable screening under a varying coefficient modelling framework, results interpretation and association to biological mechanisms. In the second study, we tackle the problem of irregularly collected continuous longitudinal data where the population under study is characterised by large heterogeneity. Multivariate questionnaire data are summarised in a single index to create frailty trajectories, a widely used surrogate of ageing. These data are then analysed using a dynamic conditional quantile regression framework while accounting for baseline participant covariates. In the third study, we go a step further by handling multivariate and irregularly collected ordinal longitudinal data collected from the same highly heterogeneous population. Here, we develop a model which handles longitudinal data with the aforementioned characteristics, but also allows the analysis of more than one underlying concepts measured through multiple questionnaires simultaneously.

This thesis is organised as follows. Chapter 2 aims to provide a better understanding of the background of the two main studies of the thesis, the cancer genomics project and the ageing project. In its first part, we briefly review the cancer research background which

motivated us to model the full dose response trajectories and explore dose-dependent associations to high-dimensional genetic data. The second part is dedicated to ageing. Specifically, we review the core studies on the ageing field which were determinative for our variable selection process and ageing domains' definition later in this work. Chapter 3 gives an overview and presents the key features of the selected longitudinal data sets. The exploratory analysis performed for the ELSA data using variables measuring all ageing domains as defined in ageing literature is the first large contribution of this thesis. In Chapter 4, we extend the methodology proposed by Chu et al. (2017) to the objective of assessing the transcriptomic effect on anti-cancer drug response, where coefficient functions are allowed to vary with dosage. This is the second contribution of this thesis. Chapter 5 adopts and extends the methodological framework of Dawson and Mueller (2018) to estimate the quantile dynamics and derive predictions for a common surrogate of ageing, frailty. The third contribution of this thesis is that by employing this methodology to model survey data and extending it to allow for time-invariant categorical and continuous covariate adjustment, we manage to address the problem of incomplete individual responses, get a comprehensive view on the distribution of frailty over time and predict frailty trajectories for population subgroups. In Chapter 6, we present the fourth contribution of the thesis. We build a bivariate latent Gaussian process model for multivariate longitudinal survey data of ordinal type. The proposed model allows for the description of changes in latent constructs measured via multiple choice questionnaires in longitudinal surveys. Finally, we conclude this work and propose directions for further extensions in Chapter 7.

How to Read this Thesis

This thesis follows the traditional Single Volume Format of Lancaster University thesis submission regulations[1]. However, parts of this document constitute a series of contributions already published as a journal paper in the *PLoS Computational Biology* journal appearing in Koukouli et al. (2021), or submitted for publication. Chapters 2 and 3 function as accompanying background chapters for the methods presented in Chapters 4, 5 and 6. Readers will notice that in both Chapters 2 and 3, we dedicate a longer part

[1] `https://www.lancaster.ac.uk/media/lancaster-university/content-assets/documents/student-based-services/asq/marp/PGR-Regs.pdf`

to ageing and the ELSA data exploration. That is due to the larger complexity of the ageing concept which was key for the evolution of this research and the larger amount of work needed to explore the available data from this study. Overall, approximately two thirds of this thesis focus on ageing and one third is focused on cancer research.

Since the main methodological chapters of the current document were developed as journal papers and handle different statistical challenges while addressing different types of complex longitudinal data, each chapter has its own introduction, literature review and discussion section. A summary section at the end of each chapter functions as a link-up for the research presented in adjacent chapters.

List of Publications

Some parts of the contributions in this thesis have been presented in the following journal publication and a submitted paper.

[i] Koukouli, E., Wang, D., Dondelinger, F. & Park, J. (2021). A regularized functional regression model enabling transcriptome-wide dosage-dependent association study of cancer drug response. PLoS computational biology, 17(1).

[ii] Koukouli, E., Park, J., Titman, A., & Doebler, S. (2022). Dynamic modelling and prediction of frailty in ageing English adults.

In all the aforementioned work, Koukouli had the main responsibility in developing the methods, developing the code for model implementation, performing the data pre-processing and analysis, and writing each paper. Revisions of the writing were incorporated by the co-authors directly or by Koukouli after discussions with the other authors. For publication [i], Wang helped with the biological interpretation of study findings. All projects were supervised by Juhyun Park. The first publication was also supervised by Frank Dondelinger, whereas [ii] (corresponds to Chapter 5) and Chapter 6, were co-supervised by Andrew Titman and Stefanie Doebler.

Software availability

All analyses have been conducted using R version 3.6.3. Simulation studies of Chapters 4, 5 and 6 were performed using The High End Computing Cluster at Lancaster University. Model implementation of Chapters 4 and 6 were also performed using The High End

Computing Cluster at Lancaster University. Code for applying the two-stage variable selection algorithm of Chapter 4 is available online as an R package at `https://github.com/koukoulEv/fbioSelect`.

Chapter 2

Background

2.1 Introduction

In Chapter 1, we discussed applications of longitudinal data analysis focusing on the methodological challenges arising due to the data generation process in each of the presented scientific fields. In this chapter, we focus on reviewing the literature in cancer pharmacogenetics and ageing to highlight current statistical needs and challenges due to the nature of the problems handled within each scientific field. The following section outlines some key concepts and highlights major developments in cancer pharmacogenetics which motivated us to build the functional regression model of Chapter 4. By discussing limitations in the previous literature, we emphasise potential research gaps and directions for future advancements. Next, in Section 2.3, we review the most extensively studied ageing concepts and their theoretical and operational definitions. Specifically, in Section 2.3.1, we outline four key ageing concepts (frailty, successful ageing, active ageing and healthy ageing) and discuss their conceptualisation and literature definitions, whereas in Section 2.3.2, we provide an overview of the measurement tools used to aid ageing quantification. In Section 2.4, we explain the value of the identified gaps and presented concepts to the current study; and, finally, in Section 2.5, we summarise the main scientific conclusions which motivated and contributed to the remaining of this thesis.

2.2 Advancements in Cancer Pharmacogenetics

Following cardiovascular diseases, cancer is the second leading cause of morbidity and mortality globally, with lung cancer being the primary cause of cancer death (Siegel et al., 2022). Cancer is a heterogeneous disease and individual tumours sometimes show very different mutational and molecular profiles. This is due to the various aberrations taking place at a cellular and molecular level previous to cancer development (Zhang et al., 2012). This uniquely complex genetic makeup of a tumour determines how it reacts to a given anti-cancer drug. However, due to lack of predictive markers of tumour response, often patients with very different tumour genetic makeup will receive the same therapy, resulting in high rates of treatment failure (Chang et al., 2003).

Large clinical trials in rapidly lethal diseases are expensive, complex and often lead to failure due to lack of efficacy at a given dosage (Cook et al., 2014). One major issue for some cancer treatments, e.g., chemotherapies, are cytotoxic effects that result in the collateral damage of the healthy host tissue (Corrie, 2008). Patient remission depends not only on the selection of the right drug but also on the determination of the optimal dosage, especially when drugs with a narrow therapeutic index, high toxicity levels or both are administered. Research has shown that genetic factors can help fine-tune the dosage for individual patients so that the minimal effective dosage can be delivered (Relling and Dervieux, 2001). Pharmacogenetics of anti-cancer drugs focuses on exactly this purpose: optimising cancer treatment to limit adverse effects while maintaining efficacy (Quaranta and Thomas, 2017). Below, we review key concepts in cancer pharmacogenetics literature and point out key findings within this field.

2.2.1 Cancer Cell Line Drug Screening

Treatment response in patients with specific cancers has been intensely examined in relation to the molecular characteristics of the tumours (Zhang et al., 2013). However, cellular heterogeneity within the tumour and the lack of standard metrics for quantifying drug response in patients can make it difficult to computationally model response as a function of molecular features. Cancer cell line drug screens can provide valuable information about the effect of genetic features on drug dose response and facilitate detection of predictive biomarkers at a preclinical stage in a controlled setting. Cancer

cell lines are cancer cell cultures that are kept under certain conditions in a laboratory to divide and grow over time. These are often treated with selected anti-cancer compounds and screened to monitor changes in cancer cell proliferation and survival. Following the advancements of high-throughput large scale cancer genomic technologies, there have been several data collection initiatives aiming to examine pharmacogenetics and pharmacogenomics relationships (Barretina et al., 2012; Yang et al., 2012; Lamb et al., 2006; Hyman et al., 2017). The constructed databases contain information from a wide range of common and rare types of human cancer cells isolated from affected tissues, grown in vitro and treated with a large number of anti-cancer inhibitors. The Cancer Cell Line Encyclopedia [CCLE; Barretina et al. (2012)], the Connectivity Map [CMAP; Lamb et al. (2006)] and the GDSC (Yang et al., 2012) are within the largest public resources for information.

2.2.2 Detection of Pharmacogenetics Interactions

Reviews on recent developments in genome-driven oncology highlight the additive value of integrating molecular information from databases in *in vitro* anti-cancer drug response prediction (Hyman et al., 2017; Ben-David et al., 2018; Vamathevan et al., 2019; Tavassoly et al., 2019). During the last decade, there have been several systematic studies that examined the relationship between genetic variants and drug response in cell lines. An example is the work of Iorio et al. (2016) where the investigators have associated cancer-driven alterations to anti-cancer drug sensitivity using cancer cell line data from The Cancer Genome Atlas [TCGA; Tomczak et al. (2015)] and the International Cancer Genome Consortium [ICGC; Zhang et al. (2011)]. In their work, they have identified hundreds of pharmacogenomics interactions, explained variation in drug response according to various molecular data types, such as gene expression data, whole-exome sequencing data, DNA methylation data etc., and employed machine learning algorithms, including elastic-net regression and random forests, to reveal key genetic features and assess the power of the built models in predicting anti-cancer drug response. One of their key findings was that gene expression data had the highest power in predicting drug response after anti-cancer drug administration reinforcing previous assumptions about the association between genetic markers and anti-cancer drug efficacy (Ross et al., 2000). As a measure for drug response, they have utilised the half-maximal inhibitory

concentration, also known as IC_{50}, and so, they have overlooked the effect of genetic features on drug efficacy under various dose concentrations. Other studies, such as Ji et al. (2009) and Delpuech et al. (2016), measured transcriptional profiles and by comparing multiple genomic features of cell lines to drug response, they were able to identify gene signatures for drug responsiveness in specific cancer types. In these studies, dose response curves were summarised using the concentration effective in producing 50% of the maximal response, also known as EC_{50}.

While the aforementioned papers inform about existing signatures of drug response and provide a way towards selecting the right drug for a patient, none of them characterise gene-dose relationships that may ultimately identify the optimal dose for a drug to use in the clinic. That is because these signatures were selected based on a single summary statistic of dose response, i.e., the IC_{50} or the EC_{50}, which, under the current framework, indicate how much of an anti-cancer agent is required to reduce cell survival by half or reach a response between the baseline and the maximum after a specified exposure time respectively (Aykul and Martinez-Hackert, 2016). Clearly, these may not always be the most useful metrics for differentiating drugs since they only provide information on one dose concentration (Keshava et al., 2019).

Falcetta et al. (2013) and Silverbush et al. (2017) have explored changes in cancer cell proliferation and survival after administering various dosages of certain anti-cancer compounds. Their work highlights the potential benefits of exploring the entire dose response curve instead of only the half maximal inhibitory or effective concentration. However, as discussed in the previous chapter, looking at the entire dose response curve introduces further challenges to the already complex problem of high-dimensional feature screening. To the best of our knowledge, there are no studies exploring dose-dependent associations between gene expression and drug response so far which is a fundamental gap to the current literature.

2.3 The Theoretical Foundations of Ageing

Due to the recent advancements in medicine, there has been a dramatic change in world population, even in the low and middle-income countries (World Health Organization, 2017). In fact, the global population aged 65 years or over has almost tripled since

1980 and is expected to double by 2050, reaching nearly 2.1 billion (United Nations, 2021). Nowadays, according to the World bank, the average global life expectancy is 72.6 years (`https://data.worldbank.org/indicator/SP.DYN.LE00.IN`), whereas the global birth rate experiences a steady drop with the current rate being around 18.5 births per 1,000 total population globally. In most western countries particularly, life expectancy exceeds 80 years. Soon the old population will outnumber the young, leading to a demographic shift with tremendous societal implications (Christensen et al., 2009).

These additional life years are not always expected to be healthy years (World Health Organization, 2017). Even though frailty and health decay are both normal parts of the ageing process, healthy behaviours and habits lead to healthier old years (Li et al., 2020). Yet, old adults often neglect them leading to higher potential for experiencing disease and disability. Simultaneously, health systems are often not aligned with the older adults' needs and, in many countries, environments are not adjusted to serve the old populations. Although efforts to adopt a positive attitude towards ageing strengthen, old age stereotypes, prejudicial views and discrimination prevail. Initiatives to tackle the inflation of the population aged 60 years or over and prevent negative experiences at old age are needed in order to sustain society's well-being and people's quality of life. To do so, it is essential to understand the dynamics of the ageing process and the relationships between the factors that affect ageing progression.

Since the second half of the last century, scientists have examined to what extent disease and disability at old age are age-associated and the factors promoting positive ageing experiences (Kusumastuti et al., 2016). Frailty, healthy ageing and successful ageing are only a few of the concepts used across the ageing literature to study and reveal the ageing dynamics (Rowe and Kahn, 1987; Searle et al., 2008; Michel and Sadana, 2017; Stephen et al., 2014; Marshall et al., 2015; McPhee et al., 2016; Gale et al., 2014).

2.3.1 Towards a Universal (Healthy) Ageing Definition

The word "ageing" is used to describe a person or thing that is getting old (Hornby, 2005) and by "old" we refer to anyone or anything that have lived or existed for many years (Hornby, 2005). Attempts aiming to find a precise and universal definition for the ageing process have come from various scientific fields such as evolutionary biology (Rose, 1991), gerontology (Strehler, 1962; Rowe and Kahn, 1987) etc. Bernard Strehler describes

ageing as "the changes which occur in the post-reproductive phase of life that result from a decrease in the ability of the body to maintain homeostasis, i.e., to regulate the functions of the body within the very precise limits required for efficient functioning and survival" in his book *Time, Cells, and Aging* published in 1962. In his work, the ageing process is perceived as the physiological changes that take place over a period of human body's gradual "collapse". Loss of body's ability to maintain health and well-being is highlighted as an indicator of getting older. However, the underlying reasons triggering this degradation are not made clear.

In 1991, an evolutionary biologist, Michael R. Rose, described ageing as the constant degeneration of health due to gradual failing of internal physiological components of the human body (Rose, 1991). This particular definition, even though successful in capturing the nature of the changes taking place while individuals are getting older, lacks the ability of pointing out any potential causal associations between internal or external factors affecting the human body and the ageing mechanisms. To fill this gap, a more recent view coming from the same author refers to the process of getting old as "a decline or loss of adaptation with increasing age caused by a time-progressive decline of Hamilton's forces of natural selection" (Rose et al., 2012). In this case, the reasons associated with this loss of ability to preserve body's health after a particular time point are associated to the forces of natural selection acting on individuals' survival and fecundity (Rose, 1991). From a biological perspective, the ageing process is, thus, viewed as an age-progressive decrease in survival and reproductive rate. This degradation, even though constant, it is believed not to be steady and, although assumed as moving towards zero level (death), there is evidence for a natural selection plateau which leads to plateaus in individual physiological state, reproduction and mortality declines (Flatt, 2012).

In this section, we review recent definitions of ageing concepts coming from gerontology, sociology and psychology. Even though, researchers in these fields recognise that negative health outcomes might be associated to advanced age, they believe that pathology is not an age-related disease and can be avoided or even reversed by adopting and adjusting individual behaviours and lifestyle. In fact, from the late 1980s, the distinction between non-pathological and pathological ageing was further enriched by a third ageing state: the one that is "successful" (Rowe and Kahn, 1987). Thereafter, positive ageing attitudes became popular and research focused on understanding why and how "successful" ageing

is achieved. The following sections discuss the theoretical foundations behind the concepts of frailty, successful, active and healthy ageing.

Ageing in Terms of Deficits Accumulation

In the biological literature ageing is often viewed as the natural process of cellular and functional senescence (Johnson et al., 1999). Recognising the heterogeneity of the ageing process, research focused on the factors that accelerate this senescence and cause loss of physiologic function and autonomy at old age. This promoted the development of a deficits-model based on which policy changes, such as early exit from the labour force, are encouraged (Boudiny, 2013). The age-associated risk of experiencing multi-system physiological changes leading to often (not always) negative health outcomes due to functionality and autonomy loss is defined in the literature as "frailty" (Hogan, 2003; Fisher, 2005). The last two decades, the concept of frailty is used as a tool to identify vulnerable individuals with increased risk of experiencing negative health outcomes or death. In other words, it is a tool to distinguish between pathological and non-pathological ageing. Frailty is conceptualised in relation to deficits accumulation in health, and by deficits we refer to symptoms or signs of poor health and decreased physical or mental functionality, disabilities and diseases (Searle et al., 2008; Rockwood and Mitnitski, 2007). However, without questioning the value of the concept of frailty in ageing studies, we can argue that even though frailty predicts mortality and successfully captures the process of senescence as people are getting old, it does not capture the positive aspects of ageing and does not encourage research on the factors that promote better old age experiences; it rather supports research only on those that protect individuals against life-threatening outcomes.

Successful Ageing

During the late 80s, research moved its focus from ageing "bad" to ageing "well" (or "successful") aiming to propose strategies that can promote living disease and disability-free lives (Rowe and Kahn, 1987; Strawbridge et al., 2002). Since then, numerous definitions and models have been proposed aiming to achieve a better understanding of what successful ageing really is and what are the factors that contribute to this "success". As shown in the work of Kusumastuti et al. (2016) and Bülow and Söderqvist (2014),

successful ageing definitions can be grouped into two main categories: the definitions allowing for the inclusion of older people opinions and those which solely consider scientists' views. Both the Katz and Havighurst clusters [as named by Kusumastuti et al. (2016)] can be divided into further sub-clusters. The variety in the first (lay definitions) is often observed due to different cultural and psychosocial backgrounds (Bülow and Söderqvist, 2014). Similarly, expert definitions often vary in different scientific fields due to different assessment tools and prioritization of study objectives. We can distinguish two multidimensional definitions of successful ageing-the biomedical and the psychosocial-and three unidimensional−the psychosocial, cognitive and biological− (Lupien and Wan, 2004).

The term successful ageing was first introduced by Rowe and Kahn (1987). In their work *Successful Aging* published in 1997, they distinguish normal ageing into two subcategories, the "successful" and "usual" ageing. Individuals are characterised as successful agers if they experience minimal physiologic loss, i.e., they do not have any disabilities or disease, they remain engaged with life and they maintain their cognitive and physical functioning (Rowe and Kahn, 1997). This definition became very popular and was used both in the future ageing research and policy making (Foster and Walker, 2015).

In 1990, Baltes and Baltes introduced the meta-model of selective optimization and compensation which views ageing as a process of adaptation, meaning that under this theory successful agers are those who are resilient and can adapt their behaviours, feelings and thinking to reach their goals in life. This is a psychosocial process-oriented approach which focuses on the gains and losses experienced at old age and the "tools" people use to obtain their desired goals (Baltes and Carstensen, 1996). Rowe and Kahn's and Baltes and Baltes models complement each other and perceive success in completely distinct ways: the first, interpret success from a biomedical perspective highlighting the importance of maintaining good health and high functioning in order to age "successfully", while, the second, argues that the successful adaptation of behaviour with regards to someone personal goals is the most important tool towards this success.

Other successful ageing models distinguish successful agers based on their cognitive performance achievements (cognitive models); life satisfaction and social network (psychosocial models); or, individual morbidity and longevity (biological models). Specif-

ically, reviews on how scientists define successful ageing show that its definitions are achieved through the direct use or combinations of the following concepts: life satisfaction, longevity, freedom from disability and illness, personal growth, active engagement with life, maintenance of social networks/support/participation, productive engagement with social activities, autonomy, maintenance of physical and cognitive functioning, mental and psychological health (Phelan and Larson, 2002; Lupien and Wan, 2004; Bowling and Dieppe, 2005; Depp and Jeste, 2006; Depp et al., 2007; Cosco et al., 2014; Martin et al., 2015; Bowling, 2007).

As opposed to scientific views, views on successful ageing coming from the older people themselves value more accomplishments in life, having good finances and living environment, maintaining physical appearance and function, cognitive function and spirituality (Faber et al., 2001; Bowling and Dieppe, 2005; Bowling, 2007; Lupien and Wan, 2004; Depp and Jeste, 2006; Phelan and Larson, 2002; Cosco et al., 2014; Kusumastuti et al., 2016). Staying active and autonomous is placed among the most important things, but, at the same time, having disabilities or disease is not synonymous to not ageing "successfully" (Faber et al., 2001; Glass, 2003). However, the cultural and societal background of the population interviewed affect the most how individuals perceive "success" at old age (Hung et al., 2010; Willcox et al., 2007).

The above views and definitions focus on the present health state of people age over 60 years old, the life-course approach of Schulz and Heckhausen (1996), however, presented a wider view of successful ageing and highlighted the importance of taking actions over the whole individual lifespan to achieve successful ageing. According to Schulz and Heckhausen (1996), a successful life-course inevitably leads to successful ageing if physical, cognitive, intellectual, affective and creative functioning are maintained throughout the person's life, along with their social relations for which the genetic and socio-cultural background play an important role. For a more comprehensive review, we refer interested readers to Bülow and Söderqvist (2014).

Due to the dissimilarities between definitions of successful ageing, its operationalisation often becomes infeasible. Different studies encourage different health-promoting practices, resource allocation and ageing strategies (Martin et al., 2015) which complicate the creation and adoption of an international plan of action to promote "success" at old age. Expert views on successful ageing are often deterministic implying that if disease

and disability occurs then becoming a "successful" ager is impossible. It is not hard to see how the above encourages ageism instead of promoting positive attitudes towards ageing which are essential to boost old people's confidence and overall health (Lupien and Wan, 2004). At the same time, flexible models built based on subjective views may lead to misleading representation of the notion (Baltes and Baltes, 1990). Adopting a multidimensional approach is essential to create a more comprehensive picture of what "successful" ageing is, however, the appropriateness of the word "successful" itself raise additional discussions (Baltes and Baltes, 1990; Bowling, 1993; Boudiny and Mortelmans, 2011).

Active Ageing

The term successful ageing was criticised as being unrealistic suggesting an ideal state of being at old age which is difficult to reach (Bowling and Dieppe, 2005; Martinson and Berridge, 2015). In fact, the model of Rowe and Kahn underestimates disease occurrence in old individuals and labels them as being unsuccessful if illness and disability occurs. That is restrictive and, as shown in the previous section, does not coincide with subjective views of successful ageing which support the idea that even if disabled or diseased an old adult can still be a successful ager if he/she stays active and socially engaged.

The concept of "active" ageing was officially defined by the World Health Organisation (WHO) in 2002, as "the process of optimizing opportunities for health, participation and security to enhance quality of life as people age". Opinions on what constitutes active ageing vary and, according to Boudiny (2013), they can be grouped into three categories: the unidimensional view where active ageing implies prolonging employment years and participation to physical activities; the multidimensional view, where active ageing become synonym to the continuous participation of older adults in several domains of life including leisure and social activities; and finally, the group of views which equate active ageing with good health, great economic circumstances, autonomy and continuing physical, psychological and social function. Even though, some authors preferred the term "active" over the term "successful", a variety of opinions was expressed regarding its definition and suitability; especially due to the fact that the word "active" can solely point to physical activity, productivity and occupation.

The concept of "active ageing" was built to promote autonomy and participation at

old age introducing a broader and less idealized ageing view (World Health Organization, 2002). The variety of active ageing definitions highlight the existing limitations around its conceptualisation. First, ongoing debate exists on whether this concept neglects the importance of maintaining high mental capacity by turning the spotlight solely on physical activity and employment overidealising the productive model of old age (Foster and Walker, 2015). Second, large discrepancies exist between lay and scientific views, with the latter setting overambitious standards failing to capture the heterogeneity of the old population (Boudiny and Mortelmans, 2011). Similar to subjective views on successful ageing, the older people themselves equate staying active with staying engaged with life, even if disability and illness exist. Overall, scientific opinions often support the use of the term "active ageing" over successful ageing (Foster and Walker, 2015), however, we observed common discussions on how the word "active" can mislead policy making and prioritize actions that are not always supportive to the group of interest (Boudiny, 2013; De São José et al., 2017).

Healthy Ageing and the HAP

Even though life is synonymous with ageing, disease and disability are not (Franco et al., 2009). Since the second half of the 20th century, research focus has shifted towards the positive aspects of ageing and the actions that individuals and countries can take to encourage a better ageing experience. In the previous sections, we briefly presented the key historical and conceptual background behind the notions "successful" and "active" ageing. In 2015 and after many years of research, the WHO formally introduced and defined the notion "healthy" ageing. However, this is not the first time this term was mentioned into the gerontological literature. In this section, we discuss the historical and conceptual background behind the notion "healthy ageing"; the definition and conceptual framework built by the WHO; how this is linked to previous research and the Healthy Ageing Phenotype (HAP) concept; and what are the current limitations and challenges around this concept.

Apart from the terms "successful" and "active", a wealth of similar terms can be found in the gerontological literature as an effort to conceptualize quality of life at advanced age. Ageing "well", "robust", "competent", "vital", "positive" and "productive" ageing are some of them (Strawbridge et al., 2002; Garfein and Herzog, 1995; Brehm, 1987; Bowling,

1993; Kerschner and Pegues, 1998; Caprara et al., 2013). Motivated by the variety of opinions and vagueness around these concepts, Peel et al. (2004) refers to "healthy ageing" and defines it as "the lifelong process optimizing opportunities for improving and preserving health and physical, social and mental wellness, independence, quality of life and enhancing successful life-course transitions". In their work, they highlighted the multidimensionality of this concept, proposed domains that could potentially describe it and explained how cultural and even chronological age differences can play a tremendous role in its interpretation.

From 2004, "healthy ageing" was often used interchangeably with "successful" and "active" ageing (Depp et al., 2007; Zaidi et al., 2017; Bousquet et al., 2015) but efforts to further clarify it strengthened. Five years later, Franco et al. (2009) presented the idea of the HAP as a fundamental concept in ageing research, which was defined as "the condition of being alive, while having highly preserved functioning metabolic, hormonal and neuro-endocrine control systems at the organ, tissue and molecular levels". The HAP was proposed not only as a means to clearly define the notion of healthy ageing but also to measure it. Five domains were identified as the building blocks of HAP−psychological and social well-being, metabolic health, cognitive function, and physical capability−highlighting its multidimensional nature (Lara et al., 2013b).

The effort to build a universal definition for "healthy ageing" was concluded in 2015 by the WHO which describes it as "the process of developing and maintaining the functional ability that enables well-being in older age". In other words, healthy ageing is defined as the multidimensional process consisting of the adaptive mechanisms that take place during the life-course to preserve functional ability, the ultimate goal of healthy ageing. Functional ability is the capacity to achieve what people have reason to value (e.g., maintaining their role in society, relationships, enjoyment, autonomy, security, and personal growth) and it is built through a combination of individual intrinsic capacity (genetic, personal−for example, gender, ethnicity, educational level, occupation etc.−and health characteristics−physiological risk factors, diseases, injuries, and other geriatric syndromes−) and environment interactions at a particular point in time (World Health Organization, 2017).

While the HAP concept focuses on the present, the definition given by the WHO, it perceives ageing as a lifelong process which allows the subject to experience the HAP

state. Linking it to the life-course approach towards successful ageing (Kuh et al., 2014; Bülow and Söderqvist, 2014), it highlights the importance of adopting action plans to support the development of intrinsic capacity early in life and the maintenance of functional ability at later life stages. The domains of functional ability presented by the WHO can be linked to the HAP domains, showing that both constructs are built under the same fundamental ideas. Specifically, maintaining the abilities to move around, build and maintain relationships, meet someone's own basic needs, learn, grow, make decisions, and, contribute is synonymous with maintaining someone's physical capability, cognitive function, social and psychological well-being and metabolic health. Defining and understanding ageing and, more specifically, "healthy ageing" is essential to overcome some of the challenges brought by the steep increase of the old population. However, these challenges are different in nature, e.g., challenges to the pensions systems and social welfare costs (macro-level) to challenges that the older people face during their everyday life (micro-level). Due to the materialistic interpretations associated with the terms "successful" and "active" ageing, the term "healthy ageing" became the gold standard during the last few years. Healthy ageing research aims to identify how old generations could be supported in becoming resilient and how old age experiences can be improved. Recently, the WHO, Member States and Partners for Sustainable Development Goals initiated a Global Strategy and Action Plan for Ageing and Health for 2016–2020 with five objectives: act to support healthy ageing both nationally and internationally; develop age-friendly environments to promote autonomy at old age; align healthcare systems to ensure they meet older adults' needs; preserve and develop long-term care systems for people of advanced age; and, promote research on healthy ageing (World Health Organization, 2017). The ultimate aim of this action was to establish evidence and partnerships to support the Decade of Healthy Ageing 2020-2030 program which focuses on: supporting national plans and actions promoting Healthy Ageing; creating an international database hosting research advances and evidence on Healthy Ageing; promoting research and data collection on Healthy Ageing; aligning health systems and developing long-term care systems; fighting ageism (discrimination against old people); supporting human resources to develop robust care systems; examining the economic impact of investment in older populations; and, promoting age-friendly cities and communities (Rudnicka et al., 2020). This plan of action is of high significance since it establishes global guidelines and

directions towards creating age-friendly communities. However, the degree of individual countries' adaptation remains uncertain and further research is needed to assess the effectiveness of the proposed plans across the seven continents.

2.3.2 Ageing Operationalisation

Even though a universal definition of "healthy ageing" has been established, a single measure to quantify it is yet under investigation (Lara et al., 2013b; Tampubolon, 2016b; Cosco et al., 2014; Bousquet et al., 2015; Michel and Sadana, 2017; Zaidi et al., 2017; Sanchez-Niubo et al., 2020). More so, operationalisation of the ageing process and individual classification can be quite complex. To recommend a representative ageing metric, the following questions should be answered:

- Is the researcher's intention to quantify ageing through measuring health "deterioration" (increasing over time) or health "sustainability" (decreasing over time)?
- What kind of information needs to be included to measure either of the two?
- What kind of metrics need to be employed to objectively and universally quantify concepts like cognitive function, social skills or mental health?
- What should the cut-off points of an ageing score be to classify individuals to different ageing categories?

These questions show the complexity surrounding ageing quantification. Nevertheless, even if the above objectives are clear, it is essential to validate any ageing metric and confirm that the given ageing representation is of high standard. The biological and conceptual complexity and heterogeneity create challenges around ageing quantification. Since there is not a unique way to understand it and conceptualise it, people tend to use surrogates such as mortality or lifespan. In the ageing literature, there are two most widely used measures of ageing, the first is a conceptualisation of frailty (refers to mortality and deficits accumulation (Searle et al., 2008)), and the second is a conceptualisation of healthy (or active or successful) ageing (refers to lifespan and positive old age experiences). Below we review how ageing operationalisation and measurement are achieved in the current pulished research.

The Frailty Index (FI)

Frailty is a state of increased vulnerability to ageing-associated adverse outcomes and is considered to be a monotonically increasing process as people are getting older (Searle et al., 2008). It is one of the most popular proxy measures of ageing (Rockwood and Howlett, 2018; Walston and Bandeen-Roche, 2015) and, as many studies have found, is a better predictor of survival compared to chronological age (Searle et al., 2008; Rockwood et al., 2007; Hao et al., 2018; Feng et al., 2017). Even though there is not a single universal definition of frailty, the most common frailty metric is the Frailty Index (FI) calculated by accumulating age-related health deficits (Searle et al., 2008; Mitnitski et al., 2001). These consist of health problems including hearing loss and poor eyesight; cognitive and psychological problems; as well as, mobility difficulties. Other frailty metrics include the modified FI, the risk analysis index, the Ganapathi indices, the electronic FI etc. and they are all used across social and medical sciences as a surrogate of ageing or illness severeness to predict morbidity and mortality (Shashikumar et al., 2020; Esses et al., 2018; Clegg et al., 2016). One of the main drawbacks of FI is its reliance on the variables used for its construction making it unsuitable for comparisons across different ageing studies. For instance, in Searle et al. (2008) the FI is built using data from the Survey of Health, Ageing and Retirement in Europe (SHARE) and physical activity level is used as one of the deficits included into the index construct whereas in Rogers et al. (2017), the same variable (extracted from the ELSA data) is used as a predictor for the FI score. Nevertheless, due to its simplicity and its well established properties it remains a concept worth examined.

Quantifying "Healthy" Ageing

Cosco et al. (2014) showed that common successful ageing constructs employ physiological, well-being, engagement with life, personal resources and environment and finances measures. Zaidi et al. (2017) built an "active ageing" measure by taking the weighted sum of four domain scores (employment, participation in society, independent, healthy and secure living, and capacity and enabling environment for active ageing). In 2013, Lara et al. identified five features of the HAP and listed a number of instruments to measure them. The selected domains were: physical capability, physiological and metabolic health, cognitive function, social and psychological well-being. Valid question-

naires and measurement tools were proposed aiming to improve result agreement among international ageing studies. Aggregated scores and well-studied scales were some of the recommended tools which have been deployed in subsequent studies (Tampubolon, 2016b). Recently, Caballero et al. (2017) developed a measure to quantify health status over time using Item Response Theory (IRT). The proposed metric summarises questionnaire data using a multilevel model which deploys both consistently and not consistently measured items used for gerontological assessment promoting homogeneity and comparisons across international healthy ageing studies (Caballero et al., 2017; Sanchez-Niubo et al., 2020). Finally, Beard et al. (2019) developed an "intrinsic capacity" index using a bi-factor model; they used it as a surrogate for "healthy ageing" and studied its properties and validity.

Common features of the aforementioned constructs are the nature of the data used (information is selected using assessment questionnaires, tests and open-ended questions) and their multi-dimensionality. Additional similarities exist at a composition level, since the majority deploy aggregated scores and weighted sums. However, the heterogeneity observed in "healthy ageing" metrics results in large disparity between study outcomes. An example is the proportion of successful agers which ranges between <1% and >90% across different ageing studies (Peel et al., 2004; Cosco et al., 2014). This variety of outcomes is partially justified by the lack of a universal "healthy ageing" definition which was until recently ambiguous (Michel and Sadana, 2017). Other reasons include the use of different cut-off points, poor data quality and, cultural and societal differences. The simplicity and inherent properties of some of the above metrics are what make them compelling. However, they still fail to portray the multifacetedness of "healthy ageing" since the information used, even though multidimensional, is combined in such a way that the resulting constructs are most of the times a single index which does not allow to study each "healthy ageing" domain separately and the time-dependent changes and relationships between them.

2.4 Discussion

Through this review, we identified a lack of studies associating dose response to transcriptional profiles of cancer cell lines. Motivated by this gap, later in this thesis (Chapter

4), we adopt a functional regression approach that takes the dose response curves as inputs, and uses them to find biomarkers of drug response. One major advantage of this approach is that it describes how the effect of a biomarker on the drug response changes with the drug dosage, which is something currently missing from the available cancer pharmacogenetics literature and can promote better decision making by informing anticancer research on drug efficacy.

With regards to ageing research, during the last decades, gerontologists diverted their focus from the age-associated declines towards the positive aspects of ageing and how these can be promoted through policy. With the old population experiencing the steepest increase ever, societal changes to support old adults needs are necessary. To do so, the WHO introduced the concept of healthy ageing and initiated efforts to promote it internationally. Nevertheless, the topic of how healthy ageing can be measured and studied to identify healthy ageing predictors and perform country comparisons to assess the effect of these actions is still cloudy. To this end, ageing studies have been designed and conducted across the globe, to understand the ageing process, how life domains evolve over time and identify the factors that contribute to healthy and active ageing. To collect data on life domains linked to positive health outcomes at old age, modern longitudinal studies on ageing use a battery of questionnaires or tests. That is because, most of the targeted life domains cannot be measured directly, resulting in a huge collection of non-continuous data, collected, often, at irregular age points. Consequently, statistical analysis and inference based on these data become very demanding and advanced statistical methods should be employed to achieve meaningful conclusions. In the remainder of this work, we propose statistical methods aiming to address some of these issues by adopting flexible approaches to model the healthy ageing domains as conceptualised by Lara et al. (2013b) and their interrelationships. This will not only provide a more comprehensive picture of ageing progression through the adoption of this multidimensional approach, but will also ensure our study's alignment to the current WHO goals on ageing.

In Chapters 5 and 6, motivated by the data collected from the ELSA, we discuss methods that could facilitate analysis of complex longitudinal data, such as those gathered in modern ageing studies. By doing so, we not only develop novel statistical frameworks that could facilitate complex longitudinal data analysis but we, also, investigate an alternative healthy ageing operationalisation which is aligned to the international guidelines.

2.5 Summary

In this chapter, we reviewed key concepts and publications around cancer pharmacogenetics and ageing which are essential to follow the rest of this thesis. On one hand, large-scale assays of drug response in cancer cell line panels are an important part of the drug discovery and drug repositioning pipeline in oncology. Current state-of-the-art methods do not allow for detection of biomarkers that act on narrow effective dose ranges. On the other hand, old population increase creates numerous challenges in society. Just recently, a universal definition of healthy ageing was established by the WHO. The ultimate aim was to adopt actions to encourage a healthier, more active and socially engaged old population internationally. However, the tools to confirm that these actions are sufficient are not yet clear. Healthy ageing operationalisation is still under research and there is currently no well established tool to quantify it. In addition, healthy ageing measures are often data-dependent and unidimensional in practice, meaning that they cannot portray how healthy ageing domains interact over time. In the following chapter, we introduce and describe the data extracted from two large databases, the GDSC and the ELSA. These collections of experimental and observational data will be used next in Chapters 4, 5 and 6. These along with the theory presented in the current chapter motivated the methodological developments presented in this thesis.

Chapter 3

The data sets

3.1 Introduction

In the previous chapter, we introduced key theories and concepts around cancer pharmacogenetics and ageing. In the current chapter, we present the two databases which motivated our study, the GDSC data and the ELSA data. Even though both consist of repeated measurements data, they differ in many ways. First, the GDSC database contains experimental data whereas the ELSA data are observational. Second, GDSC data were produced in a controlled environment and so no missing responses occurred whereas the ELSA data include thousands of missing values due to drop-outs, participants' death, lost to follow-up, incapacity or unwillingness to respond, etc. Third, even though both data sets can be considered high-dimensional, they differ in terms of the nature of this increased dimensionality; the GDSC data are paired with high-dimensional genomic data whereas the ELSA data include more tha 5,000 measured outcomes per wave and participant. Below, we go through the key aspects of these two data sets, explore their structure and present exploratory analysis results which are fundamental for Chapters 4, 5 and 6.

3.2 The Genomics of Drug Sensitivity in Cancer (GDSC)

As mentioned in Chapter 2, since the early 2000s large databases have been built to facilitate identification of potential biomarkers and molecular targets for cancer therapy. The GDSC project is a Wellcome-funded collaboration between The Cancer Genome Project at the Wellcome Sanger Institute in the UK and the Centre for Molecular

Therapeutics of the Massachusetts General Hospital Cancer Centre in the USA (Yang et al., 2012). Its main objective was to identify biomarkers that characterise anti-cancer drug suitability for treating cancer patients. The current database includes drug sensitivity and molecular measures derived from more than a thousand types of human cancer cell lines used for the screening of 518 anticancer compounds. Data are publicly available at `https://www.cancerrxgene.org/`. In the following sections, we give a brief overview of the exact data extracted from the GDSC database which were analysed in Chapter 4 and we describe the normalisation procedure undertaken prior to model implementation.

3.2.1 The GDSC1 and GDSC2 data

Drug sensitivity data and molecular measures derived from 951 cancer cell lines used for the screening of 138 anticancer compounds were downloaded from the GDSC website (release 2, July 2012). For analysis purposes in Chapter 4, we specifically focused on cell lines of cancers of epithelial, mesenchymal and haematopoietic origin treated by five BRAF targeted inhibitors (PLX-4720, Dabrafenib, HG6-64-1, SB590885 and AZ628; GDSC1 data). The maximum screening concentration for each different drug was: 10.00 uM for PLX-4720 and Dabrafenib, 5.12 uM for HG6-64-1, 5.00 uM for SB590885 and 4.00 uM for AZ628. We, additionally, used the independently generated GDSC2 data set on drugs targeting the *MEK1*, *MEK2* genes (Trametinib, Selumetinib and PD0325901) and the PI3K/MTOR signalling pathway (Alpelisib, AMG-319 and AZD8186) for validation purposes. In Chapter 4, we expound further on this matter. The maximum screening concentration for the drugs included in GDSC2 data was: 1.00 uM for Trametinib; 10.00 uM for Selumetinib; 0.250 uM for PD0325901; 10.00 uM for Alpelisib; 10.00 uM for AMG-319; and, 10.00 uM for AZD8186.

In the original experiments, the drug sensitivity measurements were obtained via fluorescence-based cell viability assays 72 hours after drug administration (Yang et al., 2012). In GDSC1 data, approximately 66% of drug sensitivity responses were measured over nine dose concentrations (2-fold dilutions) and 34% were measured over five drug concentrations (4-fold dilutions). In total, we considered 3805 cancer cell line-drug combinations. For the remainder of this thesis we refer to these cancer cell line-drug combinations as experimental units. In Figure 3.1, we show the distribution of different tissues of origin treated with the drugs. Overall, similar proportion of cell lines have been

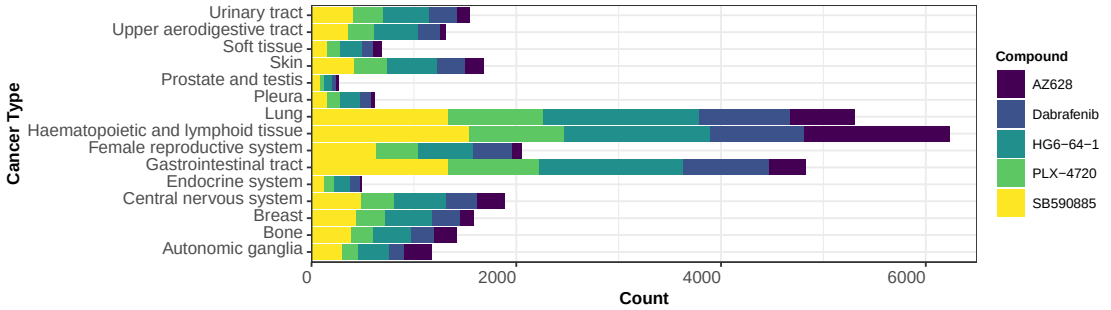


Figure 3.1: Distribution of tissue of origin across the five BRAF compounds used for cell line screening in the GDSC1 data.

treated with all of the compounds tested. A smaller number of cell lines were treated with AZ628, Dabrafenib and PLX-4720. A larger number of cell lines in the data set were originated from the lungs, the gastrointestinal tract and the haematopoietic and lymphoid tissues. Paired microarray gene expression data (17,737 genes) were available together with the dose response data.

3.2.2 Data normalisation

The dose response data also included a blank response for cells on the experimental plate that had not been seeded or treated with a drug. Blank responses have been used to adjust for the magnitude of the observation error while measuring the amount of cells in each plate. We used an affine transformation to the reported responses in order to normalise them within the drug concentration interval, 0 (0% of the maximum dosage) to 1 (100% of the maximum dosage). In particular, for the normalising procedure, we have used the formula:

$$NR_{ij} = \frac{R_{ij} - BR_i}{CR_i - BR_i} \tag{3.1}$$

where R_{ij} is the response of the ith experimental unit at the jth dosage level (i.e. amount of cancer cells after drug administration at the jth dosage), CR_i is the response under no drug administration (i.e. number of cancer cells at zero dose, $n_i = 1$), BR_i is the blank response of the ith experimental unit as described above and NR_{ij} is the new score taken from the transformation, $i = 1, \ldots, 3805, j = 1, \ldots, n_i$. After this procedure, data were analysed using the methods developed in Chapter 4.

3.3 The English Longitudinal Study of Ageing (ELSA)

Over the last two decades, there has been a huge increase in the number of longitudinal cohort studies studying ageing worldwide. Recent statistics from the Gateway to Global Ageing Data show that, globally, there are currently more than 15 longitudinal studies investigating trajectories of geriatric syndromes (Martin and Romero Ortuño, 2019). This is, in part, due to a concentrated effort at government level to optimise quality of life at old age and, if possible, prevent negative ageing outcomes through appropriate clinical practice and policy.

The ELSA is an ongoing biannual longitudinal study which collects information from people aged over 50 years old in England starting from 2002 (the total sample size exceeds 18,000 people). It is a National Institute of Ageing funded project with the additional support of a consortium of UK government departments coordinated by the Office for National Statistics, which aims to reveal age-related trends of the growing English population (Steptoe et al., 2013). It is designed to be directly comparable with other international ageing studies such as the US Health and Retirement Study in America (Juster and Suzman, 1995) and the SHARE (Börsch-Supan et al., 2013). Data collection started in 2001 as part of a pilot study (wave 0) but the main fieldwork was officially initiated in March 2002 (wave 1). The data are collected in two-yearly interviews (waves) and are currently publicly accessible through the UK Data Archive (`https://www.data-archive.ac.uk/`).

Answers to hundreds of questions have been encoded into the ELSA data set producing more than 5,000 variables per wave available for analysis. In the following sections, the reader can find a detailed overview of the data structure and content of the first eight ELSA waves. Data description is specifically focused on variables used in subsequent chapters, in particular, Chapters 5 and 6. To be more specific, in Sections 3.3.1 and 3.3.2, we give an overview of the study design and sample demographics. Section 3.3.3 presents more details about its content and data availability, whereas in Section 3.3.4, we discuss missing data and drop-outs. Sections 3.3.5 and 3.3.6 provide an in depth illustration of the data used for further analysis and their structure. Appropriate exploratory data analysis was performed at this stage which acts as a stepping stone for the analysis presented in Chapters 5 and 6.

3.3.1 Study design

The ELSA participants were drawn from households that had previously responded to the Health Survey for England (HSE) between 1998 and 2001 (Mindell et al., 2012). Due to death, loss to follow-up and participants ageing, since 2002 and until 2017, four sample refreshments took place in order to maintain balance between younger, middle and older age groups in the study (Figure 3.2). Individual and household information was collected using a Computer-Assisted Personal interview (CAPI), tests and a self-completion questionnaire. During the CAPI both core participants and their partners were interviewed. There are five main areas for which data were collected across the first eight waves of ELSA. These include demographics and economics data; health and disability measures; psychosocial measures; and, cognitive function measures. Besides, approximately every 4 years recruited nurses were visiting participants to carry out blood sample collection, physical examination and performance assays. One-off assessments such as life-history interviews (information about children and fertility, partnerships, education, employment, migration, accommodation, health events, childhood circumstances, adverse life effects), risk preferences (performance of study lotteries with the opportunity to win cash and willingness to postpone reward) and sexual activity and relationships questionnaires were additionally collected at waves 3, 5, 6 and 8 respectively. Alongside individual interview data, mortality records ("End of life" data) and data from a genome-wide association study genotyping 7,412 ELSA participants are available[1]. At this point it is important to mention that due to the evolving study objectives and data collection challenges, questionnaires tend to change over time with new questions/group of questions being added or substituting old ones, or old questions/group of questions not being measured or being measured in every future wave.

3.3.2 Sample demographic characteristics

Sample size across waves differs since the response, surviving rates and follow-up rates vary. Sample sizes fluctuate from ≈7,000 to ≈11,000 individuals, members of ≈6,000 to ≈8,000 English households (Table 3.1). Sample balance between males and females remains relatively constant across waves with more women than men participating (the

[1]All participants were of European descent and genotyping was performed at University College London Genomics in 2013-2014 using the Ilumina HumanOmni2.5 BeadChips (HumanOmni2.5-4v1, HumanOmni2.5-8v1.3).

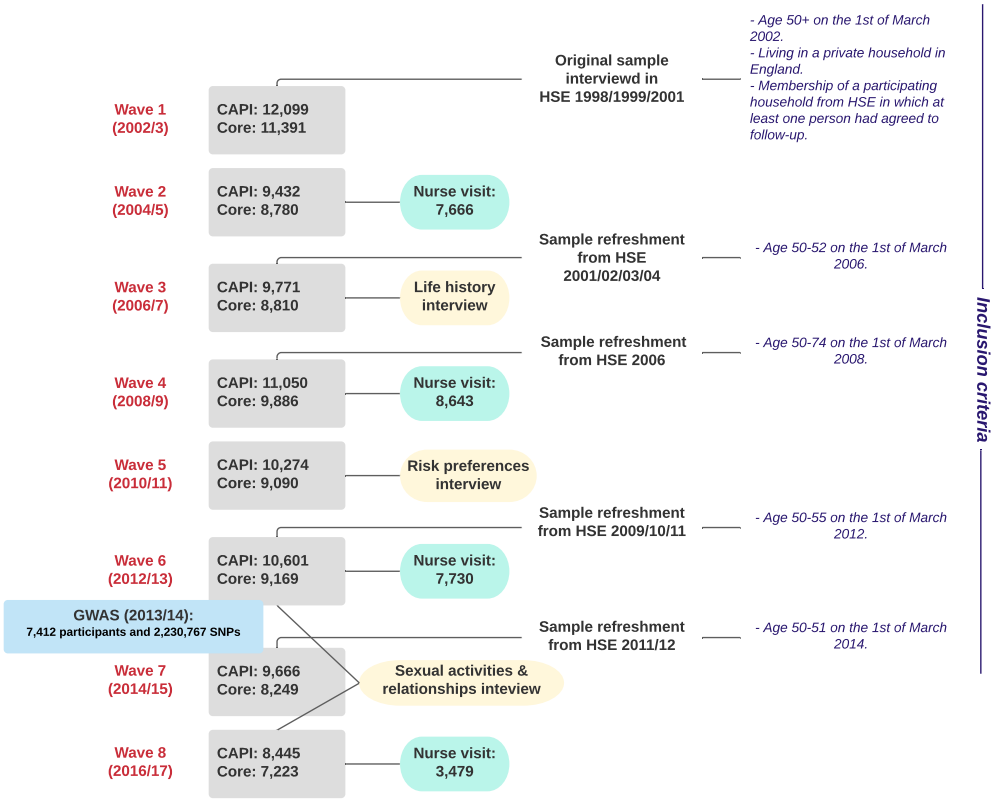


Figure 3.2: Overview of the data collection process in ELSA waves 1 to 8. Sample sizes are given for the complete study. CAPI stands for Computer Assisted Personal interview. HSE stands for the Health Survey for England from which the first and subsequent samples were collected.

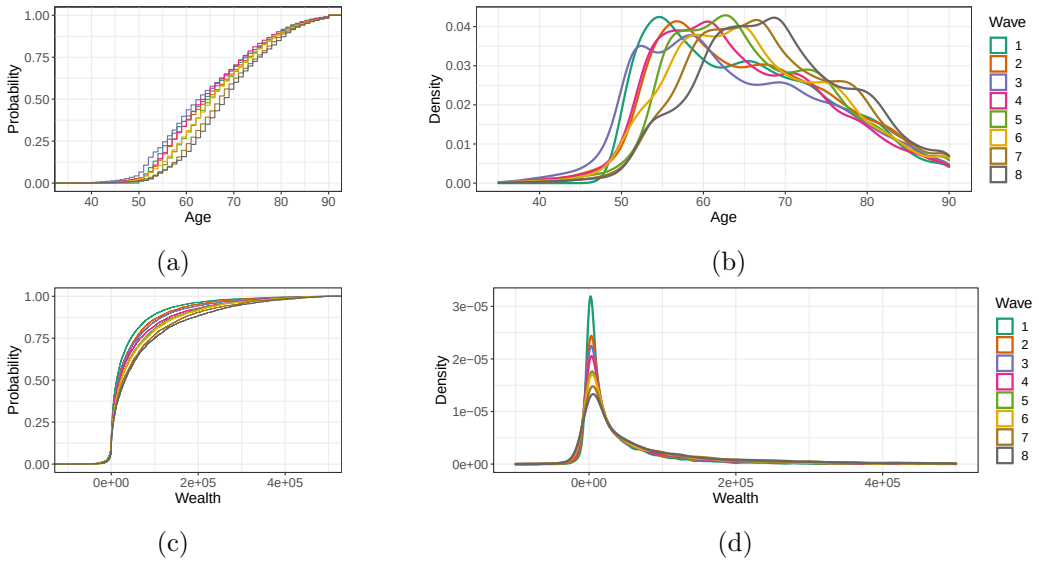
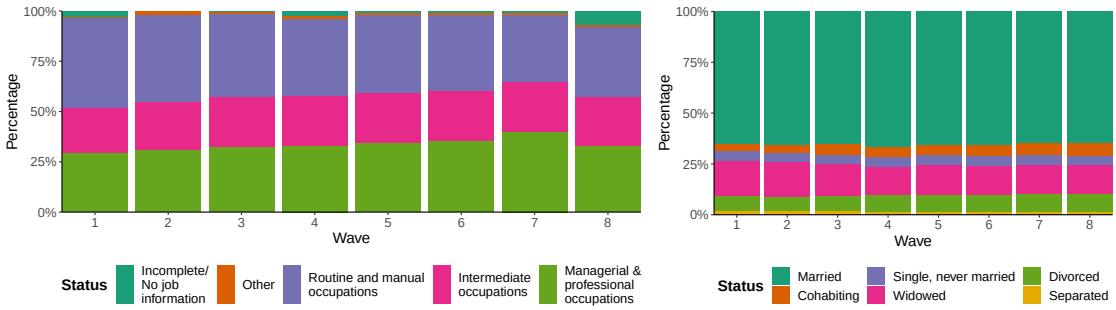


Figure 3.3: (a) Empirical cumulative distribution function for age per wave. The plot suggests that the age distribution in all waves is highly skewed and due to the differing number of participants and the sample refreshments the distribution of old people varies across waves. (b) Empirical age density for each different wave. (c) Empirical cumulative distribution function for wealth per wave. (d) Empirical wealth density for each different wave.

proportion of women across all waves varies between 54.51% and 56.40%; for men it fluctuates between 43.60% and 45.49%). This gives a representative illustration of how the real population of males and females is distributed with more women surviving for longer periods compared to men of the same age (Pinkhasov et al., 2010).

The age distribution is highly skewed towards left for all the ELSA waves with the vast majority of the sample aged between 50 to 75 years old (Figure 3.3a; skewness ranges between 0.14 and 0.45; kurtosis ranges between 2.23 and 2.62; difference between mean and median does not exceed 1.57). Even though ELSA design aims to obtain a representative sample across all different age groups, older population (more than 80 years old) remain under-represented in all waves due to possibly high rates of mortality or morbidity. In the data, over 90 years old respondents have been classified as 99 years old for confidentiality reasons. This is not depicted in Figure 3.3a since we classified all respondents over 90 years as 90 years old. For subsequent longitudinal analysis those subjects (over 90 years old) have been excluded from analysis.

Economic inequalities and population wealth distribution is well captured in the ELSA sample (Figure 3.3d and Table 3.1). The percentage of people in debt (negative wealth) is less than 10% at all waves. In particular, impoverished people percentages



(a) (b)

Figure 3.4: (a) Occupational status distribution across the eight ELSA waves (2002-2017). (b) Marital status distribution across the eight ELSA waves (2002-2017).

were: 9.18%, 8.87%, 10.00%, 9.30%, 8.61%, 8.59%, 7.39% and 7.31% for Waves 1 to 8 respectively. Between 2006 and 2010, a larger accumulation of people owing money is observed and economic inequalities become profound [interquartile range (IQR) of wealth varies between 44,600 and 110,312]. In contrast, social and educational inequalities are less apparent, especially as we move forward in time (Table 3.1 and Figure 3.4a). People being in managerial compared to manual occupations are steadily increasing over the years, whereas the number of people achieving higher educational qualifications follow a similar rise of almost 10% in 2019 compared to 2002 (1st Wave).

Most of the ELSA participants are married or widowed across all waves (Figure 3.4b). From 2002, there is an increase of 2% in divorced/separated people as well as a small decrease in married/cohabiting couples. Finally, individuals from non-white ethnic origin are underrepresented compared to white participants (less than 4% of the total sample size across all waves) which should be taken into consideration in future inference and conclusions.

3.3.3 Questionnaire content

Below, we give a brief overview of the topics covered in the ELSA questionnaires. We distinguish six key modules which are presented below.

Module 1: Health, disability and health behaviour measures

Physical and mental health have been assessed during the CAPI. Topics covered include: self-rated health; psychiatric, chronic and cardiovascular disease records (as diagnosed by a physician); long-standing illnesses; mobility impairment and physical functioning

	Wave 1	Wave 2	Wave 3	Wave 4	Wave 5	Wave 6	Wave 7	Wave 8
Individuals	11391	8780	8810	9886	9090	9169	8249	7223
Households	7912	6246	6463	7250	6687	6902	6009	5543
Proxy respondents	1.39%	1.05%	1.9%	2.97%	3.9%	4.18%	4.38%	3.97%
Gender								
Men	45.53%	44.98%	44.73%	44.77%	44.62%	44.42%	44.07%	44.05%
Women	54.47%	55.02%	55.27%	55.23%	55.38%	55.58%	55.93%	55.95%
Age								
50-59	36.57%	29.57%	34.57%	29.28%	21.65%	22.49%	14.85%	11.07%
60-69	29.83%	32.73%	29.48%	35.5%	38.62%	37.63%	39.65%	39.45%
70-79	22.51%	24.92%	23.14%	23.86%	26.56%	26.65%	30.39%	31.81%
80+	11.08%	12.77%	12.81%	11.35%	13.18%	13.22%	15.13%	17.67%
Wealth**								
Mean	44,654	55,618	62,688	75,201	75,114	89,942	93,738	106,964
Median	12,400	17,000	17,400	20,400	23,567	24,800	30,000	34,000
SD	117,263.3	163,403.1	172,106.5	241,836	176,692.2	319,484.7	239,704.2	237,567.3
Missing	1.76%	1.36%	2.68%	2.84%	2.40%	2.41%	2.67%	1.87%
Education								
No qualifications	56.18%	52.87%	43.48%	43.10%	43.48%	41.65%	41.14%	39.93%
Intermediate	21.60%	23.18%	25.27%	25.77%	25.51%	26.78%	27.30%	27.11%
Higher education	21.95%	23.84%	31.15%	30.64%	30.14%	30.74%	30.85%	30.24%
Missing	0.27%	0.11%	0.10%	0.48%	0.87%	0.83%	0.70%	2.73%
Social class based on occupation								
Managerial	29.08%	30.39%	31.8%	32.68%	33.60%	34.54%	39.20%	32.45%
Intermediate	22.50%	24.35%	24.81%	24.57%	25.35%	25.17%	24.80%	25.22%
Routine	44.69%	43.58%	41.79%	39.44%	39.36%	38.66%	34.20%	35.41%
Missing	3.72%	0.09%	0.12%	2.08%	0.59%	0.59%	0.88%	6.09%
Marital Status								
Married*/ Cohabiting	68.75%	67.84%	67.56%	68.95%	68.17%	67.77%	67.49%	67.17%
Single	5.05%	4.74%	5.21%	5.53%	5.16%	5.61%	5.16%	5.37%
Widowed	17.13%	18.30%	17.13%	15.51%	16.55%	16.05%	16.61%	16.45%
Divorced/ Separated	9.08%	9.12%	10.10%	10.01%	10.12%	10.57%	10.74%	11.01%
Non-white	2.81%	2.30%	2.72%	3.06%	3.10%	3.51%	3.41%	3.5%

Table 3.1: Demographic characteristics of the ELSA participants for waves 1 to 8. *Married category includes those in civil partnership from 2006. **Wealth is measured in British pounds and represents the net total wealth which is the sum of savings, investments, physical wealth, i.e., satisfaction-generating physical goods, and housing wealth after subtraction of any financial or mortage debt.

[difficulties in Activities of Daily Living (ADLs) and Instrumental Activities of Daily Living (IADLs)]; eyesight; hearing; quality of medical care received; fall incidents and fractures experienced; pain experienced; urinary incontinence symptoms; depression symptoms; walking speed performance (if $\geq$ 60 years old); medication intake records; respiratory symptoms; ischemic heart pain symptoms; dental health record; balance; dizziness; quality of sleep; and, menopause symptoms. However, not all of the above areas are covered in the current ELSA waves (1 to 8). In particular, only the first 12 topics have been consistently reported across all data collection waves. Smoking history, alcohol consumption and physical activity frequency were also reported across all waves, whereas information on consumption of fruit and vegetables was collected across the ELSA waves 3 to 8.

Specific tests and blood assays were used to collect data on physical examination and metabolic health during the nurse visit (NV) in waves 2, 4, 6 and 8. Individual height, weight, blood pressure, waist and hip circumference were measured by the nurses, and chair rise, balance, grip strength and spirometry tests were performed by the participants to evaluate their physical functioning. Blood samples were also collected by the trained nurses and, then, sent by post to a laboratory to be processed and analysed. Some of the biomarkers measured include: triglycerides, total and DHL-cholesterol, C-reactive protein, fibrinogen, haemoglobin and ferritin, white cell count, fasting lipids, glucose and glycated haemoglobin. Finally, DNA samples were collected along with the blood assays between 2013 and 2014 (access with fee).

Module 2: Psychosocial measures

Caring for others motivation; volunteering and unpaid help; civic, social and cultural participation; social networks (relationship with family, friends, children, partner and neighbours); social support; loneliness; use and means of transport; access to local amenities and services; and, TV watching, were some of the areas covered through the CAPI and self-reported questionnaire on social and civic participation. Additional questions were handed to evaluate respondents' relationship between effort and reward, control and demand, provision of care, altruism and collectiveness. Perceptions of respondents' life and ageing were also reported through both the self-reported questionnaire and the CAPI. Records included information on how individuals rate their social and financial

status, how they perceive life at old compared to middle age and how they experience ageing. Psychological and social well-being were measured using the CASP-19 questionnaire (Hyde et al., 2003) and the Satisfaction With Life Scale (Diener et al., 1985) at seven out of the eight waves. Additional well-being questionnaires were used in later waves to assess positive affectivity, i.e., positive emotions and self-expression, other personality facts and psychological pain.

Module 3: Cognitive function measures

From 2002 until recently, the ELSA examined many cognitive skills including memory, executive function, literacy and numeracy abilities. Memory self-rating, orientation in time, word-list learning, prospective memory and fluid intelligence tests were used for memory strength assessment. A selection of word-finding, backwards counting and object naming tests were used for assessing executive function, i.e., the set of cognitive skills needed to organise thoughts and activities, prioritise tasks, manage time efficiently, and make decisions. Literacy was assessed by a set of questions evaluating respondents understanding of the label instructions on a fabricated medication product. Numeracy was assessed by a set of five questions which required numerical calculations of varying difficulty.

Cognitive function information is not consistent across the different waves. In particular, only orientation in time, word-list learning (verbal learning and recall) and word-finding (verbal-fluency) were measured with high frequency and consistency across the eight data collection waves. Difficulties in everyday tasks showing cognitive function skills were also assessed through questions measuring difficulties in ADLs and IADLs, for example, difficulty in taking everyday medication, navigate using a map, making phone calls, etc.

Module 4: Economics data

The economics module includes information on individual and family finances; in particular, household income, wealth and debts, pension arrangements, employment, consumption and expectations about certainty of future events and financial decision-making. Most of the items in this module were measured across all eight ELSA waves providing information on respondent's earnings, benefits and assets; current and past pension plans;

job details (pay, working hours, job description); retirement details (when applicable); disabilities affecting work; other income sources or secondary jobs; work limitations due to health problems or disabilities (when applicable); housing; vehicle and durable goods ownership; purchases and fuel expenditures; health insurance, leisure and clothing expenses.

Additional modules

A life-history interview was conducted between 2004 and 2005. Its primary aim was to understand and record important events that may have played a key role on participants' health, economic circumstances and quality of life in subsequent years. A risk preferences module was also added in wave 5 to assess risk tolerance and patience of older adults in England. Risk tolerance is defined as the willingness to bear risk in pursuit of possible reward and patience is defined as the willingness to delay in return for a greater reward. Finally, a sexual health questionnaire was used in waves 6 and 8 to examine sex attitudes, sexual activities and experiences, partners and lifetime desires and experiences of older adults in the UK.

3.3.4 Missing data, drop-outs and mortality records

Since the ELSA data can be analysed both cross-sectionally and longitudinally, two types of missing data can be identified: cross-sectional missingness, i.e., missing responses in a single ELSA data wave caused by data collection errors, individual unwillingness to respond, etc.; and, longitudinal missingness, i.e., drop-outs. In ELSA, we can assume that cross-sectional missingness is a result of a missing at random (MAR) mechanism. This type of missingness can often be handled by employing specific algorithms or statistical modelling approaches during inference, e.g., Expectation-Maximisation (EM) algorithm, to produce valid parameter estimates (Molenberghs et al., 2014). Missing completely at random (MCAR) would be unrealistic as we cannot be certain that missing responses are all occurring completely randomly and that they are not related to some underlying disability, poor health or consistent data collection error. In fact, MCAR requires that missingness in a particular variable does not depend on the responses for any other variable (Little, 1988), which as shown later in this chapter, is not the case for ELSA. At the same time, we cannot be certain for a missing not at random (MNAR) mechanism

as sometimes non-response can be indeed completely random and occur due to random and indistinguishable reasons. In principle, a MNAR missingness mechanism is plausible for the ELSA data, however, it is not possible to distinguish between MAR and MNAR models from the observed data alone (Molenberghs et al., 2008). Hence, for identifiability purposes, we pursue MAR. However, longitudinal missingness should be treated with caution, especially since the sample interviewed by ELSA consists of middle aged or older individuals who are prone to experience negative outcomes, such as serious illness or death. Below, we give a detailed overview of the missingness mechanisms observed in the ELSA data.

Cross-sectional missingness

Looking at each data wave separately, one can observe that there is not any specific pattern in the missing values mechanism. Since most of the CAPI questions permit "refusal" and "don't know" answers, individuals have the chance to omit responding to some of the questions. There are cases where respondents did not complete the entire interview but the specific reasons explaining why this happened are not clear. In most of the ELSA variables missing responses are below 25% with an exception in the biomarker data where missing rates sometimes exceeded 30% due to errors during blood sample collection, delivery and analysis, Figure A.1.

In the study, sometimes, cross-sectional missingness was minimised by the study design itself or by performing imputation techniques. For instance, in the economics module, non-response was minimised through a system of "unfolding brackets" allowing respondents to make range-restricted estimates when the exact information was not given. The bracketed values were then used as a range for subsequent imputation. In Section 3.3.5, we give a more comprehensive view of the missing responses detected in the subset of the ELSA data which was used for further analysis.

Drop-outs and mortality

Drop-outs are a major issue in longitudinal data, especially in surveys where the age of individuals is advanced and there is a large chance of getting ill or dying. In ELSA, the average age of individuals ranges between 65 and 71 years across the eight waves, thus suggesting that the probability of participants experiencing adverse health outcomes

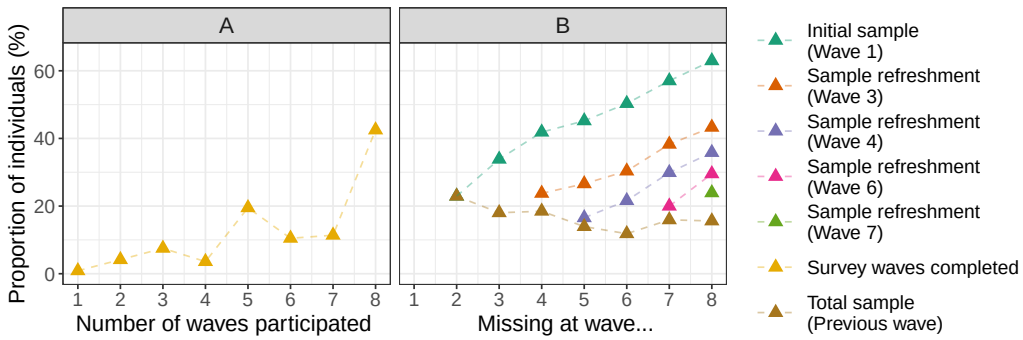
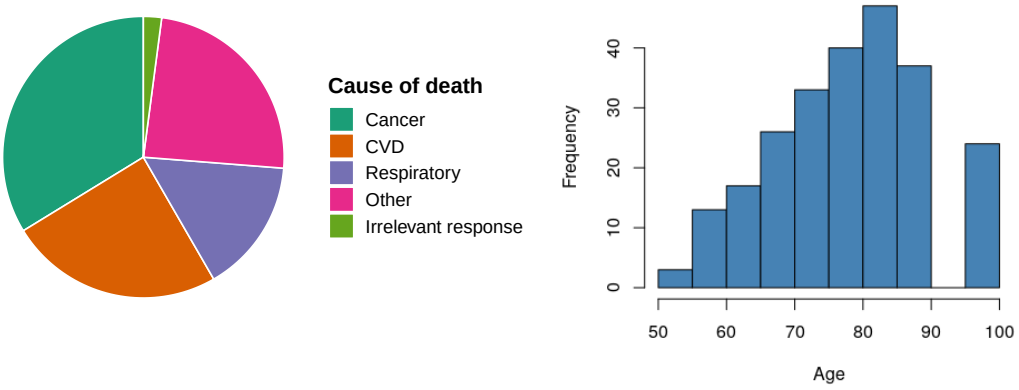


Figure 3.5: (A) Proportion of individuals who participated in 1, 2 or more ELSA data collection waves. The total number of core ELSA participants up to the eighth wave was 16,084. (B) Proportion of individuals get missing from one ELSA wave to the next. Samples presented refers to the initial core ELSA sample (cumulative proportion per future wave), sample refreshments (cumulative proportion per future wave) and total sample (initial sample + sample refreshments; proportion presented corresponds to the missing rates from previous wave to the next only).

which can make them incapable of completing the study is large. In fact, only 42.5% completed all eight ELSA waves (around 4,800 individuals−mean age at wave 1: 61; median age at wave 1: 59−, Figure 3.5A).

More than half of individuals from the initial sample became lost to follow-up by wave 8 (≈63%; mean and median age at wave 1 was 67). However, due to sample refreshments the missingness rates in between subsequent waves remained balanced between ≈15% and ≈20% (Figure 3.5B). Similar to the missing rate pattern of the initial sample, the missing rates for the sample refreshments (waves 3, 4, 6 and 7) follow a steady increase in future ELSA waves (Figure 3.5B). At this point, we should mention that individuals sometimes left the study and re-entered a few years later. The missing rates reported above refer to all individuals without accounting for any study re-entrants.

Mortality records are available in ELSA from wave 6 (2012/13), that is the "End of Life" data. This is an interview attempted with a proxy respondent (usually a family member) who could provide information on the circumstances of a deceased participant. From the 2,886 core ELSA members known to have been deceased by 2012, interview data were collected for 34% of them (Crawford and Mei, 2018). However, the data accessible for analysis include information on only 240 deaths (mean death age: ≈79 years). The primary causes of death is cancer and cardiovascular diseases (CVD), with the percentage of deaths due to cancer reaching 34% whereas for deaths due to CVDs is close to 25% (Figure 3.6a).



(a) (b)

Figure 3.6: (a) Death causes of 240 core ELSA members as reported in the "End of Life" interview conducted between 2012 and 2013. (b) Last reported age distribution of the 240 deceased core ELSA members as reported in the "End of Life" interview.

3.3.5 Variables used to study healthy ageing and frailty trends

In this section, we present the specific variables that have been used to study age-varying trends of frailty and the domains that constitute a healthy ageing phenotype. The selection was made based on variable content, consistency in data collection, proportion of cross-sectional missingness and research objectives. Apart from the demographic and economic data, we employed variables describing key elements of physical, social, mental, cognitive and metabolic health in older adults. For all variables presented below higher scores indicate better health whereas lower scores indicate worse health.

Variables measuring metabolic and physical health

In ELSA, metabolic health was directly assessed through the NV by analysing participants' blood samples and through the CAPI. We deployed biomarker, medication, cardiovascular and chronic disease records to evaluate metabolic and physiological malfunctioning in ELSA participants. Records on medication intake for controlling asthma, other lung conditions, high cholesterol levels, high blood pressure and diabetes were extracted from the CAPI data, along with anticoagulant, blood thinning medication records and Warfarin intake records when applicable. Data on participants' height and weight, blood C-reactive protein, blood triglyceride, blood HDL cholesterol, systolic and diastolic blood pressure, blood fibrinogen, glucose and glycated haemoglobin levels were also extracted

from the NV data. Using participants' up to date weight and height, we further calculated participant Body Mass Index (BMI) to be used as a body composition measure. It was computed by dividing individual weight by the square of individual height. Since height information was missing at wave eight, we used height information from the preceding waves to produce the corresponding BMI.

Cardiovascular disease, chronic disease and medication intake information were encoded using binary indicators. For biomarker and body composition data, ordinal indicators and tertiles were used similar to previous literature (Sanders et al., 2014). Due to the high sensitivity of the blood samples and the poorer consent of the ELSA participants, missing rates in biomarker data were higher compared to the rest of the variables used for further analysis. As discussed previously, for some of the metabolic health variables missing rates were approaching almost 68% in certain waves. For more details on the specific variable coding and variable missing rates readers can see the Supplementary Text A.1 in the Appendix.

Previous studies suggested that physical health can be described through a combination of five physical functions: endurance, balance, strength, dexterity and locomotion (Lara et al., 2013a). Physical ability tests, such as chair rise test (Smith et al., 2010), provide the appropriate markers for physical capability quantification. The ELSA captures physical capability in two ways: through endurance, balance, locomotion and strength tests conducted every four years during the NV interview; or, through the questions referring to difficulties in ADLs, IADLs and mobility in the self-reported questionnaire every two years. Although physical function tests better capturing physical capability, they lack collection consistency–except for the grip strength test– resulting in a high rate of missingness. Therefore, the self-completion questionnaire data were used in future applications.

The self-completion questionnaire in ELSA distinguishes three domains of physical functioning: ADLs or self-care activities; IADLs or activities necessary for independent living in a community; and mobility (here, lower and upper-limb function). During the interview, two cards were shown to respondents: the first was asking "Because of a health problem, [do you/does he/does she] have difficulty doing any of the activities on this card? Exclude any difficulties that you expect to last less than three months." (difficulties with mobility); and the second was asking "Here are a few more everyday activities.

Please tell me if [you have/he has/she has] any difficulty with these because of a physical, mental, emotional or memory problem. Again exclude any difficulties you expect to last less than three months. Because of a health or memory problem, [do you/does he/does she] have difficulty doing any of the activities on this card?" (difficulties with ADLs and IADLs). The specific list of the depicted difficulties is given in Supplementary Text A.1 of the Appendix. All responses to ADLs, IADLs and Mobility questions were binary: yes, if difficulty was present; and, no, if difficulty was absent. Since a positive answer indicates worse health, reverse coding was used.

Overall physical health variables have low missing rates (less than 2% in every wave). However, there were a number of difficulties in ADLs, IADLs and mobility that were not questioned consistently. The most frequently recorded difficulties were climbing several flights of stairs without resting and stooping, kneeling or crouching, whereas the least frequent were difficulties in eating, taking medications and recognising when in physical danger. Since all of those difficulty questions were created to detect different levels of disability in older adults, we observed smaller variability in variables referring to difficulties with ADLs and IADLs (easier tasks−most individuals answered no, i.e.,, no difficulty is experienced) and larger variability in those referring to difficulties with mobility (more demanding tasks).

Variables measuring cognitive function

Orientation in time, word-list learning (verbal learning and recall) and word-finding (verbal-fluency) can be used as guides to examine the presence or severity of cognitive impairment at old age (Lara et al., 2013a). From the ELSA data, we selected six cognitive health variables (except from wave 6 where one variable was missing) which were collected across most of the waves. The extracted data include memory-associated questions, such as whether participants were able to correctly remember current date, word recall immediately and after delay; memory illness records, such as dementia and Alzheimer's disease, and; a verbal fluency question (how many animals the participant was able to mention in 1 minute). For questions related to verbal fluency and memory, we used tertiles as cut-off points with lower scores indicating impaired cognitive functioning (specific coding can be found in Supplementary Text A.1 of the Appendix). Dementia and Alzheimer's disease records have been used to minimise missing rates in the cognitive

function variables by assigning worse scores to participants suffering from these diseases, since analysis showed that often missing cases occurred due to incapability of participants to respond caused by some cognitive disorder. Overall missing rates for cognitive function variables were less than 5% at all waves.

Variables measuring social well-being

Social well-being is a controversial concept and its operationalisation is a challenging task. Multiple metrics and questionnaires have been used in order to measure it in the literature (Lara et al., 2013a). However, four main areas related to social well-being can be identified and measured: social network, social functioning, perceived emotional and social support, and sense of purpose. In ELSA, there were multiple questions focusing on social relationships, networks, isolation and social support. In particular, social network and perceived emotional and social support were assessed by asking participants whether they have family, friends, children or partner; and, if yes, they were asked to provide more information on the quality of relationship with them and contact frequency. Societal involvement and social participation were assessed by collecting data on participation to organisations, frequency of visiting museums and art galleries (cultural participation) and frequency of going on holidays, reading newspaper or spending time in personal hobbies. Finally, as an indicator of unhealthy social functioning responses on the UCLA Loneliness Scale were used. For more information readers can look at Supplementary Text A.1. of the Appendix. These data consisted of binary and ordinal variables. Missing rates varied from 10% to less than 25% at all waves. The only exception was observed in the third wave in variables referring to cultural participation where missing rates varied between 18% and 28%. It is not clear though why missing rates are that much higher in this particular wave so we suspect that this can be due to data collection errors.

Variables measuring psychological well-being

In ELSA, psychological well-being is measured through assessing evaluative−how satisfied people are with their lives−, affective/hedonic−how people feel, refers to emotions−and eudemonic−purpose of life−well-being (Steptoe et al., 2015). Therefore, the following information was extracted from the ELSA data set: Satisfaction with Life Scale responses [SWLS; ordinal responses; Diener et al. (1985)], CASP-19 questionnaire responses [ordinal

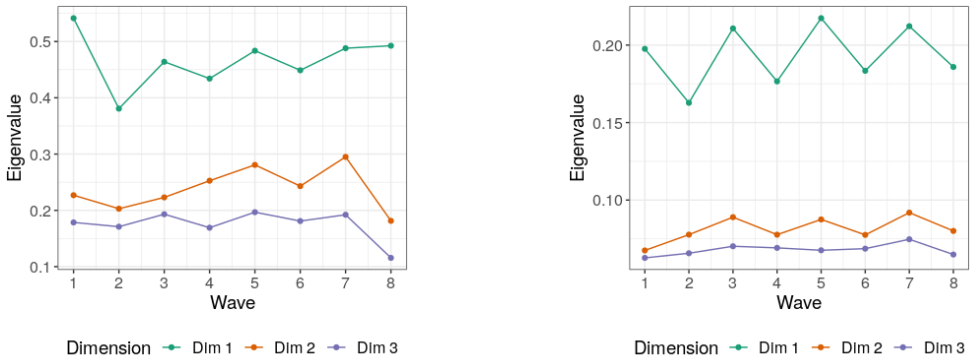
responses; Hyde et al. (2003)] and CES-Depression Scale responses [binary responses; (Radloff, 1977)]. Occurrence of any diagnosed psychiatric problem was also taken into account as recorded in the ELSA health questionnaires. Missing rates varied from 3% to 7% for the CES-D questionnaire and from 10% to 18% for the SWLS and CASP-19 questionnaires with higher missing rates being associated with the most recent waves. As for the occurrence of diagnosed psychiatric conditions, the percentage varied from ≈10% to 25% with occurrence becoming larger as we come closer to present. For more information on the exact questions included into the ELSA questionnaire, readers can look at Supplementary Text A.1. of the Appendix.

3.3.6 Targeted multivariate exploratory data analysis

The amount of data selected for further analysis exceeds 100. To do some primary data exploration and analysis we used Multiple Correspondence Analysis [MCA; Greenacre and Blasius (2006)] and Factor Analysis [FA; Rummel (1988)]. Both are dimensionality reduction techniques which allow to capture the variance in the observed variables through a smaller set. MCA is used as an alternative of the Principal Component Analysis [PCA, Dunteman (1989)], since PCA cannot be used for nominal data. Under MCA, the multi-dimensional data is projected to a 2 or (sometimes) 3 dimensional euclidean space to identify similarities between group of variables, individuals or variable levels (the group of individuals associated with two similar variable levels are themselves similar). Alternatively, FA was used to relate the variables under study to a small number of latent factors through a pre-specified measurement model.

Multiple Correspondence Analysis

In this subsection we describe the results of the MCA performed on the extracted ELSA data presented in Section 3.3.5. As a first step, we excluded variables with a cross-sectional missing rate larger than 30% (including the metabolic health data). Two types of analysis were performed since we detected patterns in the occurrence of missing responses: a full data analysis (A) and a complete case analysis (B). MCA was performed cross-sectionally for each ELSA wave separately. The total percentage of variance explained from the first two dimensions in A was between 17.55% and 24.76%, whereas for B, it fluctuated between 10.34% and 13.07% (Figures B.1 and B.2). In both analyses, dimension 1



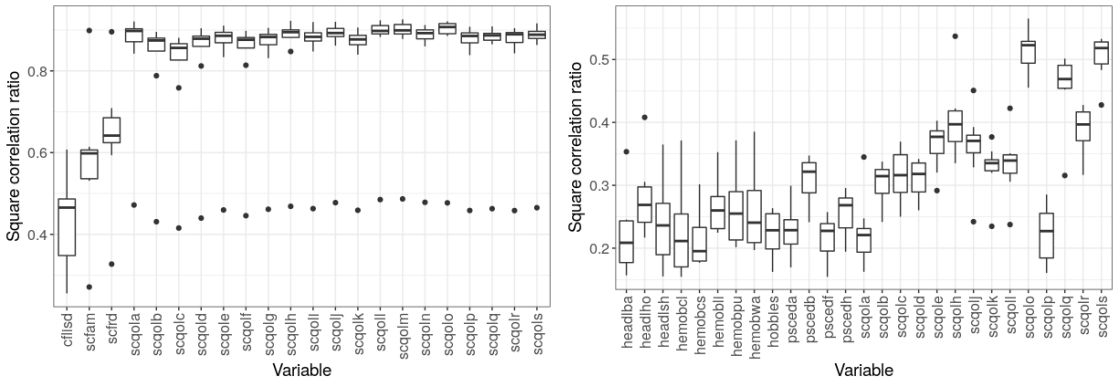
(a) (b)

Figure 3.7: (a) Multiple correspondence analysis eigenvalues for the first three dimensions. Analysis was conducted on the full data. (b) Multiple correspondence analysis eigenvalues for the first three dimensions. Analysis was conducted on the complete case data.

explained the larger percentage of variance. The eigenvalues for the first three dimensions of both analysis are presented in Figure 3.7.

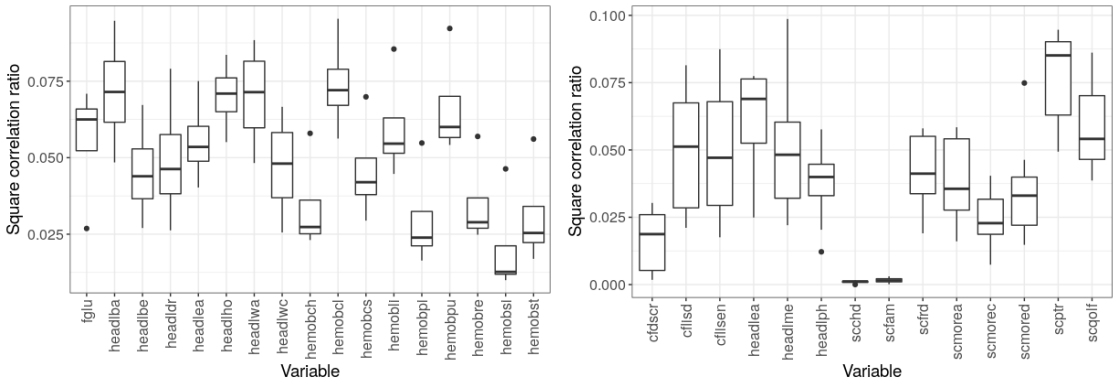
The variables with the larger and lower discriminatory power from both analysis are depicted in Figures 3.8 and 3.9. Analysis A showed that variables measuring word recalling ability after delay, social relationships with family members and friends and respondent's quality of life had the greatest discrimination power at all waves (square correlation ratio[2] greater than 0.25), whereas those measuring respondent's blood glucose levels, difficulties in ADLs, IADLs and mobility had the worst (square correlation ratio less than 0.15). This discrimination power, though, might be due to the fact that often individuals who have not responded in one of these questions, for instance, the quality of life questions, they did not respond to the rest of them. In contrast to analysis A, some of the variables measuring difficulties in ADLs, IADLs and mobility in the complete case analysis were classified amongst those with the largest discrimination power, along with those measuring depression symptoms and quality of life. Variables measuring cognitive function and social relationships with children, partner, family members and friends showed low discrimination power along with a small amount of variables measuring difficulties in eating, making a telephone call and taking medication. These results suggest that further variable screening and investigation are needed to identify the variables that might distinguish individual ageing profiles.

[2]The square correlation ratio is defined as the ratio of the between group level sum of squares over the total sum of squares; it determines the proportion of the total variability explained by the principal components for a specific variable.



(a) (b)

Figure 3.8: (a) Variables with the best discrimination power across the 8 ELSA waves (loadings greater than 0.25). Analysis performed using the full data. (b) Variables with the best discrimination power across the 8 ELSA waves (loadings greater than 0.15). Analysis performed using only the complete cases.



(a) (b)

Figure 3.9: (a) Variables with the worst discrimination power across the 8 ELSA waves (loadings smaller than 0.15). Analysis performed using the full data. (b) Variables with the worst discrimination power across the 8 ELSA waves (loadings smaller than 0.15). Analysis performed using only the complete cases.

A perceptual map of the categories of the variables investigated is illustrated in Figure 3.10 (complete case analysis). The corresponding map for the full data analysis is presented in Figure B.3, since we believe that it is not that informative apart from the fact that the geometric distance between the missing and not missing categories of variables in the two-dimensional plane suggests an association between missing responses for some individuals (individuals who have not responded in one question they did not respond to the rest of the questions). Different colours were used to show the quality of each point for the indicated dimensions based on their squared cosine value (the lower its value is the worse the quality). Overall, the quality of the perceptual map is low for most

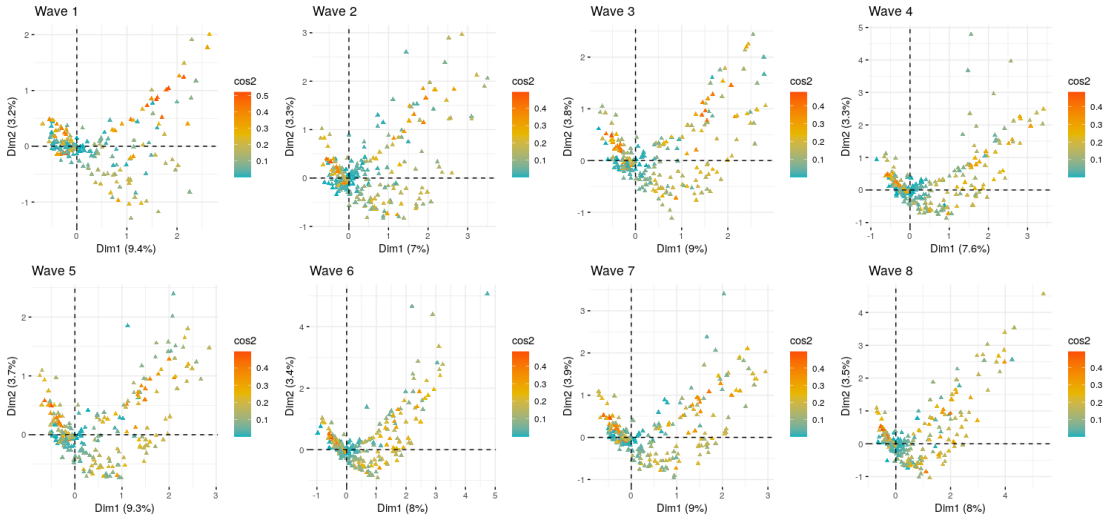


Figure 3.10: Perceptual map of categories of variables investigated. Different colours were used to show the quality of each point for the indicated dimensions based on their squared cosine value. Analysis was performed using only complete cases.

points suggesting that further analysis is needed to identify the group of variables that will be informative in further analysis. Ageing profiles cannot be easily distinguished by observing the association between the proximity of variable categories.

Multiple correspondence analysis revealed candidate variables for exclusion in future analysis. Since the data consists of answers in groups of questions targeting specific traits or behaviours, respondents who did not respond to one of the questions in one group seem to present missing responses to the rest. This is shown from the clusters observed in the results from analysis A. We conclude that we might consider excluding individuals with a large number of missing values in a large part of the variables examined.

Exploratory and Confirmatory Factor Analysis

Exploratory Factor Analysis (EFA) was performed to determine the minimum number of factors required to obtain a good fit to the data whilst maintaining meaningful interpretation of the identified factors. All variables (or items) were allowed to load to all factors and factors to be correlated with each other (oblique rotation). Furthermore, the latent structure restricted to simple structure (i.e only one factor loading per observed variable being non-zero) to ensure clear interpretation and simpler structure for later stages. A Confirmatory Factor Analysis (CFA) step was, then, performed to confirm the latent structure identified through the EFA step by restricting some factor loadings to be constrained at zero. Goodness of fit was assessed through model fit indices: the Root

Mean Square Error of Approximation (RMSEA), the Comparative Fit Index (CFI) and the Tucker-Lewis Index (TLI). An acceptable fit is achieved if RMSEA<0.05, CFI>0.9 and TLI>0.95 (Hu and Bentler, 1999; Hooper et al., 2008; Kaiser, 1960).

The exact number of variables subject to exploration was 150. To decrease model fit complexity and since the data are characterised by a natural structure (variables measuring cognitive, mental, physical, metabolic and social health−let us call them "domains"), we explored the underlying data substructure by handling data from each of those domains separately. In addition, since our study objective is to examine health changes as people are getting older, we repeatedly fit the factor model for all age points in the data[3]. Individuals aged between 50 and 90 were examined. Data was pooled across all waves. A large amount of variables were omitted from the EFA step due to high correlations with other variables loading to the same factors (constructs) or due to small loadings [less than 0.316, Tabachnick and Fidell (2001)] or cross-loadings. In total, 47 variables were selected for further analysis and 14 latent constructs were identified.

Table 3.2 shows an overall acceptable fit under the assumptions and constructs specified in CFA step. Figures 3.11 and 3.12 show the estimated measurement models for the first age group point through path diagrams. Circles represent the latent constructs and rectangles represent the observed variables. Estimated factor loadings are given on the arrows between the latent factors and observed variables. Since many of the variables had more than 2 levels, estimated thresholds are not depicted for better readability. Residual variances for the factors and observed responses are indicated by the dashed arrows. Between factor correlation is given on the arrows between them.

Psychological health was measured through 15 variables and a combination of 5 sub-constructs were identified: pleasure, control, self-realisation, satisfaction with life and negative effect. The first three latent factors were measured through items in the CASP-19 questionnaire. The rest consist of variables from the SWLS and CES-D questionnaires. For most constructs, factor loadings were relatively large (>0.8; Figure 3.11). Overall, an acceptable fit was achieved for all age points (Table 3.2).

Episodic memory and executive function were identified as sub-constructs measured through the cognitive function test results and the reported difficulties in managing money, reading a map, taking medication and making phone calls as part of the cognitive

[3]We originally had 40 age points but we decreased them to 20 age points to facilitate EFA and CFA implementation, i.e., $t_1 = [50, 51]$, $t_2 = [52, 53]$, etc.

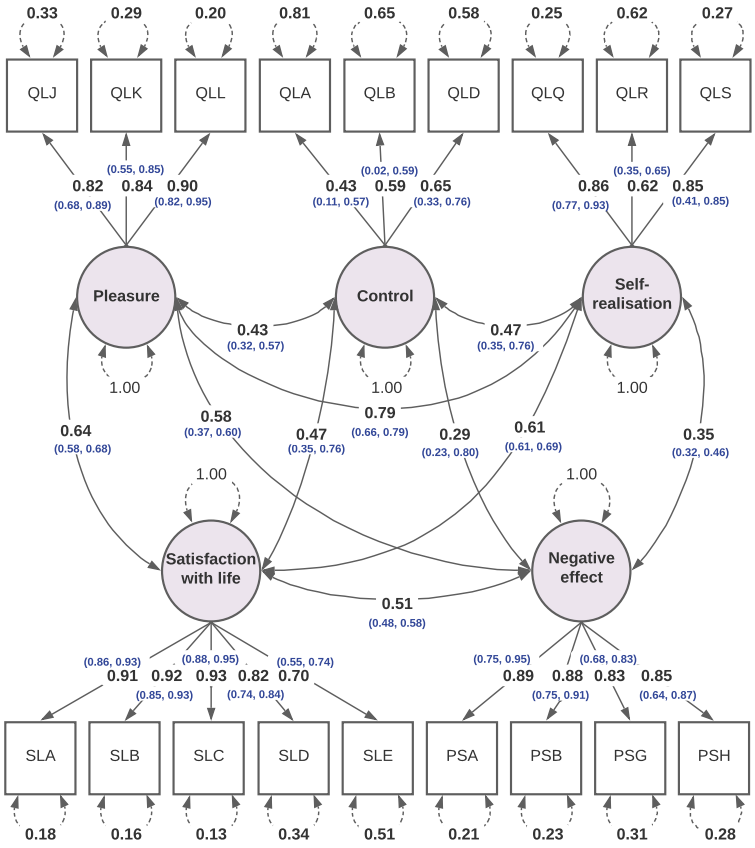


Figure 3.11: Path diagram for the estimated measurement model of psychological health for the first age point. Minimum and maximum values for each parameter estimate across all age points are shown in blue. The variables shown include information on: QLJ-how often individuals look forward to each day; QLK-how often individuals feel that their life has meaning; QLL-how often individuals enjoy the things they do; QLA-how often individuals feel age prevents them from doing things they like; QLB-how often individuals feel what happens to them is out of their control; QLD-how often individuals feel left out of things; QLQ-how often individuls feel satisfied with the way their life has turned out; QLR-how often individuals feel that life is full of opportunities; QLS-how often individuls feel the future looks good to them; SLA-whether they believe that in most ways their life is close to their ideal; SLB-whether they believe that the conditions of their life are excellent; SLC-whether they are satisfied with their life; SLD-whether, so far, they have got the important things they want in life; SLE-whether they would not change anything, if they could live their life again; PSA-whether individuals felt depressed much of the time during past week; PSB-whether individuals felt that everything they did was an effort; PSG-whether individuals felt sad during past week; PSH-whether individuals could not get going much of the time during past week.

health domain. An equality constraint was applied between items measuring episodic memory to improve model fit. All factor loadings were very large (>0.82, Figure 3.12) with small standard errors (<0.05).

Metabolic health and cardiovascular health were the latent constructs measured through a combination of items in the NV tests and CVD CAPI questionnaire as part of the general metabolic health domain. Smaller factor loadings with larger standard errors were found. Finally, five sub-constructs were identified after analysing the social and physical health data. In particular, a small amount of difficulties in mobility questions were kept whilst most of the variables measuring difficulties in ADLs and IADLs were omitted through the EFA step (small factor loadings). Social isolation, desire for social change and cultural participation were the resulting sub-constructs after performing EFA on the social health variables examined. An equality constraint was used to improve model fit, resulting in large factor loadings with relatively small standard errors. Similar results were found after performing EFA and CFA for the rest of the age points. In fact, the narrow range of the estimated factor loadings and correlations in Figures 3.11 and 3.12 show consistent patterns of the latent structure for all domains over all age points.

	Cognitive function	**Physical capability**	**Social well-being**	**Psychological well-being**	**Metabolic health**
RMSEA	0.004 [0, 0.026]	0.006 [0, 0.031]	0.028 [0, 0.046]	0.032 [0.007, 0.039]	0.032 [0.023, 0.039]
CFI	0.999 [0.997, 0.999]	0.999 [0.999, 1]	0.998 [0.995, 1]	0.997 [0.994, 1]	0.965 [0.933, 0.984]
TLI	1.003 [0.996, 1.058]	1 [0.998, 1.006]	0.997 [0.993, 1.004]	0.996 [0.993, 1]	0.951 [0.908, 0.978]

Table 3.2: The Root Mean Square Error of Approximation (RMSEA), the Comparative Fit Index (CFI) and the Tucker-Lewis Index (TLI) mean values along with minimum and maximum values in the brackets of CFA fit for the selected variables and the specified sub-constructs that have been identified through EFA acros all age points.

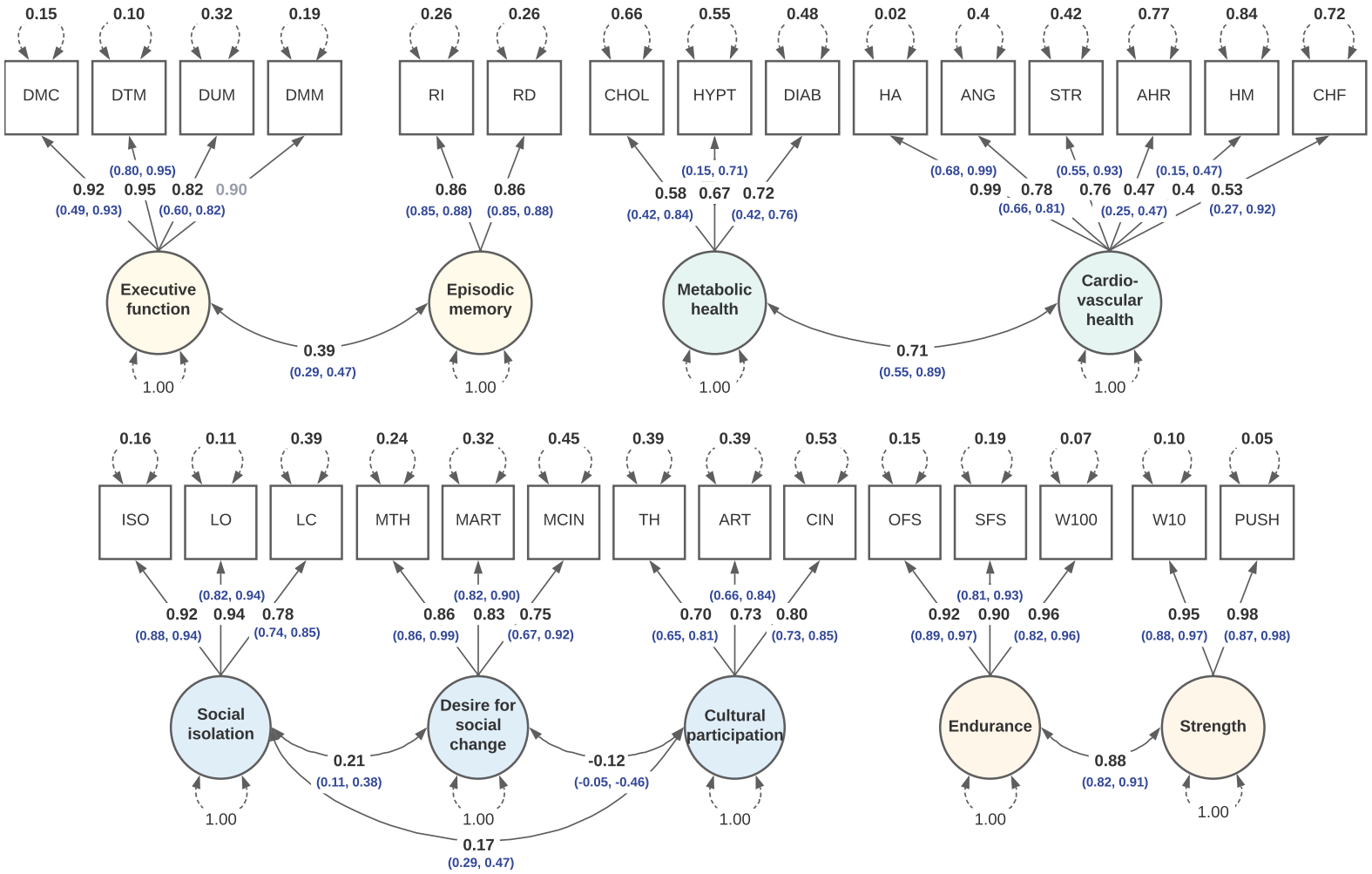


Figure 3.12: Path diagrams for the estimated measurement model of cognitive, metabolic, social, and physical health for the first age point. Minimum and maximum values for each parameter estimate across all age points are shown in blue. A single parameter which was fixed to the same value across all age points to improve model fit is shown in grey. The variables shown include information on: DMC-difficulty in making phone calls; DTM-difficulty in taking medications; DUM-dIfficulty in using a map; DMM-difficulty in managing money; RI-ability to recall words immediately; RD-ability to recall words after delay; CHOL-cholesterol levels; HYPT-whether individual has hypertension; DIAB-whether individual is diabetic; HA-whether individual has ever experience a heart attack; ANG-whether individual has angina; STR- whether individual has ever experienced stroke; AHR-whether individual has abnormal heart rhythm; HM-whether individual has ever experienced heart murmur; CHF-whether individual has ever experienced congestive heart failure; ISO-whether individual feels isolated; LO-whether individual feels left out; LC-whether individual feels lack of companionship; MTH-desire of going more often to the theatre, a concert or the opera; MART-desire of going more often to art gallery or museum; MCIN-desire of going more often to the cinema; TH-frequency of going to the theatre, a concert or the opera; ART-frequency of going to art gallery or museum; CIN-frequency of going to the cinema; OFS-difficulty in climbing one flight of stairs without resting; SFS-difficulty in climbing several flights of stairs wihout resting; W100-difficulty in walking 100 yards; W10-difficulty in lifting or carrying weights over 10 pounds; PUSH-difficulty in pulling or pushing large objects.

3.4 Summary

In this chapter, we provided a detailed overview of the data handled across the different studies of the current thesis. These data came from two major databases: the GDSC and the ELSA. In the first part of the chapter, we presented the GDSC data, an online publicly accessible resource for therapeutic biomarker discovery in cancer cells; we described the specific data fraction extracted for study purposes in Chapter 4; and, we discussed the performed data pre-processing. Subsequently, the ELSA data were presented where a selection of statistical tools and background literature were used to explore and screen the available information. This second part of the chapter gives a general overview of the quality of the ELSA data and basic characteristics. The analysis performed is a preliminary step towards revealing ageing trends over time. The ELSA sample was, first, characterised through some simple descriptive statistics. We, then, focused on a fraction of the ELSA variables which were selected based on relevance to ageing and health and some additional quality criteria. The latent structure of the data was, finally, explored and assessed using a number of multivariate statistical methods such as MCA, EFA and CFA. We concluded that even though there was a large amount of available variables in the data, due to overlapping information, it would be possible to select a subset of them for further analysis without much loss of information.

Chapter 4

A regularized functional regression model enabling transcriptome-wide dosage-dependent association study of cancer drug response

4.1 Introduction

As discussed in Chapter 2, cancer treatments can be highly toxic and frequently only a subset of the patient population will benefit from a given treatment. Previous research shows that tumour genetic makeup plays an important role in cancer drug sensitivity. Therefore, we suspect that gene expression markers could be used as a decision aid for treatment selection or dosage tuning.

With regards to the high-dimensional nature of genomic data sets, it is worth noting that highly-complex data sets with non-stationary trends are not easily amenable to analysis by classic parametric or semi-parametric mixed models. Such effects, e.g., the effect of genes on drug response over different drug dosages (dose-varying effect), can be examined using varying coefficient models which allow for the covariate effect to be varying instead of constant (Hastie and Tibshirani, 1993). Methods to estimate varying

covariate effects include global and local smoothing, e.g., kernel estimators (Wu et al., 1998; Wu and Chiang, 2000), basis approximation (Huang et al., 2004) or penalized splines (Qu and Li, 2006). Although non-parametric techniques can reduce modelling biases (Fan et al., 2011), they often suffer from the "curse of dimensionality" (Geenens et al., 2011). Inference in these models becomes impossible as the number of predictors increases, and often selecting a smaller number of important variables for inclusion into the model is clinically beneficial. Sparse regression has enabled a more flexible and computationally "inexpensive" way of choosing the best subset of predictors (Song et al., 2014). However, these methods cannot handle ultra-high dimensional problems without losing statistical accuracy and algorithmic stability, since they handle all of the predictors jointly. Consequently, there is a need of prior univariate tests focused on filtering out the unimportant predictors by estimating the association of each predictor to the outcome variable separately (Chu et al., 2016; Fan et al., 2011, 2014). The advantage of using varying coefficient models along with a variable screening algorithm on genomic data sets was first introduced to explore the effect of genetic mutations on lung function (Chu et al., 2016). Recently, Wang et al. (2020) and Tansey et al. (2018) independently proposed methods for modelling drug response curves via Gaussian processes and linking them to biomarkers. In both cases, the authors did not use their models for dosage-dependent inference of biomarker effects. Additionally, the highly non-linear neural network model in Tansey et al. (2018) makes interpretation of biomarker effects challenging.

In this chapter, we extended the methodology of Chu et al. (2016) to the objective of assessing the transcriptomic effect on anti-cancer drug response, where our coefficient functions were allowed to vary with dosage. We developed a functional regression framework to study the effectiveness of multiple anticancer agents applied in different cancer cell lines under different dosage levels, adjusting for the transcriptomic profiles of the cell lines under treatment. We considered a dose-varying coefficient model, along with a two-stage variable selection method in order to detect and evaluate drug-gene relationships, and then applied this method to data extracted from the GDSC project (Yang et al., 2012). To compare and differentiate similar treatments, we examined a case study of five *BRAF* targeted compounds under different dosages to almost 1,000 cancer cell lines. We used baseline gene expression measurements for the cancer cell lines to investigate gene-drug response relationships for almost 18,000 genes. Gene

rankings were obtained based on the estimated effects of the genes on the drug response. The resulting model describes the whole dose response curve, rather than a summary statistic of drug response (e.g., IC_{50}), which allowed us to identify trends in the gene-drug association at untested dose concentrations.

In the following section, we present an efficient two-stage variable selection algorithm for complete repeated measures response data and high-dimensional covariates. This consists of variable screening followed by sparse estimation of the dose-dependent covariate functions. This allows for the dose-dependent associations to be taken into account while selecting potential genetic markers influencing the drug response. In Section 4.3, we present simulation study results. Section 4.4 discusses all study findings highlighting their biological significance. Section 4.5 highlights the contributions of this work and study implications. Finally, we summarise the key study findings in Section 4.6. Notice that within the current chapter, symbols for genes are italicised, e.g., *BRAF*.

4.2 Methodology

Below, we introduce the two-stage variable selection algorithm used for selecting genes that are associated to anti-cancer drug dose response in human cancer cell lines. We present in great detail the considered varying coefficient model and each one of the algorithm steps. Parameter tuning and software availability are also discussed.

4.2.1 A two-stage algorithm for identification of gene-drug associations

Non-parametric techniques are a great tool for reducing modelling bias and producing data driven inference. However, flexible modelling techniques applied on high-dimensional genomic data sets can often cause real problems in statistical inference. Sparse regression techniques, such as the Least Absolute Shrinkage and Selection Operator (LASSO), can be used as dimensionality reduction techniques, but cannot handle ultra-high dimensional problems without introducing statistical inaccuracies, algorithmic instability and a huge computational burden (Song et al., 2014). Hence, the need for a feature screening algorithm which will marginally filter unimportant variables becomes essential. Below we further explain the two-stage algorithm that has been built in order to detect and explore dose-dependent gene-response associations.

Let the repeated measures data $\{(d_{ij},\ y_{ij},\ \boldsymbol{z}_i,\ \boldsymbol{x}_i) : j = 1, \ldots, n_i,\ i = 1, \ldots, n\}$, where y_{ij} is the response of the ith experimental unit (corresponds to a drug sensitivity assay of a specific drug on a specific cell line) at the jth drug dosage level d_{ij} and $\boldsymbol{z}_i$ along with $\boldsymbol{x}_i$ are the corresponding vectors of scalar (dose-invariant) covariates. The covariate vector $\boldsymbol{z}_i = (1, z_{i1}, \ldots, z_{ip})^T$ is a low-dimensional vector of predictors that should be included in the model, whereas $\boldsymbol{x}_i = (x_{i1}, x_{i2}, \ldots, x_{iG})^T$ is a high-dimensional vector, i.e., 17,737 gene expression measurements, that needs to be screened. We assumed that only a small number of x-variables (in our case, genes) are truly associated with the response while most of them are expected to be irrelevant (sparsity assumption).

To explore potential dose-varying effects between the covariates and the drug response, we consider the following varying coefficient model:

$$y_{ij} = \sum_{k=0}^{p} z_{ik}\beta_k(d_{ij}) + \sum_{g=1}^{G} x_{ig}\gamma_g(d_{ij}) + \varepsilon_{ij} \tag{4.1}$$

where $\{\beta_k(\cdot), k = 0, \ldots, p\}$ and $\{\gamma_g(\cdot), g = 1, \ldots, G\}$ are smooth functions of dosage level $d \in \mathscr{D}$, where $\mathscr{D}$ is a closed and bounded interval of $\mathbb{R}$. The errors ε_{ij} were assumed to be independent across subjects and potentially dependent within the same subject with conditional mean equal to zero and variance $\text{Var}(\varepsilon) = \sigma^2(d) = V(d)$.

Methods for estimating the coefficient functions in Eq. 4.1 include local and global smoothing methods, such as kernel smoothing, local polynomial smoothing, basis approximation smoothing etc. For computational convenience, in this application we used basis approximation smoothing via B-splines.

Let the sets of basis functions $\{B_{lk}(\cdot) : l = 1, \ldots, L_k\}$ and $\{B'_{lg}(\cdot) : l = 1, \ldots, L_g\}$ and constants $\{\zeta_{lk} : l = 1, \ldots, L_k\}$ and $\{\eta_{lg} : l = 1, \ldots, L_g\}$ where $k = 0, \ldots, p$ and $g = 1, \ldots, G$ such that, $\forall d \in \mathscr{D}$, $\beta_k(d)$ and $\gamma_g(d)$ can be approximated by the expansion

$$\beta_k(\cdot) \approx \sum_{l=1}^{L_k} \zeta_{lk} B_{lk}(\cdot) \ \text{ for } \ k = 0, \ldots, p \tag{4.2}$$

$$\gamma_g(\cdot) \approx \sum_{l=1}^{L_g} \eta_{lg} B'_{lg}(\cdot) \ \text{ for } \ g = 1, \ldots, G. \tag{4.3}$$

Substituting $\beta_k(\cdot)$ and $\gamma_g(\cdot)$ of Eq. 4.1 with Eq. 4.2 and Eq. 4.3, we approximated Eq. 4.1

by

$$y_{ij} \approx \sum_{k=0}^{p} z_{ik} \sum_{l=1}^{L_k} \zeta_{lk} B_{lk}(d_{ij}) + \sum_{g=1}^{G} x_{ig} \sum_{l=1}^{L_g} \eta_{lg} B'_{lg}(d_{ij}) + \varepsilon_{ij} \tag{4.4}$$

If $B_k(\cdot)$ and $B'_g(\cdot)$ are groups of B-spline basis functions of degree q_k and q_g respectively, and $\delta_0 < \delta_1 < \ldots < \delta_{K_k} < \delta_{K_k+1}$ and $\delta_0 < \delta_1 < \ldots < \delta_{K_g} < \delta_{K_g+1}$ are the corresponding knots, then $L_k = K_k + q_k$ and $L_g = K_g + q_g$.

Using the approximation Eq. 4.4, the coefficients $\boldsymbol{\zeta} = (\zeta_0, \zeta_1, \ldots, \zeta_p)^T$ and $\boldsymbol{\eta} = (\eta_1, \eta_2, \ldots, \eta_G)^T$ can be estimated by minimizing the squared error

$$\ell_w((\boldsymbol{\zeta}, \boldsymbol{\eta})^T) = \sum_{i=1}^{n} \sum_{j=1}^{n_i} w_{ij} \left[y_{ij} - \sum_{k=0}^{p} z_{ik} \sum_{l=1}^{K_k} \zeta_{lk} B_{lk}(d_{ij}) - \sum_{g=1}^{G} x_{ig} \sum_{l=1}^{L_g} \eta_{lg} B'_{lg}(d_{ij}) \right]^2 \tag{4.5}$$

where w_{ij} are known non-negative weights.

In cases where $p + G >> n$ though, minimisation of Eq. 4.5 is infeasible. Our aim was to identify factors of the covariate vector $\mathbf{x} = (\mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_G)^T$ (genes) that are truly associated with the response (cancer cell line sensitivity to the drug). In addition, we wanted to explore potential dose-varying effects on the drug response.

We make the following sparsity assumption: any valid solution $\hat{\boldsymbol{\gamma}}(d)$ will have $\hat{\gamma}_g(d) = 0, \forall d \in \mathscr{D}$ for the majority of components g. To detect non-zero coefficient functions, we applied a two-stage approach which incorporated a variable screening step and a further variable selection step.

Screening

The sparsity assumption applies only to components of $\boldsymbol{x}$, the high-dimensional covariate vector in Eq. 4.1.

Let the set of indices

$$\mathscr{M}_0 = \{1 \leq g \leq G : ||\gamma_g(\cdot)||_2 > 0\} \tag{4.6}$$

where $||\cdot||_2$ is the L_2-norm. In order to rank the different components of $\boldsymbol{x}$, we fitted

the marginal non-parametric regression model for the gth x-predictor:

$$y_{ij} \approx \sum_{k=0}^{p} z_{ik} \sum_{l=1}^{K_k} \zeta_{lk}^{(g)} B_{lk}^{(g)}(d_{ij}) + x_{ig} \sum_{l=1}^{L_g} \eta_{lg}^{(g)} B_{lg}^{(g)\prime}(d_{ij}) + \varepsilon_{ij}^{(g)} \tag{4.7}$$

where: $\{B_{lk}^{(g)}(\cdot) : l = 1, \ldots, L_k\}$ and $\{B_{lg}^{(g)\prime}(\cdot) : l = 1, \ldots, L_g\}$ are sets of coefficient functions; $\{\zeta_{lk}^{(g)} : l = 1, \ldots, L_k\}$ and $\{\eta_{lg}^{(g)} : l = 1, \ldots, L_g\}$ are constants to be estimated, $k = 0, \ldots, p$; and, $\varepsilon^{(g)}$ is the error term similar to Eq. 4.4. We then computed the following weighted mean squared error for each $g \in \{1, \ldots, G\}$,

$$\hat{u}_g = \frac{1}{n} \sum_{i=1}^{n} (\mathbf{y}_i - \hat{\mathbf{y}}_i^{(g)})^T \mathbf{W}_i (\mathbf{y}_i - \hat{\mathbf{y}}_i^{(g)}) \tag{4.8}$$

to quantify the importance of the gth variable. Here,

$$\boldsymbol{W}_i = \frac{1}{n_i} \hat{\boldsymbol{V}}_i^{-\frac{1}{2}} \boldsymbol{R}_i^{-1}(\hat{\boldsymbol{\phi}}) \hat{\boldsymbol{V}}_i^{-\frac{1}{2}} \tag{4.9}$$

where $\hat{\boldsymbol{V}}_i$ is the $n_i \times n_i$ diagonal matrix consisting of the dose-varying variance

$$\hat{\boldsymbol{V}}_i = \begin{bmatrix} \hat{V}(d_{i1}) & 0 & \ldots & 0 \\ 0 & \hat{V}(d_{i2}) & \ldots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \ldots & \hat{V}(d_{in_i}) \end{bmatrix} \tag{4.10}$$

and $\boldsymbol{R}_i(\boldsymbol{\phi}) = (R_{jk})$ the $n_i \times n_i$ working correlation matrix for the ith subject. By $\boldsymbol{\phi}$, we denoted the $s \times 1$ vector that fully characterizes the correlation structure. The estimate of $\boldsymbol{\phi}$, $\hat{\boldsymbol{\phi}}$, was obtained by taking the moment estimators for the parameters $\boldsymbol{\phi}$ in the correlation structure based on the residuals obtained from fitting the following model

$$y_{ij} = \sum_{k=0}^{p} z_{ik} \beta_k(d_{ij}) + \varepsilon_{ij} \quad \text{where} \quad i = 1, \ldots, n, j = 1, \ldots, n_i. \tag{4.11}$$

The variance function $V(d)$ in Eq. 4.10 was estimated using techniques described in Chu et al. (2016).

After having obtained $\{\hat{u}_g : \ g = 1, \ldots, G\}$, we sorted gene utilities in an increasing order, where smaller $\hat{u}_g$ values indicate stronger marginal associations. The x-predictors

included in the screened submodel are, then, given by

$$\widehat{\mathscr{M}_{\tau_n}} = \{1 \leq g \leq G : \ \hat{u}_g \text{ ranks among the first } \tau_n(\nu)\} \tag{4.12}$$

where $\tau_n(\nu)$ corresponds to the size of the submodel which is chosen to be smaller than the sample size n, i.e., $\tau_n(\nu) = \nu[\frac{n}{log(n)}]$ with $\nu \in \{1, 2, 3, \ldots\}$.

Variable selection using a group SCAD (gSCAD) penalty

Screening algorithms aim to discard all unimportant variables but tend to be conservative. In order to preserve only the most important x-predictors in the final model, we considered a model including the first $\tau_n(\nu)$ outranked genes and we applied a gSCAD (group Smoothly Clipped Absolute Deviation) penalty by minimising the following criterion:

$$\frac{1}{2}\sum_{i=1}^{n}\sum_{j=1}^{n_i} w_{ij}\Bigg\{y_{ij} - \sum_{k=0}^{p} z_{ik}\sum_{l=1}^{L_k}\zeta_{lk}B_{lk}(d_{ij}) - \tag{4.13}$$

$$\sum_{g\in\widehat{\mathscr{M}_{\tau_n}}} x_{ig}\sum_{l=1}^{L_g}\eta_{lg}B'_{lg}(d_{ij})\Bigg\}^2 + \sum_{g\in\widehat{\mathscr{M}_{\tau_n}}} p_{\lambda,\alpha}(||\boldsymbol{\eta}_g||) \tag{4.14}$$

where

$$p_{\lambda,\alpha}(u) = \begin{cases} \lambda u & \text{if } 0 \leq u \leq \lambda \\ -\frac{(u^2 - 2\alpha\lambda u + \lambda^2)}{2(\alpha-1)} & \text{if } \lambda \leq u \leq \alpha\lambda \\ \frac{(\alpha+1)\lambda^2}{2} & \text{if } u \geq \alpha\lambda \end{cases},$$

α is a scale parameter, λ controls for the penalty size and $||\cdot||$ is the Euclidean ℓ_2-norm. At this point, note that grouping is applied for the coefficients $\boldsymbol{\eta}_g$ that correspond to the same coefficient function. In addition, in order to reduce the bias introduced when applying a LASSO penalty, we alternatively chose the SCAD, which coincides with the LASSO until $u = \lambda$, then transits to a quadratic function until $u = \alpha\lambda$ and then it remains constant $\forall u > \alpha\lambda$, meaning that it retains the penalization and bias rates of the LASSO for small coefficients but at the same time relaxes the rate of penalization as the absolute value of the coefficients increases.

R version 3.6.3 was used to perform all analyses within the current chapter. Both

the simulation study and the data application were performed using The High End Computing Cluster at Lancaster University. R code was developed to implement the marginal screening, whereas R package grpreg was used for the implementation of the gSCAD step. doParallel and foreach R packages were also used to improve computational times of the marginal screening step of the proposed algorithm. Analysis of the GDSC1 data takes approximately five hours to complete. Code for algorithm implementation is available on GitHub as an R package (`https://github.com/koukoulEv/fbioSelect`).

4.2.2 Tuning parameter selection

For the GDSC data application, we used knots placed at the median of the observed data values along with cubic B-splines with 1 interior knot. The suitable number of interior knots was calculated using the formula $N_n = [n^{\frac{1}{2p+3}}]$ proposed and applied by Huang et al. (2004); Xue et al. (2010) and Xue and Qu (2012). Due to the computational burden this would add, we did not apply cross-validation.

As for the screening threshold τ_n, its magnitude could be determined by the fraction $\nu[\frac{n}{log(n)}]$, $\nu \in \{1, 2, 3, \ldots\}$. We conducted a pilot simulation study in order to decide the most appropriate size (for further details see Section 4.3). We also considered an automated algorithm for its selection [Greedy Iterative Non-parametric Independence Screening (Greedy INIS); Fan et al. (2011)]. Finally, the penalty size for the gSCAD step λ was determined using a 5-fold cross-validation.

4.3 Simulation study

Monte Carlo simulations were conducted to examine the ability of our model to detect the genes that are truly associated with the drug response. Due to the computational burden associated with a simulation of the same scale as the original GDSC1 dataset, we conducted a simulation using a smaller simulated dataset. This had a key role in tuning model parameters and simultaneously assessing model goodness-of-fit. Responses over different dosage levels were generated based on a subset of genes, the corresponding low-dimensional GDSC data covariates (drug and cancer type) and some prespecified smooth coefficient functions. In particular, within each iteration, a random sample without replacement of size $n_s = [p_1 n]$ experimental units and $G_s = [p_1 G]$ genes ($p_1 \in [0, 1]$) is

extracted from the original data set. Responses over different dosage levels are then generated based on a subset of G_s, $G_s^A = [p_2 G_s]$, and some low-dimensional covariates from the original GDSC data (drug type and cancer cell line histology):

$$y_{ij}^s = \sum_{k=0}^{p} z_{ik}\beta_k(d_{ij}) + \sum_{g=1}^{G_s^A} x_{ig}\gamma_g(d_{ij}) + \varepsilon_{ij} \;\text{ with }\; i = 1, \ldots, n_s, j = 1, \ldots, n_i \qquad (4.15)$$

where $\{z_{ik} : i = 1, \ldots, n_s, k = 0, \ldots, p\}$ are the dose-invariant low-dimensional covariates and $\{x_{ig} : i = 1, \ldots, n_s, g = 1, \ldots, G_s^A\}$ represent the genetic information truly associated with the drug response (active genes). Coefficient functions $\{\beta_k(d) : k = 0, \ldots, p\}$ and $\{\gamma_g(d) : g = 1, \ldots, G_s^A\}$ are, then defined as follows:

$$\beta_0(d) = 15 + 20\sin\left(\frac{30\pi d}{15}\right),\ \beta_1(d) = 15 + 20\cos\left(\frac{30\pi d}{15}\right)$$

$$\beta_2(d) = 2 - 3\cos\left(\frac{\pi(30d - 25)}{15}\right),\ \beta_3(d) = 2 - 3\sin\left(\frac{\pi(30d - 25)}{15}\right)$$

$$\beta_4(d) = 6 - 0.2(30d)^2,\ \beta_5(d) = -4\frac{(20 - 30d)^3}{2000},\ \beta_6(d) = \sin(0.3\pi) + 0.4d^2$$

$$\beta_7(d) = 3 + \frac{(1-d)^2}{5},\ \beta_8(d) = 0.1\exp\left(-\frac{(1-d)^2}{2}\right),\ \beta_9(d) = 0.1\exp\left(\frac{(1-d)^2}{2}\right)$$

$$\beta_{10}(d) = \sin\left(\frac{\pi d}{15}\right) + \cos\left(\frac{\pi d}{15}\right),\ \beta_{11}(d) = 1 + 52d + 4d^2,\ \beta_{12}(d) = d^3 - 1$$

$$\gamma_g(d) = (d+1)^{\frac{r_g}{100}} \;\text{ where }\; r_g \sim Unif(1, G_s^A),\ \forall g \in G_s^A.$$

For any $g \in G_s^I$, $G_s^I = G_s \setminus G_s^A$ we assume $\gamma_g(d) = 0$. We, also, consider a rational quadratic covariance structure $\text{Cov}(\varepsilon_i(d), \varepsilon_i(s)) = \sigma^2 \frac{1}{1+(\frac{|d-s|}{r})^2}$ and a dose-varying variance $V(d) = \sigma^2 \frac{\sin^2 d}{1+(\frac{\cos d}{10})^2} + \sin\frac{d}{5}$ where $\sigma^2 = 0.001$.

We set $p_1 = 0.05$ resulting in an analytic sample of 190 experimental units and

886 genes. We considered three different scenarios for the number of active genes: $p_2 \in \{0.1, 0.05\}$ and $G_s^A = 1$. We repeatedly sampled experimental units and genes, and generated responses over five or nine dosage levels as described above. The performance of the employed methodology has been assessed based on 1,000 simulations using different screening thresholds $[\tau_n(\nu) = [\frac{n}{log(n)}],\ \tau_n(\nu) = [\frac{2n}{log(n)}]$ and $\tau_n(\nu)$ chosen using the greedy-INIS algorithm (Fan et al., 2011)] and estimated covariance structure scenarios (independence and rational quadratic covariance structure). Cubic B-splines and knots placed at the median of the observed data values have been used for estimating the coefficient functions. To evaluate the performance of the proposed procedure we used the following summary measures: TP−number of genes correctly identified as active; FP−number of the genes incorrectly identified as active; TN−number of the genes correctly identified as inactive; FN−number of the genes incorrectly identified as inactive.

Our method has utility for drug development if it robustly identifies genes truly associated with drug response. Figure 4.1 illustrates gene selection results from simulated drug responses that were generated by varying numbers of gene predictors. Our simulation shows that the sparser the vector of high-dimensional predictors, the better the performance of the algorithm. This result is justified by the presence of sparsity assumption which needs to be satisfied in both stages of the algorithm. As for the screening threshold selection, we observed that higher thresholds demonstrate better algorithm performance. The largest differences were observed when the signal to noise ratio increases whereas the automatic screening threshold selection [Greedy-INIS algorithm; Fan et al. (2011)] seems not to be efficient enough. Small differences were observed when the covariance structure is not correctly specified, however, larger differences are expected when it comes to prediction accuracy. There were a few cases where the employed algorithm failed to identify the truly associated genes; however, this happened to less than 0.01% of the simulations.

To sum up, a screening threshold of size $[\frac{2n}{log(n)}]$ and regression weights adjusted for the covariance structure of the data were identified as the scenario where our method reached its maximum accuracy. Consequently, for the GDSC application, we chose the screening threshold to be the maximum possible, i.e., 923 genes derived from the formula $[\frac{2n}{log(n)}]$, and weights derived by assuming a rational quadratic covariance structure for the repeated measures.

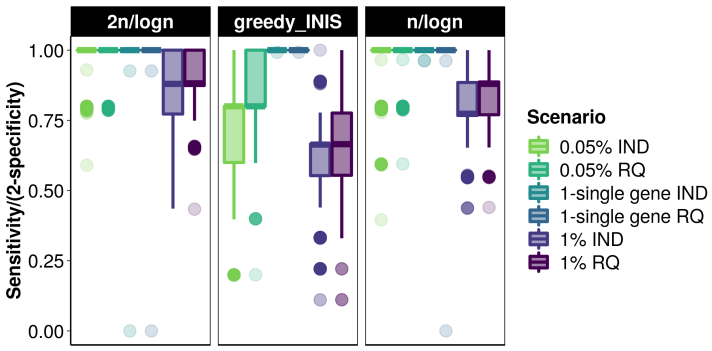


Figure 4.1: Accuracy of detecting drug-gene associations using simulated data under different screening thresholds, covariance structure and G_s^A scenarios. Screening thresholds considered were: $\frac{n}{log(n)}$; $\frac{2n}{log(n)}$, and; a threshold proposed by applying the automated Greedy-INIS algorithm (Fan et al., 2011)−here, n is the number of experimental units in the data. The covariance structure scenarios were independence (IND) and rational quadratic (RQ). The number of active genes scenarios were: either one single gene; 0.5% of the genes in the data, or; 1% of the genes in the data.

4.4 Application to the GDSC data

In the previous section, we presented simulation study results which were critical for algorithm implementation. Below, we present the results and biological interpretation of the findings, along with additional analysis to assess the biological importance and relevance. This section is organised as follows. In the first two paragraphs (4.4.1 and 4.4.2), we present the results after model implementation using the GDSC1 and GDSC2 data (both overviewed in Chapter 3), and we assess the biological significance of our findings by grouping the selected genes into common functional classes or pathways. In the third paragraph (4.4.3), we applied the variable selection algorithm to a data subset containing only cell lines with mutations activating resistance mechanisms to *BRAF* inhibitors to further highlight our method's utility. Finally, in Paragraph 4.4.4, we assess the predictive power of the proposed methodology as opposed to current state-of-the-art methods in pharmacogenetics.

4.4.1 Dose-dependent associations with gene expression in a large-scale drug sensitivity assay

We applied the two stage variable selection algorithm under the dose-varying coefficient model framework described above. Gene rankings and predicted mean drug effects over different dosage levels were obtained. Our algorithm identified 230 candidate genes associated with drug response. The effect of each of those genes was assessed with respect

to:

1. the area under the estimated coefficient curve (AUC) and its corresponding standard deviation (estimated using bootstrapping);
2. the effect on cell survival (overall positive, overall negative, mixed);
3. Spearman correlation between the coefficient function value and the dosage level;
4. the mean fold change of the expression of cell lines carrying *BRAF* mutations with respect to wild type; and,
5. the protein-protein interaction network distance between the *BRAF* gene and the selected genes using the Omnipath database (Türei et al., 2016).

The 230 genes were ranked based on the estimated AUC value (Table C.1), and the top 30 genes were highlighted for further analysis (Table 4.1). The higher the AUC, the larger the effect of the gene on the drug response. The overall effect on cell survival can be either positive, negative or vary over the different dosage levels as determined by the range of the estimated coefficient function. Spearman's rank correlation was used as an indicator of the coefficient function's monotonicity by characterising the progress of the genetic effect over different dosage levels. For instance, high expression of the *C3orf58* gene at baseline has a positive effect on cancer cell survival, which becomes stronger as the dosage increases (Spearman correlation=0.922). In other words, high expression of this gene can be an indicator of drug resistant cell lines. On the other hand, the *DLC1* gene has an overall decreasing and negative effect on cancer cell survival (Spearman correlation=-0.928) which suggested that as the dosage increases, higher baseline expression of this gene can indicate higher drug sensitivity at higher dosage. Elevated expression of *DLC1* has been observed in melanoma and is a well known tumour suppressor that could be a novel marker of *BRAF* inhibition (Yang et al., 2020). Finally, in cases where the overall effect varies (changes between positive and negative), the effect of gene expression on the drug response depends on the drug dosage. In particular, the effect of *DLX6* increases and then decreases at higher dosages (Figure 4.2). Given the biological and technical variation in drug screens, we should treat the mean effect estimates with caution and consider the confidence intervals of the coefficient functions

in order to derive conclusions about the exact effect of the selected genes on the dose response (Figure 4.2).

Gene Name	**Area**	**SD**	**Sign**	**Spearman's Correlation**	**Mean fold change in *BRAF* mutant vs wild-type cell lines**	**Protein-protein interaction network distance to *BRAF***
KIR3DL1	0.370	0.107	-	-0.874	0.978	3
CHST11	0.257	0.092	-	-0.817	0.899	NI
APOC1P1	0.247	0.09	-	-0.918	1.190	NI
PLEKHA6	0.239	0.086	-	-0.908	1.037	3
PPM1F	0.223	0.068	+	0.910	0.883	3
BFSP1	0.222	0.074	-	-0.800	1.217	NI
PPP1R3A	0.217	0.082	+	0.774	1.078	3
C16orf87	0.207	0.087	+	0.851	0.977	NI
PARVA	0.203	0.081	+	0.890	0.984	2
SLC39A13	0.202	0.079	-	-0.461	1.055	NI
UCN2	0.198	0.07	-	-0.928	0.979	NI
STMN3	0.198	0.087	+	0.834	1.201	2
RNF130	0.197	0.083	-	-0.927	1.153	NI
C3orf58	0.196	0.076	+	0.922	1.133	NI
CXXC4	0.188	0.079	+	0.866	0.995	NI
THBD	0.179	0.093	0	-0.967	1.231	4
SIRT3	0.173	0.066	-	-0.760	1.013	3
PLAT	0.172	0.092	-	-0.878	1.322	4
MPPED1	0.168	0.066	+	0.430	0.978	NI
INSL3	0.162	0.068	-	-0.973	0.965	NI
FAM163A	0.159	0.078	-	-0.983	1.106	NI
CNIH3	0.153	0.08	-	-0.918	0.938	NI
GJA3	0.153	0.067	0	-0.940	0.933	NI
BTG2	0.152	0.078	+	0.959	1.035	2
DLX6	0.152	0.059	0	0.686	0.987	NI
DLC1	0.151	0.053	-	-0.928	0.974	3
GAPDHS	0.150	0.077	+	0.886	1.232	NI
JAG2	0.149	0.069	-	-0.994	0.981	3
SMOX	0.146	0.057	0	0.816	1.070	NI
ZMYND8	0.145	0.091	+	0.907	1.020	3

Table 4.1: Gene rankings of the top 30 selected genes based on the magnitude of the genetic effect on drug response. A positive (+) sign translates to a positive effect on cell survival after drug administration, a negative (-) sign translates to a negative effect on cell survival and a mixed (0) effect translates to a varying effect on cell survival which depends on drug dosage. Spearman correlation is calculated between drug dosage and gene estimated coefficient function values as an indicator of the magnitude change of the gene effect over the increasing dosage. Area corresponds to the area under the estimated coefficient curve and the SD corresponds to the standard deviation of the area based on bootstrapping. Mean fold change is calculated between the selected gene expression values of the cell lines carrying *BRAF* mutations with respect to wild type. Protein-protein interaction network distance is computed based on the shortest interaction path between the *BRAF* gene and each of the selected genes. Here, NI denotes absence of any interaction.

Coefficient function estimates provide a lot of information about the dosage, cancer

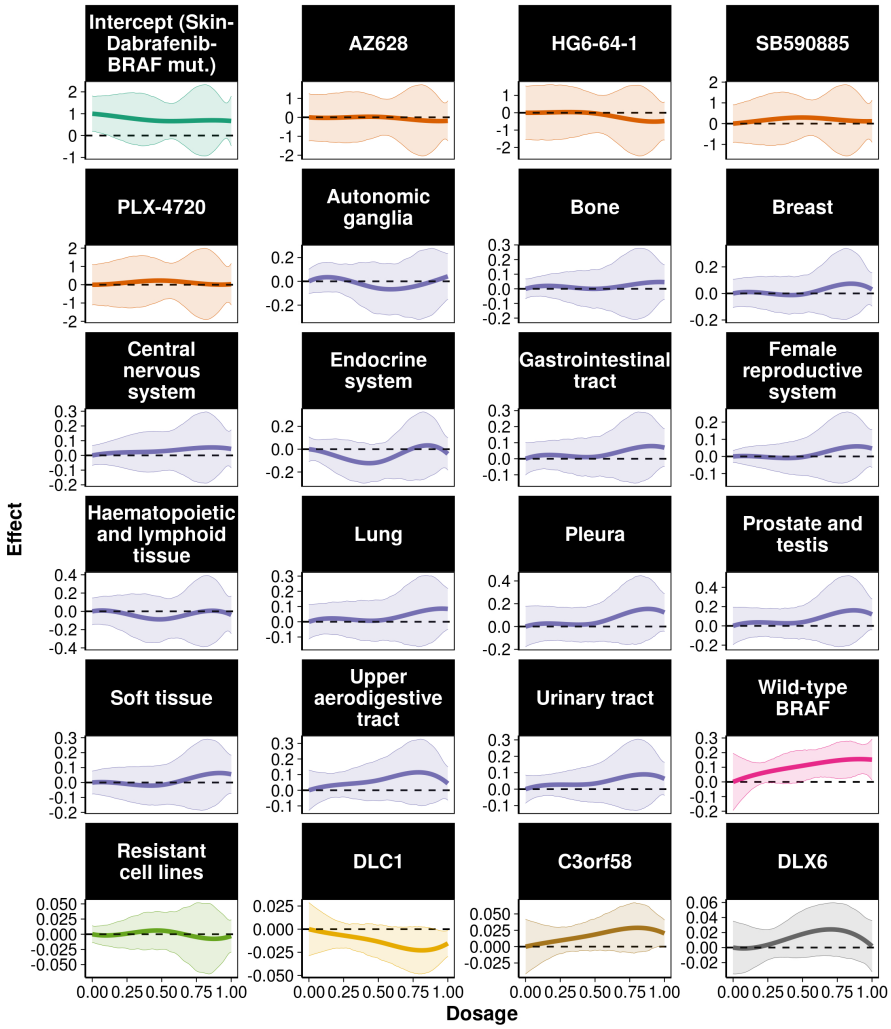


Figure 4.2: Estimated coefficient functions for the low-dimensional predictors and three of the selected genes. Estimated coefficient functions for the intercept, different drugs, tissue of origin and three of the selected genes along with 95% bootstrap confidence intervals. Baseline corresponds to *BRAF* mutant cell lines treated with Dabrafenib in skin tumours.

type and genetic effects on drug response. Figure 4.2 illustrates the estimated coefficient functions for different drugs, cancer types and three genes in relation to the model intercept, Dabrafenib response in *BRAF* mutant cell lines originating from the skin (melanoma). Except from HG6-64-1, all other *BRAF* inhibitors (AZ628, SB590885 and PLX4720) showed no additional effect compared to the intercept. Similar patterns can be observed for cancer cell lines coming from most of the tissues examined. This result indicates that the examined drugs may have similar or worse behaviour over the different dosages for most of the examined cancer types. We observed greater efficacy (negative values of the coefficient function) for cell lines originating from the endocrine system, autonomic ganglia and heamatopoietic and lymphoid tissues at lower dosages.

The observed effect in endocrine system cell lines reflects the Dabrafenib responses observed in anaplastic thyroid cancer patients (Subbiah et al., 2018). Interestingly, the drug Trametinib, taken in combination with Dabrafenib is a MEK inhibitor, and genes interacting with MEK (*MAP2K1*) were selected features from our model (Figure 4.3A). Together these results provide important insights into the effectiveness of the five *BRAF* targeted drugs examined on different cancer types, highlighting the potential for effective treatment of a wide range of cancers given the tumour genetic characteristics.

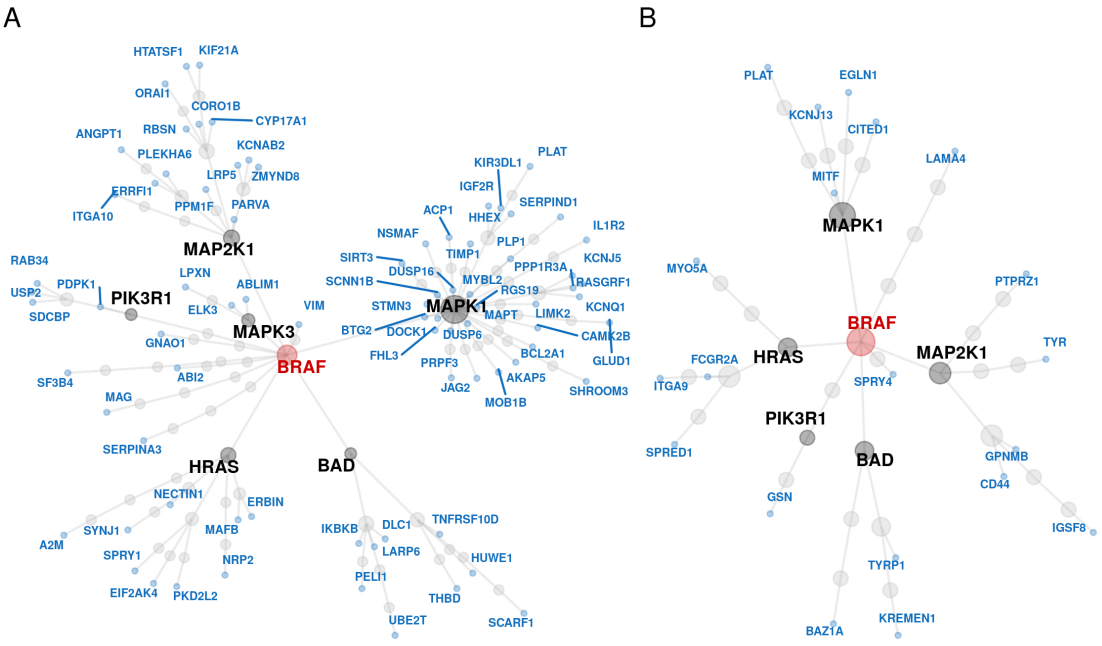


Figure 4.3: Protein-protein interaction network for the genes selected from the two-stage variable selection algorithm. (A) Undirected protein-protein interaction network between the 230 selected (blue) and the *BRAF* (red) genes (full scale analysis). (B) Undirected protein-protein interaction network between the 65 genes selected from the two-stage variable selection algorithm for the cell lines resistant to *BRAF* inhibitors (blue) and the *BRAF* (red) gene. In both panels genes depicted with black are the interaction mediators. Common mediators include the *HRAS*, *MAPK1*, *MAP2K1* and *BAD* genes.

Since the *BRAF* gene is the target of the drugs, mean fold change and protein-protein interaction network distance were used to examine whether and how the selected genes are related to the target of the inhibition. From the selected genes, 120 genes had a mean fold change greater than 1 whereas the rest had a mean fold change between 1 and 0.792. Some of the genes with the highest mean fold change of *BRAF* mutation were *PSMC3IP*, *KIF3C*, *UBE2Q2*, *SERPIND1* and *PLAT*, however only *PLAT* is displayed in Table 4.1. From the genes identified through the two-stage algorithm, 35% of them encode proteins interacting with the *BRAF* gene, though none of them directly. Most

of the selected genes interact with the *BRAF* gene via pathways mediated by *HRAS*, *MAPK1* (*ERK*), *MAP2K1* (*MEK*) and *BAD* (Figure 4.3A).

Since *HRAS* mutations are frequent in patients receiving *BRAF* targeted therapies (Sharma, 2012), we examined the mean estimated trajectory over different dosages under treatment with *BRAF* inhibitors tested in six cancer cell lines with and without *BRAF* and *HRAS* mutations (Figure 4.4). As stated previously, we observed that in most cases HG6-64-1 seems to be the most effective drug. The estimated coefficient functions facilitate drug examination and response prediction under the different dosages. In some instances, we observed different drugs having similar behaviour for lower drug dosages and larger divergence for higher dosages. In most cases, regardless of the cell line origin, our method successfully estimates the expected survival rates of the cancer cell lines for the different drugs given their gene expression information.

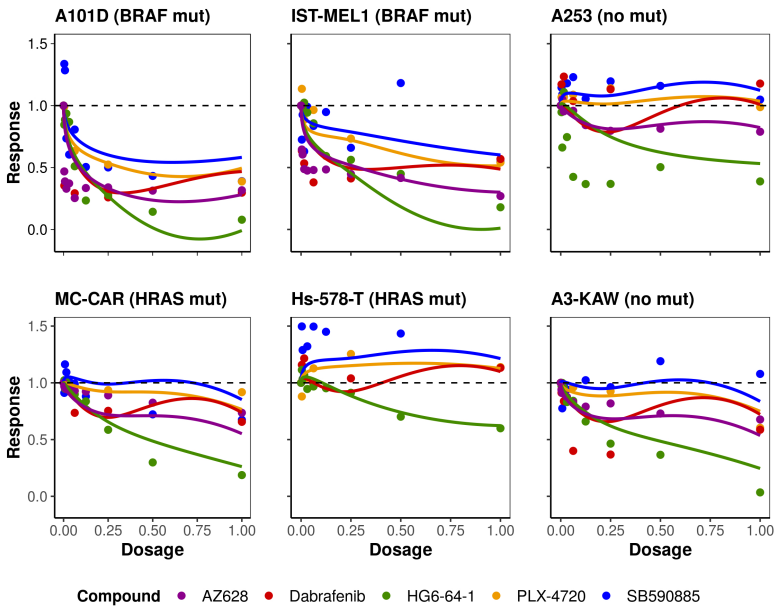


Figure 4.4: Estimated mean drug response trajectories for six cancer cell lines with *BRAF* and *HRAS* mutations. Observed responses (points) and estimated mean trajectory (lines) of cell concentration for cancer cell lines with and without *BRAF* and *HRAS* mutations after treatment with the five anticancer compounds examined.

For validation purposes, we performed the same analysis using the independently-generated GDSC2 data set, with a different set of drugs. Note that the drug set in GDSC2 only partially overlaps with the one used in GDSC1. The results are reported in Figure B.4, and show similar properties to the analysis of the *BRAF* drugs in Figure 4.4. As the GDSC2 dose responses are produced from independently-generated experiments,

the measured drug response is different for some of the drug-cell line combinations. We observe some divergences between GDSC1 and GDCS2 estimated trajectories for Dabrafenib and PLX-4720, which can be explained due to measurement error in the GDSC experiments themselves.

4.4.2 Variable selection algorithm identifies cancer pathways associated with BRAF inhibitor response

Using our functional regression approach, we identified 230 genes that were selected via the SCAD step (observed gene set). We used the Enrichr (Chen et al., 2013; Kuleshov et al., 2016) and WikiPathways (Slenter et al., 2017) databases to see if the selected genes can be grouped into common functional classes or pathways. In total, 183 were identified, of which 11 were statistically significant at 5% level, including apoptosis modulation, NOTCH1 regulation, and MAPK signalling (Table C.2). The model identified genes (*IKBKB*, *RASGRF1*, *DUSP16*, *DUSP8*, *DUSP6*, *MAPT* and *IL1R2*) downstream of the MAPK signalling pathway targeted by *BRAF* inhibitors.

Previous studies of these pathways have found associations with tumourgenesis and cancer treatment (Rangaswami et al., 2006; Sharma and Jha, 2016; Whyte et al., 2009; Mortezaee et al., 2019). Genes in more than one of these pathways include *IKBKB*, *PLAT*, *IL1R2* and *PDPK1*. The IKB kinase composed of IKBKB had previously been suggested as a marker of sensitivity for combination therapy with *BRAF* inhibitors (Colomer et al., 2019). Taken together, these results suggest that the identified associations between the drug response and the observed genes may reveal new predictive markers of tumour response to the examined *BRAF* inhibitors.

In addition to the pathway enrichment analysis, we used the Molecular Signatures Database [MSigDB database v7.0 updated August 2019; Subramanian et al. (2005)] to compute overlaps between the observed gene set and known oncogenic gene sets. Figure 4.5A and Table C.3 display the 29 overlaps found. Interestingly, we identified three instances where the observed gene set significantly overlapped with gene sets over-expressing an oncogenic form of the *KRAS* gene.

We further explored potential biologically relevant pathways using the Reactome database (Wilke, 2012; Jassal et al., 2020). More than 40 enriched pathways were identified at a 5% significance level. The top 40 pathways are depicted in Figure 4.6 along

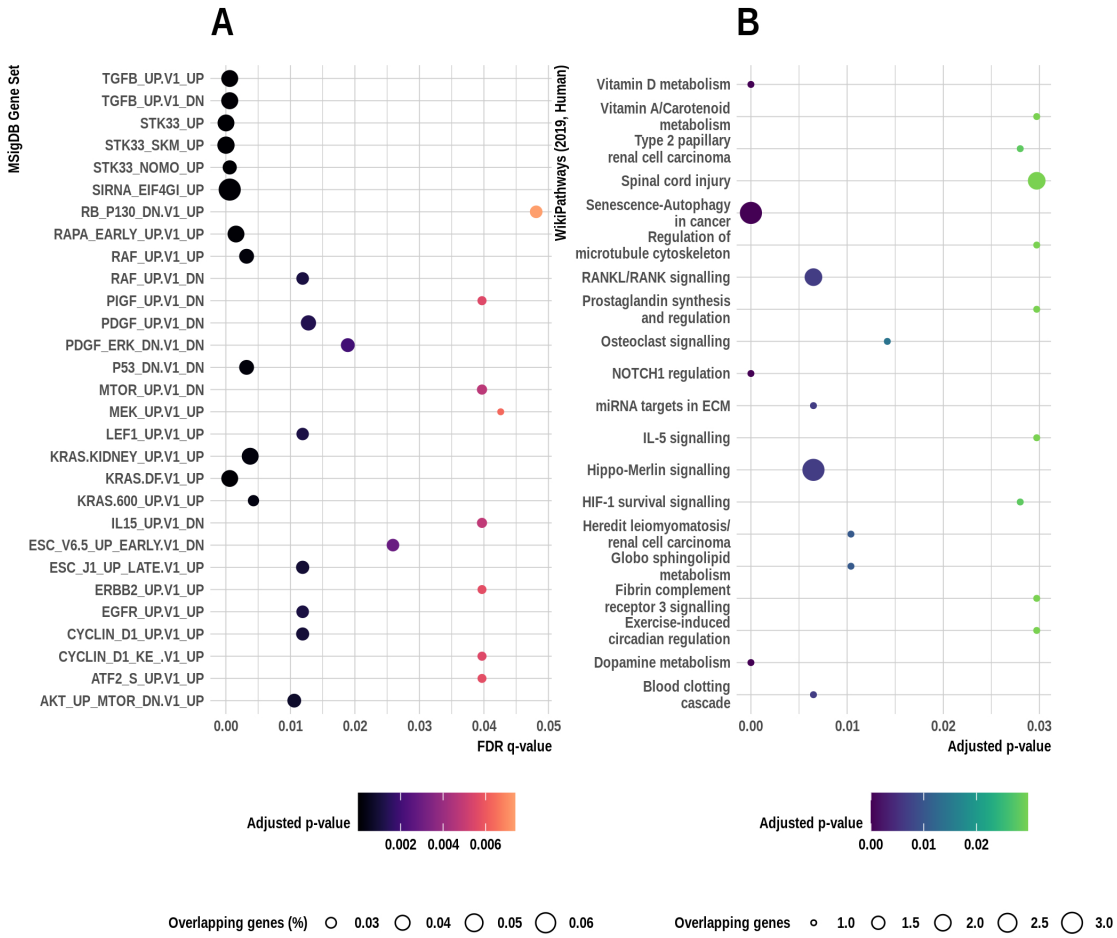


Figure 4.5: Overlaps between the observed gene set and oncogenic signatures in the Molecular Signatures Database (full data analysis); signalling pathways enriched for genes predictive of *BRAF* inhibitor response (resistant cell lines). (A) Full gene set names can be found in Table C.3. Overlaps have been detected using gene set enrichment analysis performed using a hypergeometric distribution. The false discovery rate analog of the hypergeometric p-value is displayed after correction for multiple hypothesis testing according to Benjamini and Hochberg (1995). (B) Top 20 enriched signalling pathways along with the adjusted p-values and the number of overlapping genes obtained after pathway enrichment analysis to the resistant cell line analysis results (for the full list of the pathways identified see Table C.6).

with the pathway-gene network of the top 5 pathways (for the full list, see Table C.4). Interestingly, axon guidance and VEGF signalling were among the enriched pathways, confirming relevance of the selected genes to the intended role of the examined compounds, since *BRAF* kinase activity drives axon growth in the central nervous system (O'Donovan et al., 2014) and VEGF blockade has potential anti-tumour effects when combined with *BRAF* inhibitors (Comunanza et al., 2017). Note that axon growth has not, thus far, been directly implicated in *BRAF* inhibitor activity, and, consequently, our analysis provides important evidence towards this theory.

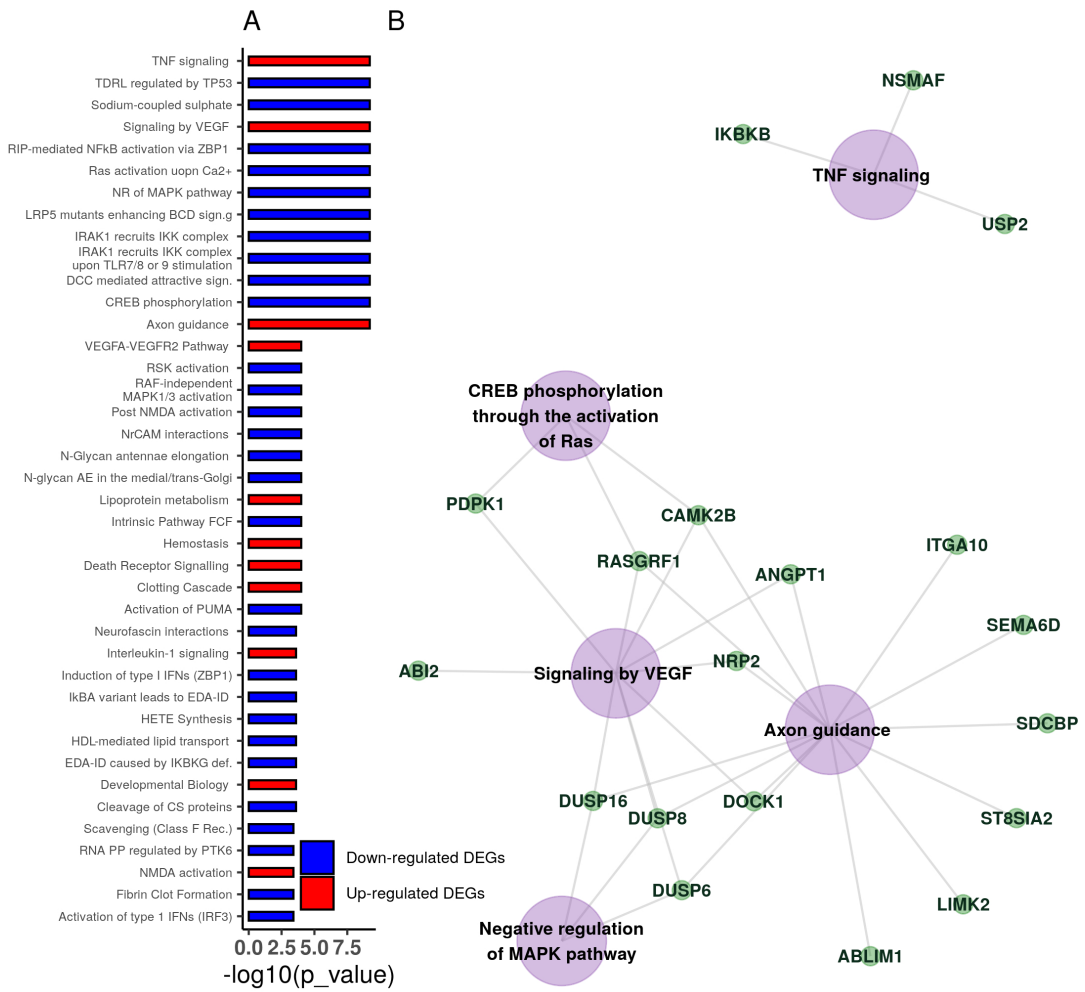


Figure 4.6: Pathway enrichment analysis using the Reactome database. (A) Top 40 enriched signalling pathways along with the adjusted p-values. For the full list, see Table C.4. (B) Pathway-gene network of the top 5 enriched signalling pathways as found using Reactome (Wilke, 2012; Jassal et al., 2020).

4.4.3 Identifying dose-dependent genes in drug-resistance conditions

Acquired resistance to *BRAF* inhibitors is often observed in the clinic (Solit and Rosen, 2011). To further examine the utility of the employed methodology, we applied the variable selection algorithm to a data subset containing only cell lines with mutations activating resistance mechanisms to *BRAF* inhibitors (Manzano et al., 2016). Out of the 951 cell lines in the data, 191 had some mutation in any of the following: *RAC1* gene, *NRAS* gene, *cnaPANCAN44* or *cnaPANCAN315*. We identified 65 genes associated with dose response, though none of them were directly associated with the MAPK/ERK pathway. However, from these, 25 genes have been found to indirectly interact with the *BRAF* gene (Figure 4.3B) and 21 to overlap with three oncogenic gene sets in the Molecular Signatures Database (genes down-regulated in NCI-60 panel of cell lines with

mutated *TP53*; genes up-regulated in Sez-4 cells (T lymphocyte) that were first starved of *IL2* and then stimulated with *IL21*, and; genes down-regulated in mouse fibroblasts over-expressing *E2F1* gene; Table C.5). Finally, we found 34 pathways enriched for genes predicting drug response of the mutated cell lines to the examined *BRAF* inhibitors, of which the top 20 are depicted in Figure 4.5B (for the full list of the pathways identified see Table C.6).

Table 4.2 presents gene rankings based on the AUC and the overall coefficient function effect (sign) for the 42 genes in either the enriched pathways, the three oncogenic gene sets discussed above or the protein-protein interaction network with the *BRAF* gene (full list available in Table C.7). Eight of the selected genes in the current implementation were also selected from the algorithm implemented on the full data: *ASB9*, *PRSS33*, *GJA3*, *PLAT*, *KLF9*, *BFSP1*, *MTARC1* and *UCN2*.

4.4.4 Predictive performance of dose-dependent models

As discussed above, the employed methodology gives a good overview of the baseline genetic effect on drug response. We assessed the overall predictive performance of our method using 10-fold cross validation under two different scenarios. For the first, we split the data into training and test set holding out the experimental units (cancer cell line-drug combinations) and for the second, holding out cancer cell lines. The Mean Absolute Error (MAE) for both cases was around 0.12. Our analysis showed robust cross-validated performance when it comes to predicting sensitivity to the administered drugs, as shown in Figure B.5 which displays the correlation between predicted and true response. This result was further validated by repeating the analysis on the independently-generated GDSC2 data set, using a different set of drugs (Figure B.6), which demonstrated comparable predictive performance.

Predictive accuracy for the dose response curves was evaluated under four different sub-scenarios: prediction of the most effective drug-dosage combination for the 951 cell lines in the data set; prediction of the most effective drug given a cell line; prediction of the most effective dosage given treatment with a particular drug and prediction of the most effective dosage range given treatment with a drug (Table 4.3). The proposed model performs well when it comes to predicting the most effective drug or dosage range ($\approx$79% in both scenarios). Results are less reliable when it comes to prediction of the

Gene Name	**Area**	**SD**	**Sign**	**Spearman Correlation**	**Mean fold change in *BRAF* mutant vs wild-type cell lines**	**Protein-protein interaction network distance to *BRAF***
MYO5A	0.531	0.261	+	0.955	1.358	4
S100A1	0.488	0.189	+	0.812	1.263	NI
GPNMB	0.424	0.196	+	1	1.169	3
ACP5	0.359	0.149	-	-0.998	1.039	NI
FCGR2A	0.341	0.158	-	-0.588	1.25	3
CITED1	0.28	0.348	0	-0.603	1.63	3
SPRY4	0.274	0.127	-	-0.611	1.228	2
CD44	0.239	0.164	+	0.868	1.413	3
RAP2B	0.236	0.179	0	0.927	1.254	NI
KCNJ13	0.205	0.094	0	-0.604	1.101	3
ALX1	0.202	0.099	-	-1	1.104	NI
PLAT	0.201	0.121	-	-0.405	1.312	4
RETSAT	0.201	0.142	0	0.689	1.127	NI
GSN	0.196	0.109	+	0.588	1.079	4
CDH19	0.185	0.102	0	0.943	0.933	NI
ATP1B3	0.178	0.115	-	-1	1.063	NI
BAZ1A	0.173	0.105	+	-0.29	1.109	4
SLC16A4	0.166	0.117	-	-0.298	1.234	NI
ST6GALNAC2	0.164	0.102	0	-0.815	1.264	NI
MFSD12	0.16	0.148	0	-0.788	1.13	NI
GJA3	0.157	0.075	0	-0.85	1.071	NI
CYP27A1	0.156	0.09	-	-0.743	1.373	NI
EGLN1	0.15	0.119	-	-0.442	1.053	3
TRPV2	0.147	0.118	0	0.769	1.074	NI
MITF	0.146	0.106	+	1	0.743	2
TBC1D7	0.146	0.118	0	-0.603	1.304	NI
SLC6A8	0.144	0.111	0	-0.263	0.941	NI
PTPRZ1	0.139	0.138	-	-0.808	1.074	4
PLOD3	0.132	0.135	0	0.696	1.166	NI
ANKRD7	0.131	0.12	+	0.92	1.241	NI
KANK1	0.107	0.113	0	-0.493	1.345	NI
GYPC	0.105	0.092	+	-0.3	1.072	NI
TYR	0.1	0.098	-	0.467	1.11	4
TYRP1	0.1	0.097	0	0.457	1.326	3
IGSF8	0.09	0.129	0	-0.668	1.313	5
SPRED1	0.067	0.116	0	-0.556	1.239	4
ITGA9	0.056	0.111	0	0.785	1.154	4
KREMEN1	0.053	0.086	0	-0.555	1.123	4
LAMA4	0.038	0.083	-	0.344	1.151	4
MLANA	0.037	0.097	0	0.534	1.147	NI
KLF9	0.011	0.074	0	0.932	1.064	NI

Table 4.2: Table notes rankings of the genes found to have some biological importance. A positive (+) sign translates to a positive effect on cell survival after drug administration, a negative (-) sign translates to a negative effect on cell survival and a neutral (0) effect translates to a varying effect on cell survival which depends on drug dosage. Spearman correlation is calculated between drug dosage and gene estimated coefficient function values as an indicator of the magnitude change of the gene effect over the increasing dosage. Area corresponds to the area under the estimated coefficient curve and the SD corresponds to the standard deviation of the area based on bootstrapping. Pearson's correlation is calculated between the selected gene microarray expression values and the *BRAF* expression across all the cell lines. Protein-protein interaction network distance is computed based on the shortest interaction path between the *BRAF* gene and each of the selected genes. Here, NI denotes absence of interaction.

exact dosage or drug-dosage combination (≈48-49% and ≈57-58% in both scenarios) but this can be due to either the large variability observed in the observed responses or due to the small number of cell lines for some predictor level combinations. Results were similar for both cross-validation scenarios (differences range from 0 to <2%, Table 4.3), meaning that as long as a cell line has similar genetic characteristics to those observed, the model can be reliable in predicting the outcome after anticancer drug administration.

Scenario	**Accuracy EU**	**Accuracy CL**
Model predicts the more effective drug-dosage combination	57.85%	57.42%
Model predicts the more effective drug given a cell-line	78.21%	78.21%
Model predicts the more effective dosage given a drug	48.44%	48.65%
Model predicts the most effective dosage range (> or ≤ 31.25% of the maximum dosage)	79.47%	79.28%

Table 4.3: Predictive performance of the employed model (mean absolute error=0.121). Table notes the predictive performance of the model based on the percentages for correctly identifying the most effective drug, dosage or drug-dosage combinations. Results obtained based on 10-fold cross-validation of the final model (based on holding out either experimental units−EU− or cancer cell lines−CL−)

We additionally compared the performance of our two-stage algorithm approach to a penalized linear (LASSO) regression for predicting the IC_{50} and area under the dose response curve (AUC). Note that our functional regression model is not directly predicting either of these values, but rather predicts the full drug response curve. As this is a harder problem, we would expect the LASSO to have a natural advantage; however, our method has the added benefit of being able to detect dose-dependent associations, which is not possible when predicting summary statistics of the dose response curve directly. We employed 10-fold cross-validation to evaluate the predictive error in terms of root mean squared error (RMSE), and we used a sigmoid curve fit for estimating the IC_{50} values from the predicted dose response curves with our two-stage method. Our method outperformed standard LASSO in terms of predicting the AUC ($RMSE_{2-stage} = 0.176$; $RMSE_{LASSO} = 0.347$) and performed well on predicting the IC_{50}, although the LASSO performed better ($RMSE_{2-stage} = 1.969$; $RMSE_{LASSO} = 1.134$). This could be expected, as estimating the IC_{50} from the predicted dose response curve adds a further level of complexity, compared to directly predicting this value using the LASSO.

4.5 Discussion

Genetic alternations and gene expression in tumours are known to affect disease progression and response to treatment. In the current chapter, we studied dosage-dependent associations between gene expression and drug response, using a functional regression approach which adjusts for genetic factors. We analysed data from the GDSC project relating to drug effectiveness for suspending cancer cell proliferation under different dosages, and examined five *BRAF* targeted inhibitors, each applied in a number of common and rare types of cancer cell lines. Our implementation of a two-stage screening algorithm revealed a number of genes that are potentially associated with drug response. Gene, drug and cancer type trajectories have been modeled using a varying coefficient modelling framework. The proposed methodology allows for dose-dependent analysis of genetic associations with drug response data. It enables us to study the effect of different drugs simultaneously, which results in high accuracy of drug response prediction. Drug comparisons using the proposed methodology could support drug repositioning, especially in diseases where existing treatment options are limited. In addition, our methodology can help to reveal unknown potential relationships between genetic characteristics and drug efficacy. Hence, the good predictive performance of our method could be due to the fact that some genes may act as proxies for unmeasured phenotypes that are directly relevant to drug sensitivity.

This work relies on two major assumptions. First, that out of tens of thousands genes regulating protein composition only a small proportion is actually associated with cancer cell survival in a dosage-dependent manner. In other words, transcriptomic profiles exert influence on disease progress after drug administration in a sparse and dynamic way. However, if a large number of genes is associated with the drug response, our method may produce biased results, and some important information about the biological mechanisms can be lost. Secondly, we assume that the different drugs are comparable on the scale of maximum dosage percentage level for our joint model. We acknowledge that different drugs have different chemical structure and maximum screening concentrations. Our focus is to identify genetic components that could be informative for dose response given drugs that belong to a particular family, for example *BRAF* targeted therapies. However, our methodology is flexible enough to allow each drug to be examined separately if it

appears to be clinically appropriate.

Drug response prediction from gene expression data has been widely studied in the literature. Sparse regression methods, gene selection algorithms such as the Ping-pong algorithm (Kutalik et al., 2008), or a combination of network analysis and penalized regression, e.g., the sparse network-regularized partial least squares method (Chen and Zhang, 2016), have all been employed to simultaneously predict drug response and select genetic factors that seem to be associated with the drug response. However, none of these methods are able to quantify the effect of drug dosage on the response, which is one of the main contributions of this work. Employing the proposed dose-varying model gives a detailed picture of different drug effects and can be extremely valuable in predicting drug response for agents with small therapeutic range and high toxicity levels. Our algorithm showed moderate predictive performance due to the complexity of predicting whole drug response curves. Methods for further enhancing the performance of the proposed methodology, such as judicious use of prior information and leveraging information sharing across multiple data sources should be explored in the future in order to overcome this issue and make good use of its full potentials.

4.6 Summary

In this chapter, we develop methodology to examine the dose-dependent associations between genes and drugs. The proposed algorithm uses the raw drug dose response data to infer the effects of interest, handles high-dimensional gene expression data and allows to obtain a more comprehensive picture of the biological mechanisms that undergo cancer treatment and the role of drug dose on it. Due to its simple structure, it, also, permits extension to different types of molecular data (e.g., RNA-seq gene expression, methylation or mutational profiles) and enrichment with further information, such as drug chemical composition.

In the following chapter, we jump to a completely different problem: ageing. We will use data from the ELSA and develop a dynamic conditional quantile modelling framework to model incomplete frailty trajectories and get predictions for ageing individuals and population subgroups.

Chapter 5

Dynamic modelling and prediction of frailty in ageing English adults

5.1 Introduction

In the previous chapter, we developed a regularised functional regression model for complete experimental longitudinal data to model non-stationary dose-varying trends in the presence of high-dimensional data. The current chapter is motivated from ageing research and focuses on incomplete longitudinal survey data where the increased dimensionality[1] lies on the outcome variables. This type of data are characterised by high subject heterogeneity and population diversity. In the current chapter, the multidimensional variables are used to construct one of the most common ageing metrics, the FI. The proposed method represents a major step forward for modelling longitudinal survey data on ageing and predicting frailty trajectories of ageing population subgroups.

Background on frailty is provided in Paragraph 2.3.2. In brief, frailty is a typical ageing surrogate and describes the state of increased vulnerability to ageing-associated diseases. It is a monotonically increasing process over age and remains one of the best predictors of morbidity and mortality (Searle et al., 2008). Advanced age, low or no physical activity, obesity, smoking, low income, low socioeconomic position and

[1] In this context, by "increased dimensionality" we mean tens of outcome variables considered for simultaneous analysis.

lack of education are only a few of the factors that contribute to frailty development and progression (Rogers et al., 2017; Niederstrasser et al., 2019; Franse et al., 2017; Szanton et al., 2010). Contrary to the aforementioned objective indicators, there is little research on how subjective experiences of inequality influence frailty. A collection of past studies have found that focusing on how individuals feel about what they have is equally important to what people have, due to the impact that self-esteem has on individual emotions, behaviour, and even mental and physical health (Smith and Huo, 2014; Mishra and Carleton, 2015). In fact, positive self-perceptions of ageing have been linked to better health at advanced age (Moser et al., 2011; Mendoza-Núñez et al., 2018; Demakakos et al., 2006), whereas feelings of relative deprivation have been associated to poor health and unhappiness (Chen, 2015). In this study, we summarise multivariate questionnaire data from ELSA to a single index to create frailty trajectories and investigate whether and how subjective indicators, in particular relative deprivation and self-perceived social class, affect health at old age and, therefore, frailty. Interested readers can look at Supplementary Text A.1 of the Appendix for more information on the exact variables extracted from the ELSA data.

Longitudinal survey data are the most common resource utilised by social scientists to extract information on person-specific and population circumstances. Common challenges when analysing data coming from these studies include: the mixed types of variables measured; the complex longitudinal trajectories with large variations between individuals; the, sometimes, large amount of missing data or dropouts; and, the fact that some variables may be measured with error (Dean et al., 2009; Blattman et al., 2016; Blackwell et al., 2017; Biemer et al., 2013). However, depending on the outcome of interest and the available information, overcoming those challenges is not always straightforward. To deepen our understanding of frailty and, in general, ageing, we require flexible statistical methodology which not only takes into account population variation and heterogeneity, but deals with the short lengths of observed trajectories relative to the period of interest. Common parametric mean models often overlook the distributional variation and trend complexity over time (Raudenbush and Chan, 1992; Rogers et al., 2017), while parametric quantile regression models are completely determined by the choice of distribution for the random effects parameters, which is often assumed to be a Gaussian distribution (Koenker et al., 1994). Incomplete longitudinal data introduce additional complexity and standard

non-parametric smoothing methods are not suitable since estimation of the covariance function becomes challenging and, under severe sparseness scenarios, even impossible (Yao et al., 2005; Kraus, 2015; Kneip and Liebl, 2020; Liebl and Rameseder, 2019).

A novel approach for limited longitudinal data has been proposed by Dawson and Müller (2018) under the assumption of monotone trajectories. Directly utilizing the functional relation between the trend and its slope, a dynamic conditional quantile modelling framework suitable for experimental longitudinal data without covariates was developed. Motivated by this work, we treat the observed frailty trajectories as being partially observed functional data and assume that they come from realisations of a monotone stochastic process where the instantaneous time where frailty is computed is not necessarily informative for the observations' rate of change. We propose the conditional quantile modelling framework to resolve the problem of missing values in longitudinal survey data and extend it to allow for covariate adjustment. We investigate the performance of the proposed model under varying missing at random scenarios and sample sizes. Model implementation is then performed using the ELSA data. Our objective is to obtain a comprehensive view on the distribution of frailty over time and predict frailty trajectories for individual groups based on their self-perceived social and relative deprivation status. Our results highlight the relationships between frailty progression and socioeconomic inequalities and illustrate the utility of the employed methodology in assessing and detecting ageing adults' needs in a dynamic and personalised manner.

The rest of the chapter is structured as follows. In the following section, we present the ELSA data, describe how the FI is built and develop the extended conditional quantile methodology. Simulation studies in Section 5.3 illustrate the strengths and utility of this methodology under multiple data scenarios. The results from our analysis are presented in Section 5.4, whereas findings, study implications, limitations and future work are discussed in Section 5.5.

5.2 Methodology

This study uses longitudinal data from the ELSA study. A detailed overview of the background, objectives and data of the study has already been presented in Chapter 3.

In this chapter, we use CAPI and NV data from the first 8 waves of the study covering the period 2002 to 2017.

5.2.1 Participants

We consider only individuals recruited in 2002 (ELSA Wave 1). Individuals with missing demographic information as well as proxy respondents have been excluded from the analysis. The total number of individuals available in Wave 1 was 10,602 subjects. However, this number decreases over time due to deaths, patient lost to follow-up and data collection errors.

5.2.2 The Frailty Index (FI)

To build the FI, we followed the methodology presented in Searle (2008). In total, 56 variables were selected based on five inclusion criteria and previous literature (Rogers et al., 2017; Niederstrasser et al., 2019; Searle et al., 2008), and referred to the following categories: mobility and daily life activities' difficulties; self-reported health; depressive symptoms; self-reported health conditions including cardiovascular and chronic diseases, eyesight and hearing capabilities; and cognitive function. A detailed overview of the variables used along with their coding within ELSA and cut-off points can be found at Supplementary Text A.2 of the Appendix. Generally, for binary variables a score of 1 was assigned if a deficit was present and 0 if not. For continuous, ordered or nominal variables, cut-off points between 0 and 1 were assigned describing the degree of health decline present. The frailty score was created using only individuals with a sufficient number of variables (at least 30 out of 56 variables).

5.2.3 Dynamic modelling of the FI scores

Below, we present the conditional quantile methodology introduced by Dawson and Müller (2018) and extend it to allow for the conditional quantiles to be dependent on time-invariant covariates. As shown in previous literature, frailty monotonically increases over time. However, we can argue that time (either chronological age or instantaneous time when frailty is measured) is not always informative of the rate of frailty increase (Clegg et al., 2013). In particular, if the observed longitudinal trajectories are very short compared to the time interval of interest, the available data may not carry

information beyond the local level and slope. The authors argued that, even under these circumstances, if the underlying process is monotonic, one can still obtain trajectory information over time as long as information about local level and slope is known for a subject (Abramson and Müller, 1994; Vittinghoff et al., 1994). This is a major strength compared to other modelling approaches used for the analysis of frailty longitudinal data, e.g. growth curve modelling, since it allows the recovery of the underlying functional and quantile dynamics of frailty by modelling longitudinal measurements which are short relative to the domain of interest (chronological age span which ranges from 50 up to 90 years old). In that manner, the problem of missing data is resolved whilst information of the underlying functional dynamics is maintained.

Let the set of longitudinal data of the scores be denoted by

$$\{(y_{nj}, t_{nj}) : n = 1, \ldots, N, j = 1, \ldots, J_n\} \tag{5.1}$$

where y_{nj} refers to the jth response of the nth subject associated with time $t_{nj} \in \mathscr{T}$. We consider the case where the observed window of time $[t_{n1}, t_{nJ_n}]$ where the longitudinal responses have been collected is too narrow to recover the underlying functional dynamics of the process under study over the interval of interest. By transforming these data to pairs of levels and slopes of the underlying target process (e.g. frailty) for each subject n, we can obtain insights into the probabilistic dynamics of this underlying data generation process.

Model specification

Let us assume that the original observed trajectories in Eq. 5.1 are generated by an underlying $(k+1)$-times continuously differentiable random process $Y : \mathscr{T} \to \mathscr{J}$, where $k \geq 1$ and both $\mathscr{T}$ and $\mathscr{J}$ are closed and bounded subsets of $\mathbb{R}$. We further assume that the measurement times (T) are independent of the value of the process (Y). Individual level and slope data can, then, be generated by obtaining

$$X_n := Y_n(T_n) \text{ and } V_n := Y_n'(T_n) \tag{5.2}$$

at some potentially unobserved random subject-specific time T_n, $n \in \{1, \ldots, N\}$. For $n_1 \neq n_2$ and conditional on the trajectories, $(X_{n_1}, V_{n_1}, T_{n_1})$ have a joint distribution

and are independent of $(X_{n_2}, V_{n_2}, T_{n_2})$. For the rest of this section, we assume that the quantities X_n, V_n are directly observed. If not, then surrogates can be used. For instance, estimates for individual local level and slope can be obtained through least-square line fits to the observed individual trajectories (Dawson and Müller, 2018). Later in this work, this type of surrogates are used to obtain local level and slope information from the observed longitudinal trajectories within ELSA.

Under the aforementioned assumptions, the conditional distribution of slope given the level can be expressed in a way that does not explicitly depend on T. In fact,

$$F(v|x) = P(Y'(T) \leq v | Y(T) = x) = P(Y'(Y^{-1}(X)) \leq v | X = x)\,, \tag{5.3}$$

which holds since the process Y is assumed monotone. This conditional distribution informs about where the subjects generally are headed in the immediate future given that they are currently at a certain level, not only in the mean but in distribution.

For full functional trajectories available, we would aim to estimate the cross-sectional distribution of Y for some amount of time $s \in \mathscr{T}$ after $Y(T) = x$, that is

$$G_s(y|x) = P(Y(T+s) \leq y | Y(T) = x), \tag{5.4}$$

and the cross-sectional α-quantile trajectory

$$q_{\alpha,x}(T+s) = G_s^{-1}(\alpha|x) \tag{5.5}$$

where $0 < \alpha < 1$. However, the estimation of Eq. 5.4 explicitly depends on observing $Y(T+s)$ for all $s \in \mathscr{T}$, therefore, direct estimation is an infeasible target.

Dawson and Müller (2018) introduced the instantaneous α-quantile of the slope for a given level x as

$$\xi_\alpha(x) = F^{-1}(\alpha|x)\,, \tag{5.6}$$

and the longitudinal α-quantile trajectory, $z_{\alpha,x}(T+\cdot)$, as the solution function that

satisfies

$$\frac{dz_{\alpha,x}(T+s)}{ds} = \xi_\alpha(z_{\alpha,x}(T+s)). \tag{5.7}$$

Here, $z_{\alpha,x}$ depends only on α and x, hence, given an estimate for ξ_α, we can obtain an estimate for the full longitudinal α-quantile trajectory by solving the differential equation Eq. 5.6. It is worth mentioning that in most cases $z_{\alpha,x}$ and $q_{\alpha,x}$ do not coincide. However, under the smoothness and uniqueness assumptions of our framework and Proposition 1 in Dawson and Müller (2018), we have that

$$q_{\alpha,x}(T+s) = z_{\alpha,x}(T+s). \tag{5.8}$$

This demonstrates that we can model the conditional quantile trajectories by estimating the conditional distribution of slope given the process local level.

Estimation of the Conditional Distribution and Quantiles (CQ)

In this section, we extend the above framework to allow the rate of change of the process to be dependent on time and other time-invariant covariates. In this regard, we consider the set of longitudinal data introduced in Eq. 5.1 along with some subject specific time-invariant covariates

$$\{(y_{nj}, t_{nj}, \boldsymbol{c}_n) : n = 1, \ldots, N, j = 1, \ldots, J_n\} \tag{5.9}$$

where y_{nj} refers to the jth response of the nth subject, and $\boldsymbol{c}_n$ is a d-dimensional vector of time-invariant covariates (continuous, discrete or both) of subject n.

In order to estimate the α-quantile trajectory, we need information on individual trajectory's slope and level. We represent the observed individual trajectories as pairs of level and slope, (X_n, V_n) for all $n = 1, \ldots, N$. To do so, we regress individual responses $\boldsymbol{y}_n = (y_{n1}, y_{n2}, \ldots, y_{nJ_n})^{\mathrm{T}}$ against time $\boldsymbol{t}_n = (t_{n1}, t_{n2}, \ldots, t_{nJ_n})^{\mathrm{T}}$ using linear least-squares fits. Individuals' slope and level are, then, obtained using the corresponding trajectory slope and responses' mean. Specifically for each subject, we get $v_n = \hat{\beta}_1$ and $x_n = (1/J_n)\sum_{j=1}^{J_n} y_{nj}$ obtained by fitting the model $y_{nj} = \beta_{0n} - \beta_{1n} t_{nj}$. As follows, we

get the set of data points

$$\{(x_n, v_n, \boldsymbol{c}_n) : n = 1, \ldots, N\}. \tag{5.10}$$

We can, now, estimate the conditional distribution of slope given the level and the covariates by kernel smoothing methods with adaptive weights incorporated. As the weights are normally defined on the continuous scale, the smoothing methods need some adjustments to incorporate discrete variables. Specifically, let the covariate vector $\boldsymbol{c}_n = (\boldsymbol{c}_n^o, \boldsymbol{c}_n^m)$ where $\boldsymbol{c}_n^o \in \mathbb{R}^q$ is a q-dimensional continuous random vector, $\boldsymbol{c}_n^m \in \mathbb{R}^r$ is a r-dimensional discrete random vector with both ordinal and nominal variables and $\boldsymbol{c}_n \in \mathbb{R}^d$ where $q + r = d$. The estimator of the conditional distribution of the slope is

$$\hat{F}(v|x, \boldsymbol{c}) = \frac{\sum_{n=1}^{N} H_{h_H}(v - v_n) K_{h_K}(x - x_n) W_{h_W}(\boldsymbol{c}_n^o, \boldsymbol{c}^o) L_{h_L}(\boldsymbol{c}_n^m, \boldsymbol{c}^m)}{\sum_{n=1}^{N} K_{h_K}(x - x_n) W_{h_W}(\boldsymbol{c}_n^o, \boldsymbol{c}^o) L_{h_L}(\boldsymbol{c}_n^m, \boldsymbol{c}^m)} \tag{5.11}$$

where $H(\cdot)$ is a smooth distribution function scaled with a bandwidth h_H, $K(\cdot)$ is a smooth and symmetric kernel function scaled with a bandwidth h_K, in particular, $K_{h_K}(u) = {h_K}^{-1} K(u/h_K)$ with $K(\cdot)$ being a Gaussian kernel, and $W(\cdot)$ along with $L(\cdot)$ are kernel functions for the covariates in the model with variable-specific bandwidths h_W, h_L. Following Li and Racine (2008), we define the kernel

$$W_{h_W}(\boldsymbol{c}_n^o, \boldsymbol{c}^o) = \prod_{s=1}^{q} K_{h_W}(c_s^o - c_{ns}^o) \tag{5.12}$$

where K_{h_W} is a kernel function as defined before and c_{ns}^o is the sth continuous covariate of the nth subject. In addition, for the discrete variables in $\boldsymbol{c}_n$ we define the following kernel function

$$L_{h_L}(\boldsymbol{c}_n^m, \boldsymbol{c}^m) = \prod_{s=1}^{r_1} h_{Ls}^{(|c_{ns}^m - c_s^m|)} \prod_{s=r_1+1}^{r} h_{Ls}^{\mathbb{I}(|c_{ns}^m \neq c_s^m|)} \tag{5.13}$$

where the first kernel product corresponds to the first r_1 ordered variables in $\boldsymbol{c}_n^m$ and the second corresponds to the rest (nominal covariates).

Given estimators $\hat{F}(v|x, \boldsymbol{c})$, an estimate for the instantaneous α-quantile trajectory

ξ_α can be obtained. That is,

$$\hat{\xi}_\alpha(x|\boldsymbol{c}) = \inf\{v : \hat{F}(v|x, \boldsymbol{c}) \geq \alpha\}. \tag{5.14}$$

Subsequently, for the estimation of $z_{\alpha,x,\boldsymbol{c}}(T+t)$, we use $\hat{\xi}_\alpha$ as a plug-in estimate for ξ_α in Eq. 5.14 and then we solve the resulting differential equation using numerical methods, such as the Euler's method, with initial condition $z_{\alpha,x,\boldsymbol{c}}(T+0) = x$.

R version 3.6.3 was used to perform all analyses within the current chapter. The simulation study was performed using The High End Computing Cluster at Lancaster University. R code was developed to implement the CQ model. Analysis of the ELSA data and estimation of the lower, median and upper quantile trajectories takes approximately 20 minutes to complete using a laptop (Intel Core i5-8265U CPU 1.60GHz, RAM 16GiB, Ubuntu 18.04.6 LTS). However, computational time changes relative to the bandwidth choice and number of covariates.

5.2.4 Bandwidth selection

We used the least squares cross-validation method proposed by Li and Racine (2008) to select the kernel function bandwidths. Since there does not exist an automatic data-driven method for optimally determining the bandwidths when estimating the conditional cumulative distribution function, they suggest to use a data-driven least squares cross-validation method which is based on the conditional probability distribution function estimation (Li and Racine, 2008; Hall et al., 2004). This methodology not only suggests optimal bandwidth values but also identifies which variables are needed or not to be included into the final model by over-smoothing of the irrelevant variables.

5.3 Simulation study

To ensure CQ methodology provides valid estimates and explore the method's overall performance under multiple data scenarios, we simulated longitudinal data and investigated estimation accuracy under varying lengths of the observed individual trajectories; estimates' validity in the presence of time-invariant covariates; and, predictive ability when compared to state-of-the-art methods, such as random effects models.

We assessed the degree to which the conditional quantile estimates vary from the true quantile trajectories by simulating longitudinal data from the following exponential process $Y(t) = \exp\{-bt - 1\}$, where b is a random variable and has a Uniform distribution over the interval $[0.3, 0.5]$ and $t \in [0, 8]$. First, a sample of N curves $Y_n(t)$ was generated. Next, for each n, the length of the longitudinal observations' window Δ_n was determined along with the corresponding sampling time T_n, which was drawn uniformly and independently over the interval $[0 + \frac{\Delta_n}{2}, 8 - \frac{\Delta_n}{2}]$. Over the observation window $[T_n - \frac{\Delta_n}{2}, T_n + \frac{\Delta_n}{2}]$, we, then, drew J_n equally spaced measurements $Y_n(t_{n1}), Y_n(t_{n2}), \ldots, Y_n(t_{nJ_n})$, where J_n was an integer in $\{2, 3, \ldots, 8\}$ and depended on the length of Δ_n, i.e. the longer the observation window, the larger the number of repeated measurements observed. We did that to simulate data with a structure similar to what we have observed in ELSA. Finally, independent random noise $\varepsilon \sim N(0, \sigma^2)$ was added to each $Y_n(t_{ij})$ measurement. The performance of the CQ methodology was evaluated under varying noise, sample size and Δ_n scenarios. For the sample size and noise, we considered $N \in \{100, 1000, 5000\}$ and $\sigma \in \{0, 0.001, 0.01\}$ respectively. For the Δ_n, we considered four sub-scenarios (S1-S4). For the first three (S1-S3), $\Delta_n = \Delta$ for all n and $\Delta \in \{0.8, 0.16, 4\}$. For the fourth sub-scenario (S4), we wanted to see how estimation and prediction will be affected if the simulated data have identical observation window lengths to what was observed in ELSA, i.e. 17% of the sampled Δ_n equal to 0.4 (5% of the length of the interval under study), 12% equal to 0.8, 10% equal to 1.2, 10% equal to 1.6, 12% equal to 2.4, 15% equal to 2.8 and 25% equal to 3.2 (40% of the interval under study). All estimation trajectories were conditioned on a starting level of $x = 0.36$ and the considered quantile levels α were $0.10, 0.25, 0.50, 0.75$ and 0.90. We chose $h_H = 0.001$ and $h_K = 0.01$. For each of the above scenarios, we replicated the simulation $R = 1000$ times using different seeds. An illustration of the generated data under the fourth Δ_n scenario, $N = 100$ and varying σ can be found in Figure B.7 of the Appendix.

As a measure of estimation quality, we used the Sum of the Mean Integrated Squared Error (SMISE) across all data replications, that is

$$\text{SMISE} = \sum_{r=1}^{1000} \int_0^8 (z_{\alpha,x}(s) - \hat{z}_{\alpha,x}^{(r)}(s))^2 ds \tag{5.15}$$

where $\hat{z}_{\alpha,x}^{(r)}(s)$ is the estimate of $z_{\alpha,x}(s)$ in the rth replication. Tables 5.1 and 5.2 illustrate

Noise scenario (σ)	Δ	**Sample size (N)**	$\alpha = 0.10$	$\alpha = 0.25$	$\alpha = 0.50$	$\alpha = 0.75$	$\alpha = 0.90$
0.000	0.8	100	1.136	1.074	0.748	0.323	0.205
		1000	0.976	1.160	1.024	0.581	0.273
		5000	0.692	0.937	0.837	0.474	0.288
	1.6	100	4.250	4.132	2.497	1.881	0.659
		1000	4.428	4.352	3.125	1.415	0.907
		5000	3.296	3.265	2.353	1.229	0.919
	4.0	100	28.526	24.924	15.441	11.198	8.159
		1000	26.654	24.144	18.351	11.958	8.164
		5000	27.064	24.800	18.843	11.917	8.811
0.001	0.8	100	3.086	3.393	1.450	0.447	0.189
		1000	0.592	0.754	0.734	0.321	0.265
		5000	0.550	0.798	0.744	0.542	0.390
	1.6	100	5.740	7.196	5.550	2.170	1.144
		1000	4.105	4.121	3.368	1.867	1.048
		5000	3.794	4.114	3.030	1.695	0.961
	4.0	100	40.437	39.620	27.666	17.697	13.403
		1000	33.626	31.950	24.674	16.421	12.042
		5000	25.619	23.547	17.449	11.149	7.650
0.010	0.8	100	2.172	1.823	0.958	3.544	10.830
		1000	1.572	0.774	0.586	3.696	14.359
		5000	1.887	0.820	0.400	4.069	15.171
	1.6	100	4.118	3.229	1.377	1.016	2.569
		1000	2.756	2.838	2.396	2.347	3.404
		5000	2.887	3.166	2.606	2.222	3.647
	4.0	100	23.969	21.268	14.253	8.421	5.958
		1000	29.586	27.096	20.896	14.678	11.377
		5000	22.733	20.605	15.689	10.285	7.582

Table 5.1: SMISE across all replications for the scenarios where $\Delta_n = \Delta \in \{0.8, 0.16, 4\}$. Here, R = 1000.

Noise scenario (σ)	**Sample size (N)**	$\alpha = 0.10$	$\alpha = 0.25$	$\alpha = 0.50$	$\alpha = 0.75$	$\alpha = 0.90$
0.000	100	0.242	0.261	0.287	0.349	0.493
	1000	0.042	0.071	0.144	0.129	0.130
	5000	0.038	0.064	0.143	0.141	0.142
0.001	100	0.681	0.329	1.225	4.162	16.628
	1000	0.411	0.075	0.134	0.462	1.974
	5000	0.397	0.075	0.146	0.494	1.767
0.010	100	43.920	63.932	134.132	260.942	284.034
	1000	30.547	50.491	109.869	263.497	449.031
	5000	15.327	18.719	80.845	304.990	528.119

Table 5.2: SMISE across all replications for the scenario where the simulated data have the same observation window lengths to what was observed in ELSA. Here, R = 1000.

that violations of the length of longitudinal trajectories relative to the interval of interest do not significantly affect the method's robustness, even though estimates become slightly less accurate. Notice that for $\Delta = 4$, i.e. half of the overall interval length, the method's performance significantly worsens. This is possibly due to model misspecification when the exponential (individual) trajectories are summarised via linear least squares fits to obtain the level and slope data. Nevertheless, results suggest that, even though performance is worse, the model still manages to capture the underlying functional dynamics of the process. Another factor that has a detrimental effect on method's accuracy is the noise level and mostly for the upper quantile trajectories. That is because for highly noisy data the monotonicity of the upper quantile curves is violated creating ceiling effects for the corresponding estimates. Increasing sample size has a positive impact on accuracy, especially for the lower quantiles and the scenario with mixed lengths of Δ_n. This boosts our confidence that the obtained estimates after fitting the model to the ELSA data will be robust even in cases where some individual trajectories cover more than 25% of the interval of interest.

The main methodological contribution of this work is the extension of the CQ methodology to allow for covariate-adjustment. We explored estimates' validity in the presence of time-invariant explanatory variables of mixed types. Three sub-scenarios of the initial data generation process where the introduced covariates had a multiplicative effect on the overall process dynamics were considered (S5-S7). For the first (S5), we assumed that $Y(t) = \exp\{-bt - 1 + 0.5\mathbf{1}_{x=1}\}$ and $x \sim \text{Bernoulli}(\frac{1}{2})$. For the second (S6) and third (S7) sub-scenarios, we assumed $Y(t) = \exp\{-bt - 1 + 0.5x\}$ where $x \sim B(3, \frac{1}{3})$ and $x \in [0, 5]$ respectively. In all three, b is considered a Uniform random variable over the interval $[0.3, 0.5]$ as before. Individual observation windows and measurements were sampled similar to S4. Here, performance was measured using the mean integrated squared error (MISE) across 1000 data replications and compared over multiple sample size scenarios, $N \in \{100, 1000, 5000\}$. Notice that

$$\text{MISE} = \frac{1}{1000} \sum_{r=1}^{1000} \int_0^8 (z_{\alpha,x}(s) - \hat{z}_{\alpha,x}^{(r)}(s))^2 ds. \tag{5.16}$$

The noise level was assumed to be fixed at $\sigma = 0.01$ and the bandwidths were selected using the cross-validation methodology presented in Section 5.2.4. Results show that

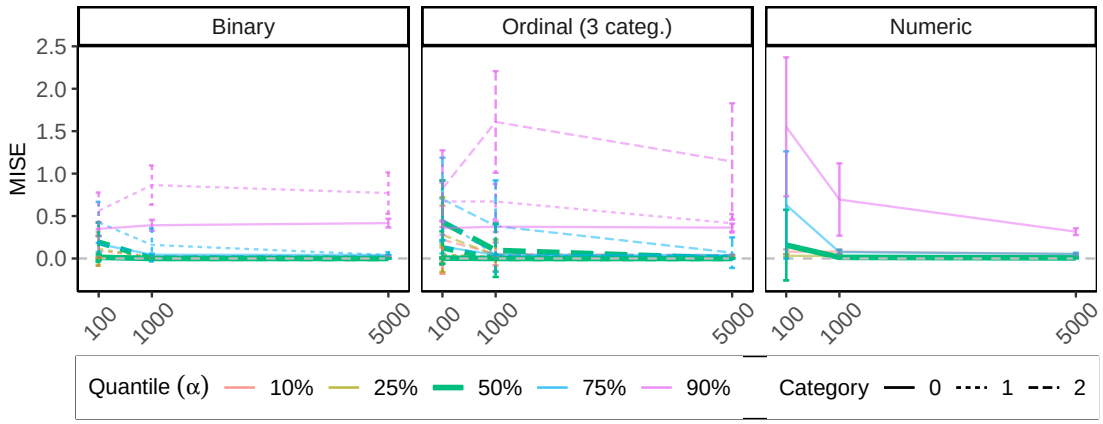


Figure 5.1: CQ methodology performance when a multiplicative covariate effect is assumed under multiple sample size scenarios. The mean integrated squared error (MISE) was calculated for α equal to 0.10, 0.25, 0.50, 0.75 and 0.90. Noise level was assumed to be fixed and high at 0.01. Covariate bandwidth was determined using cross-validation as described in Section 5.2.4. In particular, for the binary variable bandwidth was chosen to be equal to 0.7, for the ordinal variable bandwidth was chosen to be equal to 0.9 and for the numeric that was 4.5. For the numeric covariate scenario, estimates were obtained after having conditioned on $x = 2.5$. The number of data replications was R = 1000.

valid quantile estimates are produced independent of covariate type and large sample sizes promote quantile trajectory estimation accuracy for all quantiles (Figure 5.1). In particular, for the 50% quantile the average MISE across all data replications is found to be between 4.30×10^{-3} and 4.30×10^{-1} for all covariate types and $N = 100$; and, between 3.21×10^{-4} and 6.76×10^{-3} for all covariate types and $N = 5000$. Due to the high noise assumed ceiling effects are observed for the upper quantile estimates which decrease the overall 90% quantile estimation accuracy for all sample size and covariate type scenarios.

Finally, 10-fold cross validation was implemented to compare how the CQ method competes to state-of-the-art methods utilised for longitudinal data analysis in social sciences when it comes to predictive ability. In particular, we compared the CQ method to linear mixed effects (LME) and multilevel quadratic growth curve (QGC) modelling. Three sub-scenarios were considered. For the first (S8), the data generation process was similar to S4, consequently both LME and QGC were misspecified. In these models, apart from the fixed effects, a random intercept and random slope were assumed. A log transformation was, also, used before fitting the LME. For the second (S9), longitudinal trajectories were produced assuming a linear mixed effects model, i.e. $Y_{\text{linear}}(t) = -b + (a + 5.2)t$, where $a \sim N(0, 1)$, $b \sim N(0, 1.5)$ and a independent of b. For the last

(S10), $Y_{\text{quadratic}}(t) = -b + (a + 3.2)t + 0.5t^2$, with , where $a \sim N(0, 1)$, $b \sim N(0, 0.5)$ and a independent of b. In all above scenarios, individual measurements were drawn according to S4 and bandwidths were selected using cross-validation. For each fold, the training set was selected by randomly sampling 90% of the data, leaving the rest 10% for testing. Performance was assessed using the MAE under multiple noise and sample size scenarios. Due to the computational burden involved N ranged from 100 up to 1000. When the linear and quadratic models were misspecified, results showed that the CQ method outperformed the state-of-the-art methods and produced high quality predictions, especially in scenarios where noise is high (Figure 5.2). That is possibly due to the use of least squares fits to individual trajectories which can get rid of some of the noise present. On the contrary, when the underlying models were correctly specified our method shows worse predictive ability. This is expected since both linear mixed effects and quadratic growth curve models can effectively deal with missing at random data problems and, when it comes to prediction, they perform competently.

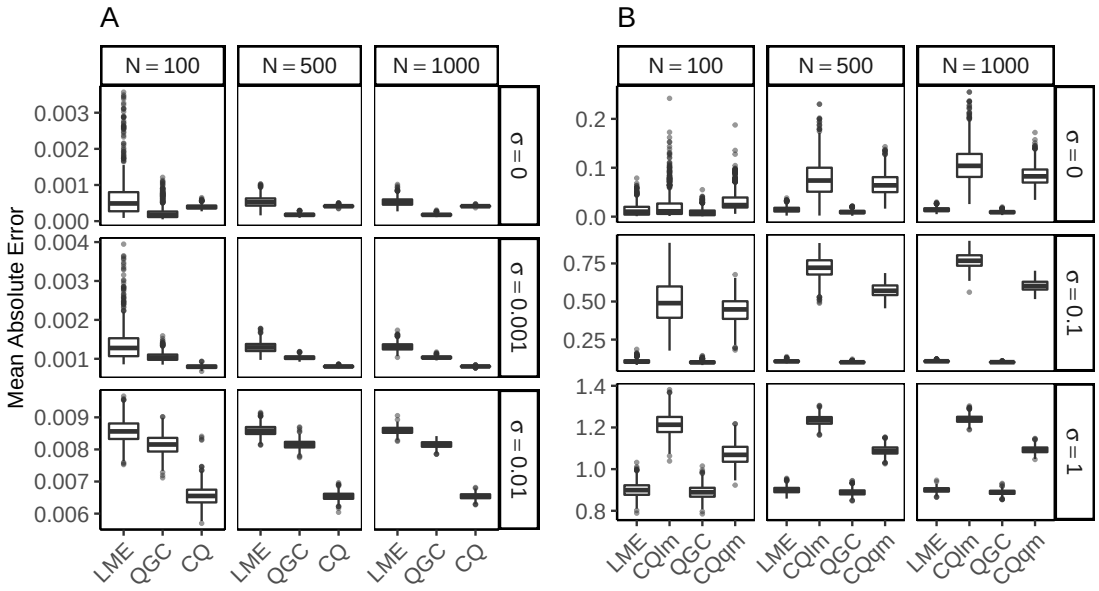


Figure 5.2: Prediction MAE for multiple sample size and noise scenarios for the simulated data were the length of trajectories are determined based on the length trajectories observed in the ELSA data. The number of replications was R= 1000. The left panels (A) correspond to S8 results (misspecified model for LME and QGC) whereas the right panels (B) correspond to S9 and S10 results (correctly specified LME and QGC models). LME stands for linear mixed-effects model, QGC stands for quadratic growth curve model, CQ stands for conditional quantile methodology and CQlm and CQqm correspond to conditional quantile regression results for data generated through S9 and S10 respectively. Under the CQ methodology, the estimated quantile for each individual was used to obtain individual predictions, i.e., the quantile on which we estimate that the subject is travelling.

5.4 Application to the ELSA data

We defined the frail subjects as those having a baseline FI larger than 0.25 (Searle et al., 2008; Rogers et al., 2017). We analysed data from the non-frail at baseline individuals to avoid introducing additional survival bias into our results. That is because adults who were frail at baseline had a smaller number of repeated observations overall (median number of repeated measurements were 3 compared to 6 for the non-frail group; Figure B.8) and were expected not to have survived long enough to have complete longitudinal trajectories. We also omitted individuals who participated in only a single wave, resulting in an analytic sample of 6,563 individuals. An overview of the sample at wave 1 and how the balance between individual subgroups evolved across later waves is shown in Table 5.3. Age group, gender, whether in couple, self-perceived social class and self-perceived relative deprivation are all considered as reported at baseline. An average FI increase of 0.05 is observed over the 16 year period (mean FI of 0.09 at wave 1 versus a mean FI of 0.14 at wave 8). Younger people (those in the 50-59 group at baseline) remained in the study for longer and the balance between males and females was maintained. The majority of individuals reported themselves as being in couple, middle class and non-deprived at baseline. No major changes were observed in the following years with fairly similar proportions of individuals carrying on with the study from all baseline subgroups.

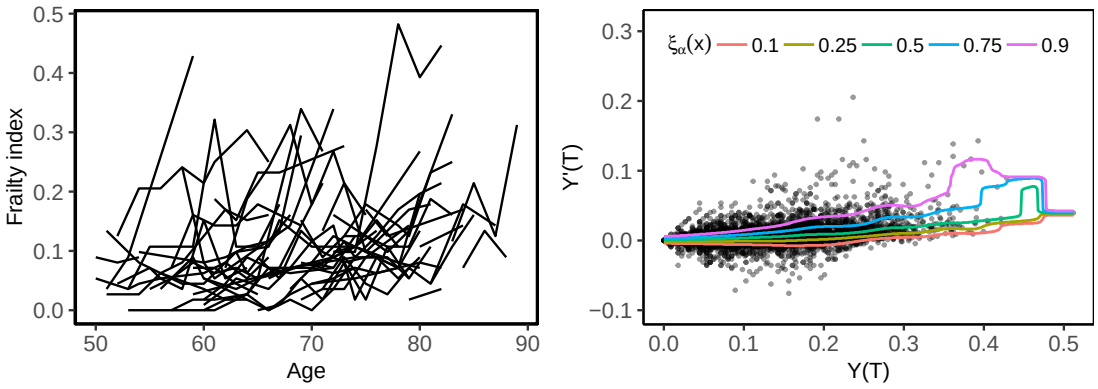


Figure 5.3: Frailty trajectories for a sample of 100 non-frail at baseline individuals (left) and corresponding estimates of $\xi_\alpha(x)$ for $\alpha \in \{0.10,\ 0.25,\ 0.50,\ 0.75,\ 0.90\}$ (bottom to top) using the rule-of-thumb bandwidths $h_H = 2.38 \times 10^{-3}$ and $h_K = 1.68 \times 10^{-2}$ (right).

Wave	1 (N = 6,563)	2 (N = 5,941)	3 (N = 4,929)	4 (N = 4,456)	5 (N = 4,318)	6 (N = 3,817)	7 (N = 3,354)	8 (N = 3,044)
FI								
Min	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
Median (IQR)	0.07 (0.04, 0.12)	0.08 (0.04, 0.14)	0.08 (0.04, 0.14)	0.09 (0.04, 0.15)	0.09 (0.05, 0.16)	0.09 (0.04, 0.16)	0.10 (0.05, 0.17)	0.11 (0.07, 0.18)
Mean ± SD	0.09 ± 0.06	0.10 ± 0.08	0.11 ± 0.09	0.11 ± 0.09	0.12 ± 0.09	0.12 ± 0.09	0.12 ± 0.09	0.14 ± 0.10
Maximum	0.25	0.61	0.61	0.60	0.59	0.71	0.64	0.61
Age Group								
50-59	2,767 (42%)	2,497 (42%)	2,136 (43%)	1,966 (44%)	2,017 (47%)	1,885 (49%)	1,741 (52%)	1,660 (55%)
60-69	2,157 (33%)	1,969 (33%)	1,643 (33%)	1,530 (34%)	1,533 (36%)	1,365 (36)	1,222 (36%)	1,102 (36%)
>70	1,639 (25%)	1,475 (25%)	1,150 (23%)	960 (22%)	768 (18%)	567 (15%)	391 (12%)	282 (9%)
Gender								
Male	3,061 (47%)	2,762 (46%)	2,285 (46%)	2,049 (46%)	1,968 (46%)	1,722 (45%)	1,505 (45%)	1,344 (44%)
Female	3,502 (53%)	3,179 (54%)	2,644 (54%)	2,407 (54%)	2,350 (54%)	2,095 (55%)	1,849 (55%)	1,700 (56%)
In couple								
No	1,664 (25%)	1,478 (25%)	1,246 (25%)	1,075 (24%)	1,020 (24%)	868 (23%)	759 (23%)	662 (22%)
Yes	4,899 (75%)	4,463 (75%)	3,683 (75%)	3,381 (76%)	3,298 (76%)	2,949 (77%)	2,595 (77%)	2,382 (78%)
Social status*								
Lower	1,132 (17%)	993 (17%)	805 (16%)	702 (16%)	676 (16%)	585 (15%)	505 (15%)	437 (14%)
Middle	2,994 (46%)	2,712 (46%)	2,234 (45%)	2,032 (46%)	1,930 (45%)	1,686 (44%)	1,461 (44%)	1,326 (44%)
Upper	2,437 (37%)	2,236 (38%)	1,890 (38%)	1,722 (39%)	1,712 (40%)	1,546 (41%)	1,388 (41%)	1,281 (42%)
Deprivation status*								
Deprived	1,594 (24%)	1,396 (24%)	1,198 (24%)	1,048 (24%)	1,027 (24%)	924 (24%)	783 (23%)	730 (24%)
Non deprived	4,964 (76%)	4,540 (76%)	3,731 (76%)	3,408 (76%)	3,291 (76%)	2,893 (76%)	2,571 (77%)	2,314 (76%)

Table 5.3: Sample overview. Age group, gender, whether in couple, social status and deprivation status are all shown as they reported at baseline. *Social status refers to self-perceived social class and deprivation status refers to the self-perceived relative deprivation.

Frailty is described as a biological process which monotonically increases with age through the accumulation of age-related deficits in health (Searle et al., 2008). However, there is considerable individual heterogeneity in frailty trajectories (Figure 5.3) and, as it has been shown in past studies, chronological age is not their only determinant (Niederstrasser et al., 2019; Clegg et al., 2013). Before proceeding with the CQM implementation, we performed exploratory analysis to confirm the age dependence hypothesis and explore whether our assumption regarding the association of socioeconomic self-evaluation and frailty holds.

Least-square fits on each subject's measurements were used to compute the local frailty level X_n and slope V_n for all n (Figure 5.3). Exploratory analysis showed that both level and local slope change significantly over chronological age suggesting that dependence on age should be accommodated into the final model. We found that adjusting for individual's age at baseline significantly improved model fit when regressing slope on level (F-statistic: 62.8 and ; p-value equal to < 0.001). We also segmented the data set according to frailty level and mean age and compared the distribution of slopes V_n for different ages. Evidence indicated that the slope is dependent on chronological age at baseline and needs to be adjusted into the model (Figure B.9; Kolmogorov-Smirnov test p-value ranged between $< 10^{-8}$ and 1.6×10^{-3}).

We assessed the relationship of the slope with two socioeconomic variables after having conditioned on level and age group at baseline; these included baseline self-perceived social class and baseline self-perceived relative deprivation. Exploratory data analysis showed that differences between the rate of change of frailty across social class and deprivation groups were weak once conditioned on age group, even though level differed for all subgroups (Figure 5.4). Association of the rate of change of frailty (slope) with all aforementioned variables (level, baseline age group, baseline social class and baseline deprivation status) was confirmed through Kolmogorov-Smirnov tests and the automatic cross-validation method discussed in Section 5.2.4, so we proceeded with their inclusion in all forthcoming analysis. We also explored quantile estimates' sensitivity under undersmoothing or oversmoothing conditions. Analysis showed that especially for the middle quantiles even a 100% change in the bandwidth magnitude results in minor changes to the resulted estimates, with level bandwidth determination being crucial to model fit (Figure B.10). On top of that, middle and upper quantiles appeared more robust

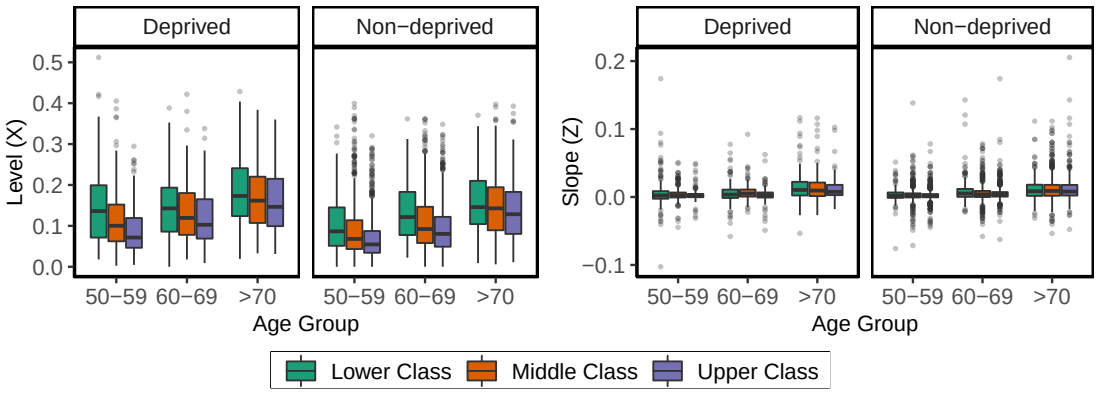


Figure 5.4: Boxplots of FI level for individual subgroups based on baseline age group, social class and deprivation (left). Boxplots of FI slope for individual subgroups based on baseline age group, social class and deprivation (right).

compared to the quantiles towards the lower tail of the distribution. Undersmoothing also led to larger changes compared to oversmoothing for most of the covariates.

Figure 5.5 shows the estimated lower, middle and upper population conditional quantile trajectories for different subgroups of individuals until they become frail given that they started at the median frailty level for their age group at baseline (left panels). Pointwise 95% bootstrap confidence intervals confirmed our prior finding that differences between the social class and deprivation subgroups are small after having conditioned on level and age group. Overall, the median FI of older people was larger compared to the younger groups ($\approx$0.15 for those aged over 70 years old) and conditional on this FI level, frailty (a score over 0.25) is expected to be reached within 10 years for those in the middle quantile while no significant difference was found across the different social class and deprivation subgroups. FI acceleration rates change for those at younger age groups, with people in the 60-69 group at baseline being expected to reach frailty within 25 years conditional to their median FI level ($\approx$ 0.11) and people in the 50-59 group being expected to reach frailty in double this time (middle quantiles). On the contrary, smaller differences were found between the 60-69 and 50-59 age groups for the upper and lower quantiles.

In the right panels of Figure 5.5, we plotted the difference between the estimated 50% quantile FI trajectories for the non-deprived versus deprived individuals along with 95% confidence intervals. Feeling non-deprived was not found to be protective against reaching frailty soon in life. In addition, people in the upper social class are expected to reach frailty faster compared to the lower class group. Bootstrap confidence intervals showed

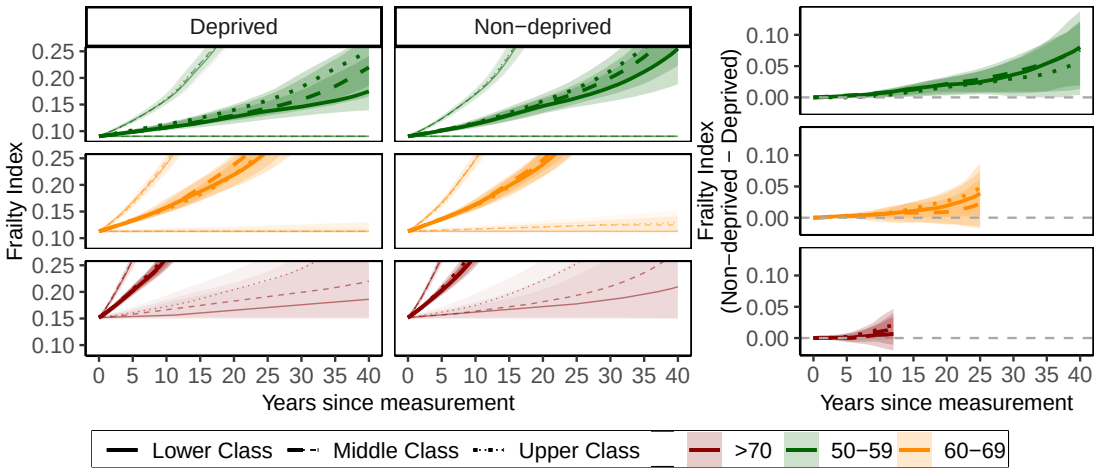


Figure 5.5: The left panels show the estimated quantile trajectories conditioned on the median per age group level along with 95% pointwise bootstrap confidence intervals. For the estimation, we used the cross-validation bandwidths: $h_H = 3.3\times10^{-4}$, $h_K = 1.7\times10^{-2}$ and $h_L = \{7.5\times10^{-2}, 4.4\times10^{-1}, 3\times10^{-1}\}$ for the age group, social status and deprivation variables. From top to bottom, α varies over 0.75, 0.50 (middle) and 0.25. The right panels show the difference in trajectories between non-deprived and deprived groups for all age-group and social class subgroups, along with 95% pointwise bootstrap confidence intervals for the difference.

that these differences are not strong and are expected to occur at later time points where estimation is affected by the number of people surviving (or completing the study) up to this point. These results are counter-intuitive and should be treated with caution since the amount of data for specific subgroups of individuals, etc. deprived participants at lower class over 70 years old, was not large. At the same time, for the upper and lower quantiles as well as later time points uncertainty increases which is supported from the illustrated confidence intervals. As discussed in Dawson and Müller (2018), looking at the median trajectory helps in extracting more robust conclusions, however looking at the difference between the lower and upper quantiles allows conclusions to be drawn for the FI population distribution over time. We further explored the effect of self-perceived social class and deprivation by allowing for interaction effects with age group, however the effect was not significant (Figure B.11).

The left panels of Figure 5.6 provide information on the severity of a given subject's trajectory. In particular, we randomly selected 6 middle class individuals with two FI measurements and plotted their observed trajectories over a range of estimated quantile trajectories conditional on their first observation and remaining characteristics. By comparing the observed individual trajectory to the estimated $z_{\alpha,x}$ values we can assess

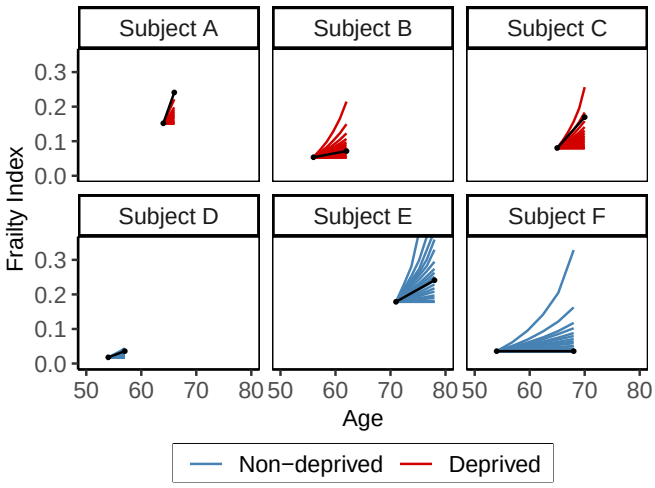


Figure 5.6: A subset of 6 individual trajectories from middle class participants (half deprived, half non-deprived) and their corresponding $z_{\alpha,x}$ trajectories starting from the first observation point for each subject. The quantile α ranges from 0.05 (bottom) to 0.95 (top) and $J_n = 2$ for all subjects illustrated. The dots are the observed FI values.

how severe the condition of an individual is. For instance, subjects F and B are in relatively good condition compared to the rest of the population since their observed FI measurements coincide with the lowest estimated quantile trajectories. On the other hand, subjects A and C from the deprived group are in a more severe condition (observed trajectory coincide to the estimated upper quantiles), suggesting that they might be in need of additional support sooner than expected.

Predictions of an individual's future trajectories can be obtained by estimating the conditional quantile trajectory starting from the subject's fitted value at the last time point of measurement. In this way, we can get not only a spectrum of future scenarios but also an estimate of where the subject is headed based on his/her observed slope and last fitted FI measurement. In essence, after having determined the quantile on which the subject is travelling α^* by comparing the subject's slope to the estimated conditional distribution $\hat{F}(v|y_{i,n_i})$, we estimate a subject's future trajectories by employing the quantile trajectories for a wide range of α values under the constraint that early in the prediction, each α must be close to α^*. We refer interested readers to Dawson and Müller (2018) for more details on how predictions can be achieved if no further covariates apart from local level are included into the model. Figure 5.7 shows the observed trajectories and future scenarios for 18 non-deprived individuals of lower and upper social class. Those who did not experience a significant increase of their frailty score in the past are expected to follow an optimistic scenario and there is less uncertainty around their frailty

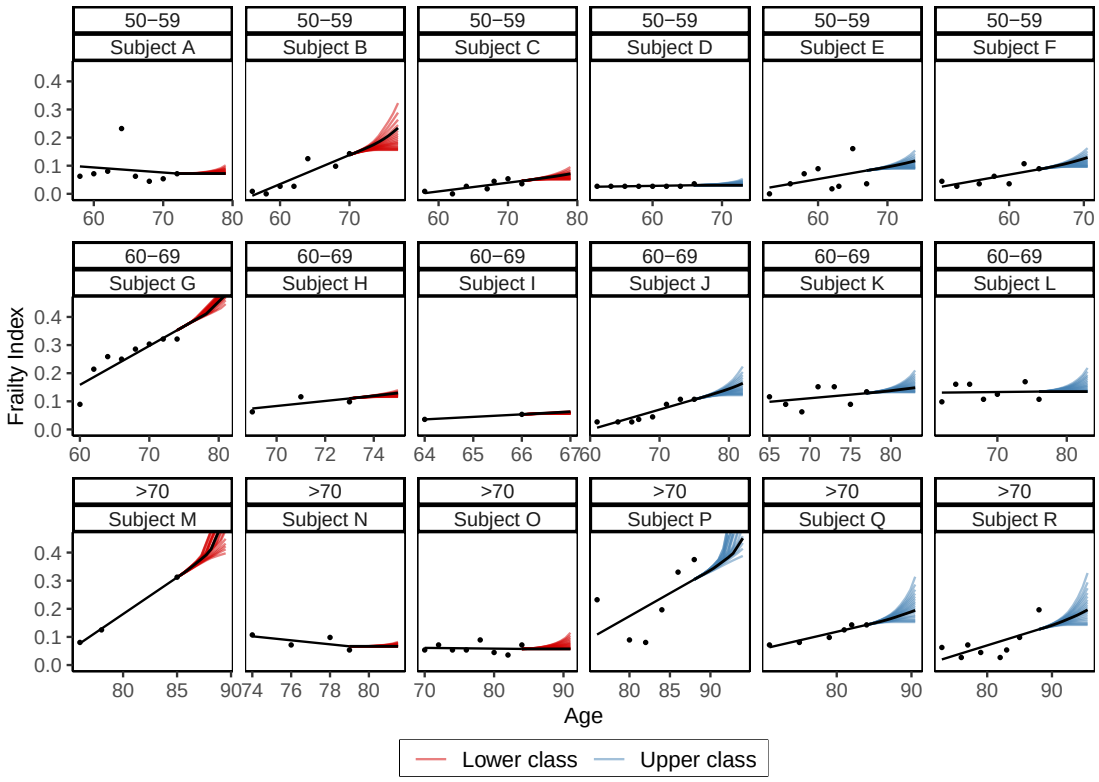


Figure 5.7: Prediction trajectories for a random sample of 18 non-deprived at baseline individuals where the length of the prediction period is chosen to be half the length of timespan of the subject's longitudinal trajectory. The dots are the observed frailty scores. The solid black line up to the last observation time point represents the linear regression line obtained from the least squares fit on each individual trajectory. After the last time point of observation is the estimated α-quantile prediction trajectory conditioned on the subject's last fitted value if the subject was to remain at the same quantile where he/she was travelling in the past (black solid line). Additionally, pessimistic and optimistic future trajectory scenarios are depicted with red (for lower class subjects) and blue (for upper class subjects).

deterioration in the near future (e.g. subject A, subject D, subject I, etc.). At the same time, there are cases where even though the past trajectory was relatively stable, there is much more uncertainty associated to their prediction (e.g. subject B, subject Q, subject R, etc.).

Finally, to further assess the utility of the current method, we compared its predictive validity as compared to the state-of-the-art methods, i.e. a QGC model and a LME model, both common choices for analysing this type of longitudinal data (Rogers et al., 2017). We built the latter using the same covariates (all of which were statistically significant) and an interaction term between age group and wave. We performed 10-fold cross-validation by randomly taking segments of individual observations to create the training and test sets for every iteration. As in Section 5.3, we evaluated performance based on the MAE.

Model	Mean	SD	Min.	25% Quant.	75% Quant	Max.
CQ	5.65×10^{-2}	4.47×10^{-2}	1×10^{-8}	1.12×10^{-2}	6.84×10^{-2}	7.26×10^{-2}
LME	3.46×10^{-2}	3.36×10^{-2}	2.14×10^{-6}	1.13×10^{-2}	4.70×10^{-2}	3.94×10^{-2}
QGC	3.46×10^{-2}	3.36×10^{-2}	3.22×10^{-6}	1.13×10^{-2}	4.67×10^{-2}	3.95×10^{-2}

Table 5.4: Prediction accuracy based on the MAE after performing 10-fold cross-validation using the Conditional Quantile (CQ) methodology, Linear Mixed Effects Modelling (LME) and the Multilevel Growth Curve (QGC) model.

As demonstrated in Table 5.4, our method (CQ) had similar performance to both LME and QGC methods (MAE equal to 5.65×10^{-2} as compared to 3.46×10^{-2} for both LME and QGC respectively), highlighting its additional strength in providing some sort of "personalised" prediction for each individual.

5.5 Discussion

Frailty progression is known to be affected by various factors such as biological, behavioural and social factors (Rogers et al., 2017; Niederstrasser et al., 2019). Analysing longitudinal data on frailty often encompass a lot of difficulties due to population variation, individual heterogeneity and the short lengths of the observed trajectories relative to the period of interest which often covers the second half of human lifespan. Here, we studied the association between frailty and self-perceptions regarding individual socioeconomic status after controlling for age at baseline using a data driven conditional quantile regression approach which allows for dynamic modelling of the short observed frailty trajectories. We analysed data from the ELSA and constructed a FI score based on 56 variables. Our implementation of the conditional quantile methodology revealed counter-intuitive relationships between self-perceived social class and relative deprivation, accentuating that once controlled for frailty level and age at baseline, the effect of negative socioeconomic self-perceptions on frailty progression is minor. We, however, highlight that the results shown should not be perceived as marginal effects of self-perceived social class and relative deprivation as the rate of frailty change is mainly affected from the instantaneous frailty level. So, if there are not many individuals reaching high FI levels for a particular subgroup, those who do will be expected to experience faster increases in their FI over time. We, finally, compared our method to multilevel quadratic and

linear mixed effect modelling, which emphasised the added value of our work for getting individual predictions facilitating the detection of vulnerable individuals and leading to a personalised approach against supporting ageing adults. The robustness and prediction performance of the proposed method were, further, confirmed through simulation studies under multiple data scenarios.

Our work relies on four major assumptions. First, that frailty is a monotonic process which is true according to previous literature for the application on target (Searle et al., 2008; Clegg et al., 2013). However, this methodology is not expected to be particularly useful for non-monotonic processes. In the ELSA application, frailty is an increasing process, so an individual with an observed trajectory with a negative slope will fall into the lowest estimated quantile conditional on his/her local trajectory level and any predictions will indicate no expected change on individual frailty level in the near future. Second, we assume that the observed trajectories do not carry information beyond the local level and slope of the process, as a result if a larger amount of repeated measures is available we expected to lose information by relying only on the local level and slope. In this application the average amount of trajectory coverage was 12 years relative to the examined period of 40 years, and for more than half the individuals the available trajectory information covered less than a quarter of the period under study. As illustrated from our simulation studies, the method's results remain valid and accurate even if this assumption is violated for some of the observed individual trajectories. Third, we assume that the missing data are missing at random and if not, then some survival bias could potentially affect our estimates. For the purpose of this study, we do not explore how this affects the reported results but we leave it for future work. Finally, it needs further investigation to what extent our results are biased due to measurement error in the FI scores. It has been previously shown and confirmed by our simulation studies, that under noisy data scenarios, the estimation of the lower and upper quantiles can be considerably affected. Improving the conditional quantile methodology so that it accurately estimates the upper and lower tail quantiles whilst allowing for measurement error goes beyond the scope of the current study but we leave that for further investigation.

Modelling incomplete longitudinal data has been widely studied in the literature. Random effect models (Niederstrasser et al., 2019; Franse et al., 2017; Szanton et al., 2010; Diggle et al., 2002), growth curve modelling (Rogers et al., 2017) as well as

functional analysis methods (Fitzmaurice et al., 2008) have all been employed to study complex processes over time while quantile regression methods have been developed to study distributional changes on top of the mean trend exploration (Koenker, 2004). However, none of the methods used to study ageing have explored frailty progression without making strong assumptions about the underlying trend of the process and its distributional properties. This is one of the main contributions of this work. We employed a conditional quantile modelling framework to explore the distributional evolution of frailty over time. At the same time, we assessed the relationship between these changes and self-perceived social class and relative deprivation. This is another strength of our work since even though there are numerous studies looking at the relationships between socioeconomic factors and frailty, research on how self-perceptions affect the ageing process are limited. Our investigation gave interesting insights into a potential negative effect of positive self-perceptions with frailty. There are numerous factors which might affect this relationship, however further investigation is needed to confirm those findings since the differences found were not very strong.

5.6 Summary

The main purpose of this chapter is to examine the frailty progression over time and how self-perceived social class and relative deprivation affect this process. The proposed methodology allows a more comprehensive picture of the mechanisms that undergo frailty progression to be obtained, highlighting the need to support not only frail individuals but also non-frail since the higher the frailty level and age, the larger the expected increase in frailty over time. This framework provides a dynamic way of estimating individual future trajectories which can further help in identifying those in need before they even become frail. The proposed method can be extended to several other longitudinal applications of ageing with severe missingness and when the available information is very short relative to the interval of interest. We believe that the findings from this study along with the additional benefits of the conditional quantile methodology can deepen our understanding of ageing and frailty and can potentially be used as a dynamic tool to facilitate policy making.

In the following chapter, we build on the current findings by handling the multidimen-

sional outcome variables differently. In particular, we model them directly by assuming that they collectivelly measure the latent ageing domains presented in Chapters 2 and 3. To do so, we develop a bivariate latent Gaussian process model which not only models each of these domains separately, but also reveals interelationships between them.

Chapter 6

A Bivariate Latent Gaussian Process Model for Analysing Ordinal Panel Data

6.1 Introduction

In the previous chapter, we analysed longitudinal questionnaire data by aggregating the multivariate responses into a single continuous index, the FI. However, our study showed that the produced metric does not reveal or depict the multidimensional nature of ageing which could potentially promote targeted strategies towards supporting old people. In this chapter, we move a step further by focusing on the raw questionnaire data collected in ELSA. The data handled are, thus, multivariate longitudinal and, similar to the previous chapter, irregularly collected over age. This study highlights not only the statistical challenges involved but also creates new opportunities towards exploring the dynamics of the ageing process.

Panel studies have become very popular in social sciences, as they enable to study change in human behaviour and society over time, and explore potential determinants of this change. Due to the fact that the behaviours and/or traits under study often cannot be measured directly, it is common to use questionnaires and tests to repeatedly collect individual observations through time, thus leading to a large collection of multivariate and potentially non-continuous longitudinal data. Popular methods to analyse this type of data when they measure a single trait include hierarchical mixed-effects modelling

[MEM; Chapter 14 in Diggle et al. (2002); Chapter 16 in Fitzmaurice et al. (2008); Proust et al. (2006); Cagnone et al. (2009); Proust-Lima et al. (2013); Verbeke et al. (2014); Li et al. (2017)], latent growth curve modelling [LGCM; Chapter 7 in Bollen and Curran (2006); Byrne and Crombie (2010); Masyn et al. (2014); Burant (2016)] and longitudinal Item Response Theory [LIRT; Andrade and Tavares (2005); Van der Linden (2016); Wang and Nydick (2020)]. Multiple traits can be simultaneously examined using extensions of LGCM for repeated measures categorical data and longitudinal multidimensional IRT [LMIRT; Bollen and Curran (2006); Marvelde et al. (2006); Reckase (2009); Kline (2015); Van der Linden (2016); Wang and Nydick (2020)].

All the aforementioned approaches are similar in the sense that they use one or more latent variables to reduce the dimensionality of the multivariate vector of outcomes and model change over time at both population and individual level (Muthén, 2002; Skrondal and Rabe-Hesketh, 2007; Verbeke et al., 2014). Differences are mainly detected with regards to the assumptions that are made for the dependence structure between measurements within an item, between items at the same time point and between items at different time points, as well as, the assumptions regarding the functional form of the latent concept measured through these items (linear or non-linear). MEM assumes that observations are realisations of a latent subject specific trajectory and introduces an hierarchical (or multilevel) random effects model for both individuals and items which allows a lower-dimensional parametrisation of the variance-covariance matrix of the data. Even though this method easily handles data with varying number of observations per subject collected at individually varying time points and enables modelling of complex time trends, a large number of items quickly leads to computational problems (Verbeke et al., 2014). Recent work focuses on overcoming these computational problems as well as modelling longitudinal outcomes of mixed types, however, none of them explores modelling data involving groups of items which measure more than one underlying concepts (Cagnone et al., 2009; Proust-Lima et al., 2013; Li et al., 2017).

Both LGCM and L(M)IRT can be viewed as special categories within the structural equation modelling (SEM) family (Chapters 13 and 15 in Kline, 2016) where attention concentrates on the latent factors over time, measurement error and parameter interpretation. LGCM assumes that the repeated outcomes are multiple indicators for latent growth factor values for each individual. Latent growth curve models consist of a

measurement and a structural model. The first describes the relationship between the observed outcomes and individual growth factor values through time capturing the within individual change, whilst, the second models the population distribution of the latent growth factors (between individual change over time) (Bollen and Curran, 2006; Masyn et al., 2014; Duncan et al., 2013). This methodology has been extensively studied and applied for categorical outcomes (Duncan et al., 2013; Masyn et al., 2014; Lee et al., 2018). Nevertheless, complex non-linear trends over time cannot be easily accommodated (Bollen and Curran, 2006), which is also the case for L(M)IRT. IRT was initially introduced in the 1980s and mainly used in educational and psychological research to facilitate the design, analysis, and scoring of tests, questionnaires, and similar instruments measuring examinee's abilities and attitudes (van der Linden and Hambleton, 2013). In fact, LGCM and L(M)IRT only differ in the measurement model, since the second makes additional assumptions regarding the item parameters or introduces additional guessing parameters to accommodate for test or questionnaire-specific characteristics affecting individual responses (Van der Linden, 2016). Especially when it comes to multivariate longitudinal data analysis these two methods are often used interchangeably since the L(M)IRT can be seen as a natural extension of LGCM (Hsieh et al., 2010; Paek et al., 2016; Wang and Nydick, 2020).

Recent developments allow increased flexibility in investigating how behaviours evolve over time. For instance, Proust-Lima et al. (2013) introduced a latent process model for multivariate mixed longitudinal data which handled flexible modelling of time trends, individual varying measurement times and mixed outcomes (including categorical and non-Gaussian data) but it was designed to enable modelling of items referring to a single latent process. Hsieh et al. (2010) extended the unidimensional LIRT model for modelling multidimensional traits but it did not allow for flexible modelling of the underlying concepts over time. Within the SEM framework, it is difficult to model data considering the exact (continuous) time of measurement (e.g., individuals' age), especially if data are collected over a wide time interval and longitudinal measurements are sparse within it. Within the MEM framework, even though modelling non-linear relationships over the exact (continuous) time of measurement is achievable, modelling categorical data measuring more than one latent constructs becomes cumbersome.

Current approaches to analyse intensive longitudinal, functional and time-series

data employ latent Gaussian processes (LGPs) to flexibly model non-linear outcome trends over time (Shi and Choi, 2011; Wang and Shi, 2014; Greven and Scheipl, 2017; Vandenberg-Rodes and Shahbaba, 2015; Gao et al., 2018; Chen and Zhang, 2020). Within the machine learning research community, Gaussian process latent variable modelling is used for regression, dimensionality reduction and clustering purposes due to its inherent flexibility and distribution-free form which can accommodate the increased complexity of image and speech recognition, information retrieval and recommender systems data (Li and Chen, 2016; Rasmussen and Williams, 2006). Hall et al. (2008) employed a LGP model to analyse sparse and irregularly collected longitudinal data and extended functional principal components analysis for non-Gaussian repeated measures data by perceiving longitudinal data as sparse non-Gaussian functional data. In another study, Wang and Shi (2014) built a LGP framework for sparse non-Gaussian functional data. Recent work extended the LGP model introduced by Chen and Zhang (2020) for multivariate intensive longitudinal data to multivariate panel data (Karch et al., 2020). Their work generalises all LGCM, LIRT and MEM for modelling latent constructs measured in panel studies allowing classical statistical inference and predictive modelling of the individual heterogeneous latent trajectories.

In this work, we extend the LGP methodology proposed by Chen and Zhang (2020) to handle multivariate ordinal longitudinal data measuring more than one underlying concepts. By relying on a bivariate LGP (BiLGP) approach defined in continuous time, we manage to handle times and number of measurements that vary from one subject to another. This allows consideration of the exact time of measurement rather than the time of follow-up. We validate the proposed model using simulation studies and we implement it using data from the ELSA to explore how four, out of the five ageing domains introduced in Chapters 2 and 3, evolve and interrelate over time.

The remainder of this chapter is organised as follows. In the following section, we provide the necessary background to introduce our LMGP model for ordinal panel data in Section 6.3. Inference and identifiability considerations are also discussed in Section 6.3. Simulation studies and the application using the ELSA data are presented in Sections 6.4 and 6.5 respectively. Finally, we conclude with a discussion in Section 6.6 and a summary of what has been discussed in Section 6.7.

6.2 Preliminaries

Throughout this chapter, we will stay within the framework of ordinal longitudinal data and Gaussian processes. In this section, we provide the background for our BiLGP model which is discussed in Section 6.3. First, in Section 6.2.1, we present the standard measurement model for ordinal longitudinal data. Then, in Sections 6.2.2 and 6.2.3, we discuss univariate and bivariate Gaussian processes. Finally, in Section 6.2.4, we give a short overview of common techniques to build valid and flexible cross-covariance functions.

6.2.1 Measurement Model for Ordinal Multivariate Longitudinal Data

Suppose there are N individuals responding to I questionnaire items (questions). The response vector of individual n at some continuous time t will be

$$\boldsymbol{Y}_n(t) = (Y_{n1}(t), \ldots, Y_{nI}(t))^{\mathrm{T}},$$

where $t \in [0, T]$ and $T \in \mathbb{R}_{>0}$. Let observations

$$\boldsymbol{y}_n(t) = (y_{n1}(t), \ldots, y_{nI}(t))^{\mathrm{T}}$$

be collected at some distinct time points, $t \in \{t_{n1}, \ldots, t_{nJ_n}\}$, where J_n is the total number of measurements for individual n. As in Bartholomew (1980), Wang and Nydick (2020) and Chen and Zhang (2020), if we assume that there exists an underlying stochastic relationship

$$P(\boldsymbol{y}_n(t_{n1}), \ldots, \boldsymbol{y}_n(t_{nJ_n}) | \theta_n(t), t \in [0, T])$$

that relates the observed responses $\boldsymbol{y}_n(t)$ with a latent curve $\theta_n(\cdot) = \{\theta_n(t) \ : \ t \in [0, T]\}$, that is the unobservable latent trait that varies over time, to model the relationship between the observed responses and the unobservable stochastic process, we make the following assumptions:

(A1) the observed responses between two individuals are independent to each other, i.e., $P(\boldsymbol{y}_1(t_{11}), \ldots, \boldsymbol{y}_1(t_{1J_1}), \ldots, \boldsymbol{y}_N(t_{N1}), \ldots, \boldsymbol{y}_N(t_{NJ_N}) | \theta_1(t), \ldots, \theta_N(t), t \in [0, T]) = \prod_{n=1}^{N} P(\boldsymbol{y}_n(t_{n1}), \ldots, \boldsymbol{y}_n(t_{nJ_n}) | \theta_n(t), t \in [0, T])$;

(A2) the latent trait level at any other time point is conditionally independent of the observed responses, given the latent trait level at the corresponding time points of observation, i.e., $P(\boldsymbol{y}_n(t_{n1}), \ldots, \boldsymbol{y}_n(t_{nJ_n})|\theta_n(t), t \in [0,T]) = P(\boldsymbol{y}_n(t_{n1}), \ldots, \boldsymbol{y}_n(t_{nJ_n})|\theta_n(t_{n1}), \ldots, \theta_n(t_{nJ_n}))$;

(A3) within-individual responses are assumed to be independent given the corresponding latent trait levels, i.e., $P(\boldsymbol{y}_n(t_{n1}), \ldots, \boldsymbol{y}_n(t_{nJ_n})|\theta_n(t_{n1}), \ldots, \theta_n(t_{nJ_n})) = \prod_{j=1}^{J_n} \boldsymbol{P}(\boldsymbol{y}_n(t_{nj})|\theta_n(t_{nj}))$, and;

(A4) responses between different items are assumed to be independent given the underlying latent trait, i.e., $\boldsymbol{P}(\boldsymbol{y}_n(t_{nj})|\theta_n(t_{nj})) = \prod_{i=1}^{I} P(y_{ni}(t_{nj})|\theta_n(t_{nj}))$.

Hence, someone can write

$$\begin{aligned} P(\boldsymbol{Y} = \boldsymbol{y}) &= \int_{-\infty}^{\infty} \cdots \int_{-\infty}^{\infty} \prod_{n=1}^{N} \prod_{j=1}^{J_n} \prod_{i=1}^{I} P(y_{ni}(t_{nj})|\theta_n(t_{nj})) f(\boldsymbol{\theta}) d\boldsymbol{\theta} \\ &= \mathrm{E}\Big[\prod_{n=1}^{N} \prod_{j=1}^{J_n} \prod_{i=1}^{I} P(y_{ni}(t_{nj})|\theta_n(t_{nj})) \Big] \end{aligned} \tag{6.1}$$

where $\boldsymbol{\theta} = [\boldsymbol{\theta}_1, \boldsymbol{\theta}_2, \ldots, \boldsymbol{\theta}_N]^{\mathrm{T}}$ and $\boldsymbol{\theta}_n = [\theta_n(t_{n1}), \theta_n(t_{n2}), \ldots, \theta_n(t_{nJ_n})]^{\mathrm{T}}$.

Now, if $Y_{ni}(t) \in \{0, 1, \ldots, C_i\}$ for $n \in \{1, 2, \ldots, N\}$, $i \in \{1, 2, \ldots, I\}$ and $t \in [0, T]$, where $0 < 1 < \ldots < C_i$ are the $C_i + 1$ ordered categories of item i, we can express an individual's probability of answering c ($c \in \{0, 1, \ldots, C_i\}$) to the item i at time t given his/her latent trait curve $\theta_n(t)$ as

$$\begin{aligned} P(Y_{ni}(t) = c|\theta_n(t)) &= P(Y_{ni}(t) \geq c|\theta_n(t)) - P(Y_{ni}(t) \geq c+1|\theta_n(t)) \\ &= P(Y_{ni}(t) < c+1|\theta_n(t)) - P(Y_{ni}(t) < c|\theta_n(t)) \end{aligned} \tag{6.2}$$

with $P(Y_{ni}(t) \geq 0|\theta_n(t)) = 1$ and $P(Y_{ni}(t) > C_i|\theta_n(t)) = 0$ meaning that any individual is able to give the minimum response in any question (item) and none can answer more than the highest category (C_i). The ordinal response model is, then, defined as

$$P(Y_{ni}(t) \leq c|\theta_n(t)) = F(b_{i,c} + \alpha_i \theta_n(t)) \tag{6.3}$$

where $F(\cdot)$ can be either the standard normal or logistic distribution. The parameters α_i and $b_{i,c}$ represent the corresponding item loadings and category thresholds for the ordinal responses; in particular, for identifiability purposes, α_i is constrained in $(0, +\infty)$

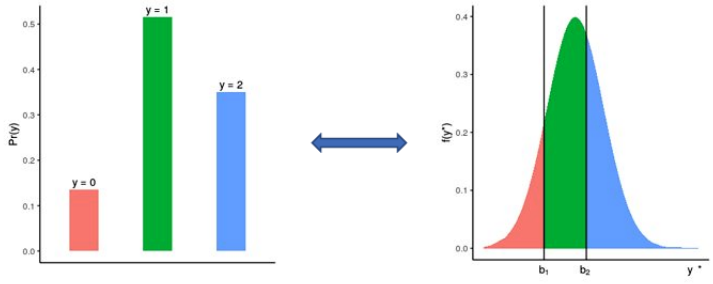


Figure 6.1: Latent variable formulation of ordinal data. We consider the scenario where $C_i = 2$. Then, $Y = c$ if $b_c \leq Y^* < b_{c+1}$ where $c \in \{0, 1, 2\}$.

and $b_{i,c} \in (-\infty, +\infty)$ with $-\infty = b_{i,0} < b_{i,1} < b_{i,2} < \cdots < b_{i,C_i} < b_{i,C_i+1} = \infty$ with $i \in \{1, 2, \ldots, I\}$ and $c \in \{0, 1, \ldots, C_i\}$. Therefore, under the probit model for the ordinal responses, we get

$$P(Y_{ni}(t) = c|\theta_n(t)) = \Phi(b_{i,c+1} + \alpha_i\theta_n(t)) - \Phi(b_{i,c} + \alpha_i\theta_n(t)). \tag{6.4}$$

Note that Eq. 6.4 can be alternatively reformulated as

$$Y^*_{ni}(t) = -\alpha_i\theta_n(t) + \varepsilon_{ni}(t) \tag{6.5}$$

where α_i and $b_{i,c}$ as in Eq. 6.4, $\varepsilon_{ni}(t) \sim \mathrm{N}(0, 1)$ and $Y^*_{ni}(t)$ is a continuous latent variable with $n \in \{1, 2, \ldots, N\}$, $i \in \{1, 2, \ldots, I\}$ and $t \in [0, T]$. Under this formulation, the response $Y_{ni}(t)$ can be seen as a coarsened version of $Y^*_{ni}(t)$ obtained by

$$Y_{ni}(t) = c \text{ if } b_{i,c} \leq Y^*_{ni}(t) < b_{i,c+1}. \tag{6.6}$$

Figure 6.1 provides a simple illustration of the above idea under the scenario where $C_i = 2$.

6.2.2 Univariate Gaussian Processes

Let the probability space be denoted by $(\Omega, \mathscr{F}, P)$ and the index set S. We define a stochastic process as the finite or real valued function $\Theta(t, \omega)$ which for every fixed $t \in S$ is a measurable function of $\omega \in \Omega$ and can be written as $\{\Theta_t(\omega) : t \in S\}$. A Gaussian process is, then, a stochastic process where all finite-dimensional distributions $F_{t_1,t_2,\ldots,t_k}$ are themselves multivariate normal distributions for any choice of $k \in \mathbb{N}$ and

finite set of indices $\{t_1, t_2, \ldots, t_k\} \subseteq S$ and it can be fully specified by its mean function $\mu : S \longrightarrow \mathbb{R}$ with $\mu(t) = \mathrm{E}[\Theta(t)]$, and kernel function $K : S \times S \longrightarrow \mathbb{R}$ with $K(t, t') = \mathrm{Cov}(\Theta(t), \Theta(t')) = E[(\Theta(t) - \mu(t))(\Theta(t') - \mu(t'))]]$ (Billingsley, 2008). We, then, write

$$\Theta(t) \sim \mathcal{GP}(\mu(t), K(t, t')).$$

Consequently, a continuous time stochastic process on a time interval $[0, T]$ is a Gaussian process if for every $k \in \mathbb{N}$ and $t_1, t_2, \ldots, t_k \in [0, T]$

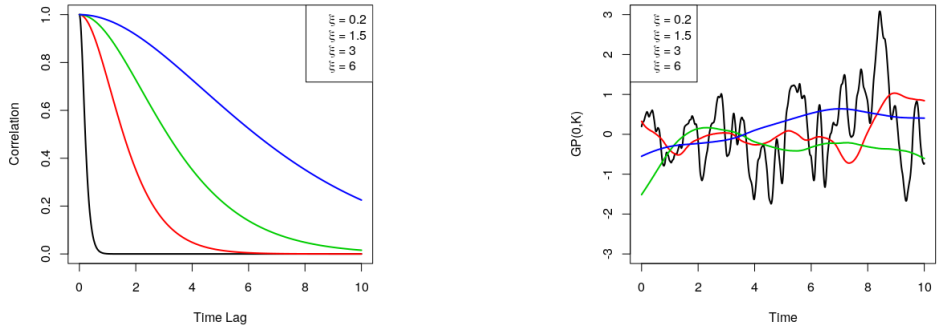
$$\boldsymbol{\Theta} = (\Theta(t_1), \Theta(t_2), \ldots, \Theta(t_k))^{\mathrm{T}} \sim N_k(\boldsymbol{\mu}, \Sigma)$$

where $\boldsymbol{\mu} = (\mu(t_1), \mu(t_2), \ldots, \mu(t_k))^{\mathrm{T}}$ and Σ is a covariance matrix with (i, j)-entry $K(t_i, t_j)$, $i, j = 1, \ldots, k$.

Note that the covariance function of a Gaussian process describes how similar neighbouring data points are, however, for $K(\cdot, \cdot)$ to be a valid covariance function, the generated matrix should be positive semi-definite for any choice t and t'. Furthermore, a kernel function, and subsequently, the Gaussian process which it specifies, will be *stationary* if $K(t, t') = K(v)$, where $v = t - t'$ for any $t, t' \in S$; and, *isotropic*, if it depends on the distance between t and t' alone, i.e., $K(t, t') = K(|v|)$ where $|v| = |t - t'|$. Common examples of stationary and isotropic kernels are the squared exponential, Matérn, rational quadratic, Ornstein-Uhlenbeck etc. Linear combinations of the above kernels can be also result in valid kernel functions. The choice of kernel function is crucial, since it determines the qualitative beliefs about the underlying signal. For instance, the most common kernel function is the squared-exponential kernel, which is infinitely differentiable and results in smoother processes but can be unrealistic for modelling a range of physical processes (Stein, 2012); while the Ornstein-Uhlenbeck is non-differentiable and it has been firstly introduced to model the velocity of a particle undergoing Brownian motion (Rasmussen and Williams, 2006). We refer the interested readers to Chapter 4 of the Rasmussen and Williams (2006) book for a detailed overview of the topic.

In this work, we employ the Matérn class of covariance functions (Figure 6.2), that is

$$M(v|\xi, \nu) = \frac{2^{1-\nu}}{\Gamma(\nu)} \left(\frac{\sqrt{2\nu} r}{\xi} \right)^{\nu} B_{\nu} \left(\frac{\sqrt{2\nu} r}{\xi} \right) \tag{6.7}$$



(a) Correlation decay over time.

(b) Generated Gaussian process.

Figure 6.2: Matérn correlation and Gaussian process generated using a Matérn kernel for different scale parameter ξ values. Here, $\nu = 5/2$.

where $v = |t - t'|$, ν is a positive smoothing parameter, ξ is a positive scale parameter and B_ν is a modified Bessel function. Due to its inherent flexibility and nice properties, e.g., for $i = 0, 1, \ldots$ and $\nu = \frac{1}{2} + i$ the Matérn covariance becomes a product of an exponential and a polynomial of order i function, the Matérn is preferred over other types of covariance functions for numerous data applications (Stein, 2012; Rasmussen and Williams, 2006).

6.2.3 Multivariate Gaussian Processes

Now, let a P-variate Gaussian process be denoted by $\boldsymbol{\Theta}(t) = (\Theta_1(t), \Theta_2(t), \ldots, \Theta_P(t))^{\mathrm{T}}$. Similar to the univariate case, it can be fully specified through its mean $\boldsymbol{\mu}(t) = \mathrm{E}[\boldsymbol{\Theta}(t)]$ and its cross-covariance matrix function $\boldsymbol{K}(t, t') = \mathrm{Cov}(\boldsymbol{\Theta}(t), \boldsymbol{\Theta}(t')) = \{K_{pq}(t, t')\}_{p,q=1}^{P}$ where

$$K_{pq}(t, t') = \mathrm{Cov}(\Theta_p(t), \Theta_q(t')), \; t, t' \in T \tag{6.8}$$

for $p, q = 1, \ldots, P$. Here, $K_{pq}(\cdot, \cdot)$ can be seen as the marginal-covariance function for $p = q$ and as the cross-covariance function for $p \neq q$. For the above cross-covariance function to be a valid the mapping $\boldsymbol{K} : T \times T \longrightarrow \mathbf{M}_{P \times P}$, where $\mathbf{M}_{P \times P}$ is the set of $P \times P$ real-valued matrices, must be *non-negative definite*, i.e., for the covariance matrix

$\boldsymbol{\Sigma}$ of the random vector $(\boldsymbol{\Theta}(t_1)^{\mathrm{T}}, \ldots, \boldsymbol{\Theta}(t_k)^{\mathrm{T}})^{\mathrm{T}} \in \mathbb{R}^{kP}$,

$$\boldsymbol{\Sigma} = \begin{pmatrix} \boldsymbol{K}(t_1, t_1) & \boldsymbol{K}(t_1, t_2) & \cdots & \boldsymbol{K}(t_1, t_k) \\ \boldsymbol{K}(t_2, t_1) & \boldsymbol{K}(t_2, t_2) & \cdots & \boldsymbol{K}(t_2, t_k) \\ \vdots & \vdots & \ddots & \vdots \\ \boldsymbol{K}(t_k, t_1) & \boldsymbol{K}(t_k, t_2) & \cdots & \boldsymbol{K}(t_k, t_k) \end{pmatrix} \tag{6.9}$$

and any vector $\boldsymbol{a} \in \mathbb{R}^{kP}$ we should have $\boldsymbol{a}^{\mathrm{T}}\boldsymbol{\Sigma}\boldsymbol{a} \geq 0$.

Stationarity and *isotropy* can be defined similarly to the univariate case. In particular, a multivariate Gaussian process is *stationary* if there is a mapping $K_{pq} : T \longrightarrow \mathbb{R}$ such that

$$\mathrm{Cov}(\Theta_p(t), \Theta_q(t')) = K_{pq}(v) \text{ where } v = t - t' \text{ for any } t, t' \in T \tag{6.10}$$

and *isotropic*, if it is stationary and invariant under rotations and reflections; in other words, if there is a mapping $K_{pq} : \mathbb{R}_+ \cup 0 \longrightarrow \mathbb{R}$ such that

$$\mathrm{Cov}(\Theta_p(t), \Theta_q(t')) = K_{pq}(|v|) \tag{6.11}$$

Additional properties of the covariance and cross-covariance functions can be found in Genton and Kleiber (2015); Gneiting et al. (2010); Rasmussen and Williams (2006) and Abrahamsen (1997).

6.2.4 Cross-covariance Functions for Multivariate Gaussian Processes

There are multiple techniques for building valid and flexible cross-covariance functions. In this section, we give a brief summary of these techniques, however interested readers can look at Genton and Kleiber (2015) for a thorough review.

Univariate covariance models can be combined to build valid cross-covariance functions using the linear model of coregionalisation, convolution techniques or by using the latent dimensions technique. In particular, when the linear model of coregionalisation is employed, the multivariate random process is expressed as a linear combination of a number of independent processes with some pre-specified stationary covariance function. For this approach, only the univariate covariances must be specified avoiding direct specification of the cross-covariance function. However, if a large number of processes is used, the number of parameters to be estimated increases vastly while the cross-covariance

smoothness is only controlled through the smoothness of each univariate process.

Convolution is another technique which can be used to build valid covariance functions. Authors distinguish two different types: kernel convolution, where all processes are generated by the same underlying process; and, covariance convolution. Both necessitate numerical integration, while the first has the additional disadvantage that it imposes strong assumptions about the dependence structure between the pairs of the univariate processes and produces parameters difficult to interpret. A special case of covariance convolution is the multivariate Matérn covariance model defined as

$$\mathrm{Cov}(\Theta_d(t), \Theta_q(t')) = \sigma_d \sigma_q \beta_{dq} M(v|\nu_{dq}, \xi_{dq}), \tag{6.12}$$

where $M(\cdot|\nu_{dq}, \xi_{dq})$ is univariate Matérn covariance as in Eq. 6.7, $\sigma_d,\ \sigma_q$ are the standard deviations of the two marginal processes, $d, q \in \{1, 2, \ldots, D\}$ and β_{dq} is a cross-correlation coefficient, representing the strength of correlation between Z_d and Z_q when $v = t - t' = 0$. Conditions on model parameters ν_{dq}, ξ_{dq} and β_{dq} for all $d, q \in \{1, 2, \ldots, D\}$ should be specified for Eq. 6.12 to result in a valid covariance function. As shown in Eq. 6.12, the resulting cross-correlation can be written in a closed form and each constituent process of the multivariate Gaussian process is allowed to have a marginal Matérn covariance, with Matérns also composing the cross-covariance structures.

Finally, another popular method introduced by Apanasovich and Genton (2010), employs latent dimensions to transform the cross-covariance function of the multivariate random field to a valid univariate covariance function. Common extensions of this technique often allow modelling non-stationary processes (Bornn et al., 2012), however this is beyond the scope of the current work.

6.3 Methodology

In this section, we introduce the BiLGP modelling framework for ordinal multivariate longitudinal data measuring more than one concepts over time. In Section 6.3.1, we extend the measurement model described in Section 6.2.1 when instead of a latent curve $\theta(t)$ we have a two-dimensional latent curve vector $\boldsymbol{\theta}(t) = (\theta_1(t), \theta_2(t))^{\mathrm{T}}$. We, then continue, by presenting the bivariate extension of the structural model discussed in Chen and Zhang (2020) and Karch et al. (2020). Parameter and individual trajectory estimation

are also discussed in this section. Identifiability, parameter tuning and computational intensity considerations are discussed in Sections 6.3.2, 6.3.3 and 6.3.4.

6.3.1 Model Specification and Inference

The Measurement Model

Consider the multivariate longitudinal data introduced in Section 6.2.1. We assume that each individual response vector $\boldsymbol{Y}_n(t) = (Y_{n1}(t), \ldots, Y_{n1}(t))^{\mathrm{T}}$ is, now, associated with a latent curve vector $\boldsymbol{\theta}_n(t) = (\theta_{1n}(t), \theta_{2n}(t))^{\mathrm{T}}$ so that we can express an individual's probability of answering c ($c \in \{0, 1, \ldots, C_i\}$) to the item i at time t under the probit model for the ordinal responses, as

$$P(Y_{ni}(t) = c|\boldsymbol{\theta}_n(t)) = \Phi(b_{i,c+1} + \sum_{d=1}^{D} \alpha_{id}\theta_{dn}(t)) - \Phi(b_{i,c} + \sum_{d=1}^{D} \alpha_{id}\theta_{dn}(t)) \tag{6.13}$$

where $D = 2$, α_{id} and $b_{i,c}$ represent the corresponding item loadings and category thresholds for the ordinal responses; in particular, $\alpha_{id} \in (0, +\infty)$ and $b_{i,c} \in (-\infty, +\infty)$ with $c \in \{0, 1, \ldots, C_i\}$, $-\infty = b_{i,0} < b_{i,1} < b_{i,2} < \cdots < b_{i,C_i} < b_{i,C_i+1} = \infty$, $d \in \{1, 2, \ldots, D\}$ and $i \in \{1, 2, \ldots, I\}$. Alternatively, one can reformulate Eq. 6.13 using another level of latency, as in Section 6.2.1, i.e.,

$$Y^*_{ni}(t) = -\sum_{d=1}^{D} \alpha_{id}\theta_{dn}(t) + \varepsilon_{ni}(t) \tag{6.14}$$

with $D = 2$, α_{id} and $b_{i,c}$ similar to Eq. 6.13 and $\varepsilon_{ni}(t) \sim \mathrm{N}(0, 1)$ with $n \in \{1, 2, \ldots, N\}$, $i \in \{1, 2, \ldots, I\}$ and $t \in [0, T]$.

The Structural Model

Consider the framework introduced in Sections 6.2.1 and 6.2.2. We can assume that $\theta_n(\cdot)$ in Eq. 6.4 is a Gaussian process and we can write

$$\theta_n(t) = \mu(t) + \bar{\theta}_n(t), \tag{6.15}$$

where $\mu(t)$ is a smooth mean function that can be interpreted as the population mean trait trend, and $\bar{\theta}_n(t) \sim \mathcal{GP}(0, K)$ is another zero mean Gaussian process with the kernel function K describing individual deviation from the population mean function $\mu(t)$. Both

$\mu(t)$ and K can be parametrised according to the application needs, taking either known parametric forms or by allowing more flexibility using alternative parametrisations (Rasmussen and Williams, 2006). For example, one can adopt a parametrisation of $\mu(t)$ and $\bar{\theta}_n(t)$ using basis functions, i.e.,

$$\mu(t) = \gamma_0 + \gamma_1 B_1(t) + \ldots + \gamma_H B_H(t)$$

and

$$\bar{\theta}_n(t) = \sum_{h=1}^{H_1} \omega_h Z_{nh} \Phi_h(t)$$

where $\{B_h(\cdot) : h = 1, \ldots, H\}$ and $\{\Phi_h(\cdot) : h = 1, \ldots, H_1\}$ are pre-specified basis functions on $[0, T]$; $\gamma_h, h = 1, \ldots, H$, and $\omega_h, h = 1, \ldots, H_1$, are model parameters and $Z_{nh}, h = 1, \ldots, H_1$, are i.i.d. standard normal random variables.

The above framework is extensively discussed in Chen and Zhang (2020); Karch et al. (2020) and Wang and Shi (2014). By allowing $\boldsymbol{\theta}(\cdot)$ in the measurement model to be a bivariate Gaussian process, we enable modelling multivariate longitudinal data which measure more than one latent traits over time. Below, an appropriate cross-covariance function is specified and a computational algorithm is developed to estimate model parameters and individual trajectories.

In particular, the latent bivariate random process $\boldsymbol{\theta}_n(t)$ is assumed to be bivariate Gaussian, which is fully described by its mean and cross-covariance function, i.e., $\boldsymbol{\mu} : T \longrightarrow \mathbb{R}^2$ and $\boldsymbol{K} : T \times T \longrightarrow \boldsymbol{M}_{2\times 2}$ respectively. Both need to be specified as shown in Section 6.2.3. Similar to Eq. 6.15, the bivariate Gaussian process $\boldsymbol{\theta}_n(t)$, $t \in [0, T]$, can be written as

$$\boldsymbol{\theta}_n(t) = \boldsymbol{\mu}(t) + \bar{\boldsymbol{\theta}}_n(t) \tag{6.16}$$

where $\bar{\boldsymbol{\theta}}_n(t) \sim \mathcal{MGP}_2(\mathbf{0}, \boldsymbol{K})$ is a zero mean bivariate Gaussian process with a covariance function $\boldsymbol{K}$.

We parametrise $\boldsymbol{K}$ using a bivariate Matérn model which is defined in Section 6.2.4. The mean function $\boldsymbol{\mu}$ of each constituent process is, then, assumed to be a smooth function of time and can be parametrised accordingly. For instance, if we use basis

functions, the mean vector can be written as

$$\boldsymbol{\mu}(t) = \begin{bmatrix} \mu_1(t) \\ \mu_2(t) \end{bmatrix} = \begin{bmatrix} \gamma_{10} + \gamma_{11}B_{11}(t) + \ldots + \gamma_{1H_1}B_{1H_1}(t) \\ \gamma_{20} + \gamma_{21}B_{21}(t) + \ldots + \gamma_{2H_2}B_{2H_2}(t) \end{bmatrix} = \begin{bmatrix} \boldsymbol{B}_1(t)\boldsymbol{\gamma}_1^{\mathrm{T}} \\ \boldsymbol{B}_2(t)\boldsymbol{\gamma}_2^{\mathrm{T}} \end{bmatrix}, \tag{6.17}$$

where $\{B_{dh}(\cdot) : h = 1, \ldots, H_d,\ d = 1, 2\}$ are pre-specified basis functions on $[0, T]$ of some degree with H_d number of knots and γ_{dh} are model parameters with $h = 1, \ldots, H$ and $d = 1, 2$.

Population Level Inference

Let the parameter vector $\boldsymbol{\Psi} = (\{\alpha_{id} : i = 1, \ldots, I,\ d = 1, 2\}, \{b_{i,c_i} : i = 1, \ldots, I,\ c_i = 1, \ldots, C_i\}, \{\gamma_{dh} : h = 1, \ldots, H,\ d = 1, 2\}, \{\sigma_d : d = 1, 2\}, \{\xi_{dd} : d = 1, 2\}, \{\nu_{dq} : d = 1, 2,\ q = 1, 2\}, \{\beta_{dq}; d = 1, 2,\ q = 1, 2\})$. We, additionally, let $\boldsymbol{Y} = [\boldsymbol{Y}_1, \boldsymbol{Y}_2, \ldots, \boldsymbol{Y}_N]^{\mathrm{T}}$ and $\boldsymbol{\Theta} = [\boldsymbol{\theta}_1, \boldsymbol{\theta}_2, \ldots, \boldsymbol{\theta}_N]^{\mathrm{T}}$, where $\boldsymbol{Y}_n$ is the $J_n \times I$ matrix of individual n responses with $i \in \{1, 2, \ldots, I\}$ and $t_1, t_2, \ldots, t_{J_n} \in [0, T]$; and, $\boldsymbol{\theta}_n = [\boldsymbol{\theta}_{1n}, \boldsymbol{\theta}_{2n}]^{\mathrm{T}}$ with $\boldsymbol{\theta}_{dn} = [\theta_{dn}(t_1), \theta_{dn}(t_2), \ldots, \theta_{dn}(t_{J_n})]^{\mathrm{T}}$, $d \in \{1, 2\}$.

Under the above formulation, the marginal likelihood $\mathcal{L}(\boldsymbol{\Psi}; \boldsymbol{y})$ is expressed as

$$\mathcal{L}(\boldsymbol{\Psi}; \boldsymbol{y}) = \int P(\boldsymbol{Y} = \boldsymbol{y}|\boldsymbol{\Theta}, \boldsymbol{\Psi}) f(\boldsymbol{\Theta}|\boldsymbol{\Psi}) d\boldsymbol{\Theta} \tag{6.18}$$

which is a high-dimensional integral which cannot be solved analytically (intractable). To overcome this issue and maximise Eq. 6.18, we can implement an Expectation-Maximisation (EM) algorithm. However, considering the joint likelihood of $\boldsymbol{y}$ and $\boldsymbol{\Theta}$ for $D = 2$,

$$\mathcal{L}(\boldsymbol{\Psi}; \boldsymbol{y}, \boldsymbol{\Theta}) = \prod_{n=1}^{N} \Bigg\{ \prod_{i=1}^{I} \prod_{j=1}^{J_n} \Bigg[\Phi\Big(b_{i,y_{ni}(t_{nj})+1} + \sum_{d=1}^{D} \alpha_{id}\theta_{dn}(t) \Big) - \Phi\Big(b_{i,y_{ni}(t_{nj})} + \sum_{d=1}^{D} \alpha_{id}\theta_{dn}(t) \Big) \Bigg] \Bigg\} \times$$
$$(2\pi)^{-\frac{DJ_n}{2}} |\Sigma|^{-\frac{1}{2}} \exp\Big[-\frac{1}{2} (\boldsymbol{\theta}_n - \boldsymbol{\mu}_n)^T \Sigma^{-1} (\boldsymbol{\theta}_n - \boldsymbol{\mu}_n) \Big] \tag{6.19}$$

the conditional distribution of $\boldsymbol{\Theta}$ given $\boldsymbol{y}$ is not easy to find. Hence, instead of the regular EM-algorithm we implement a Stochastic EM-algorithm (StEM; Nielsen (2000)). To do so, we use the formulation in Eq. 6.14 to obtain the augmented joint log-likelihood

$$\ell(\boldsymbol{\Psi}; \boldsymbol{y}, \boldsymbol{y}^*, \boldsymbol{\Theta}) = \sum_{n=1}^{N} \Bigg\{ \sum_{i=1}^{I} \sum_{j=1}^{J_n} \mathbf{1}_{b_{i,y_{ni}(t_{nj})} < y^*_{ni}(t_{nj}) \leq b_{i,y_{ni}(t_{nj})+1}} \times$$

$$\log\Bigg[N(y^*_{ni}(t_{nj}); -\sum_{d=1}^{D}\alpha_{id}\theta_{dn}(t_{nj}))\Bigg]\Bigg\} + \log\Bigg[MVN_{DJ_n}(\boldsymbol{\theta}_n; \boldsymbol{\mu}_n, \Sigma)\Bigg]. \tag{6.20}$$

For the StEM-algorithm, we then iterate between a StE-step, where we sample from the conditional distribution of Θ using Gibbs Sampling, and a M-step where we maximise the joint log-likelihood obtained from Eq. 6.19. A detailed overview of the algorithm can be found in Supplementary Text A.3 of the Appendix. Note that for the maximisation step, we used a limited-memory modification of the BFGS quasi-Newton method, i.e., the L-BFGS-B algorithm (Zhu et al., 1995). The algorithm allows box constraints, i.e., generates a sequence of strictly feasible solutions to the maximisation problem, whilst promoting computational efficiency when the number of parameters is large.

Individual Level Inference

To get estimates for an individual latent trait level at a particular point in time (individual D-variate Gaussian process at time t), we use the Expected a Posteriori estimates for $\boldsymbol{\theta}_n(t),\ t \in [0,T],\ n \in N$, i.e.,

$$\hat{\boldsymbol{\theta}}_{n,EAP} = \mathrm{E}[\boldsymbol{\theta}_n|\boldsymbol{Y}_n, \hat{\boldsymbol{\Psi}}] = \int \boldsymbol{\theta}_n P(\boldsymbol{\theta}_n|\boldsymbol{Y}_n, \hat{\boldsymbol{\Psi}}) d\boldsymbol{\theta}_n.$$

To compute the above integral we use Monte Carlo integration (Gibbs-step of the StEM-algorithm) to draw M samples from the posterior $P(\boldsymbol{\theta}_n|\boldsymbol{Y}_n, \hat{\boldsymbol{\Psi}})$.

Given those estimates, to get estimated process values $\boldsymbol{\theta}_n(t)$ for all $t \in [0,T]$, we interpolate the missing $\boldsymbol{\theta}_n(t)$ values at $t \in [0,T] \setminus \{t_{n1}, t_{n2}, \ldots, t_{nJ_n}\}$ as follows:

$\hat{\boldsymbol{\theta}}_n(t_{new}) = \hat{\boldsymbol{\mu}}_{\boldsymbol{\theta}_n(t_{new})|\boldsymbol{\theta}_n(t_{n1}),\boldsymbol{\theta}_n(t_{n2}),\ldots,\boldsymbol{\theta}_n(t_{nJ_n})} = \hat{\boldsymbol{\mu}}_{\boldsymbol{\theta}_n(t_{new})} + \hat{\Sigma}_{\boldsymbol{\theta}_n(t_{new})}\hat{\Sigma}^{-1}_{\boldsymbol{\theta}_n}(\hat{\boldsymbol{\theta}}_n - \hat{\boldsymbol{\mu}}_{\boldsymbol{\theta}_n})$

since

$$\begin{bmatrix}\boldsymbol{\theta}_n(t_{new})\\ \boldsymbol{\theta}_n\end{bmatrix} \sim MVN_{2(J_n+1)}\Bigg(\begin{bmatrix}\hat{\boldsymbol{\mu}}_n(t_{new})\\ \hat{\boldsymbol{\mu}}_{\boldsymbol{\theta}_n}\end{bmatrix}, \begin{bmatrix}\hat{\sigma}^2 & \hat{\Sigma}_{\boldsymbol{\theta}_n(t_{new})}\\ \hat{\Sigma}^T_{\boldsymbol{\theta}_n(t_{new})} & \hat{\Sigma}_{\boldsymbol{\theta}_n}\end{bmatrix}\Bigg). \tag{6.21}$$

6.3.2 Identification and Other Constraints

To ensure model identifiability and covariance function validity, we applied constraints on both the Univariate LGP (ULGP) and BiLGP models. In particular, similar to structural equation modelling, we fix each $\gamma_{d0} = 0$ for all $d \in \{1,2\}$ (avoids mean shift) and $\sigma_{11} = \sigma_{22} = 1$ (fixes the scale of the latent process). Notice that, alternatively, someone can fix the first factor loading of each dimension to 1 and the first threshold

of one of the items from each dimension to 0 (an example is presented in Section 6.4). For the bivariate model, in particular, we further constrain the Matérn covariance and assume

$$v_{12} = \frac{v_{11} + v_{22}}{2}, \quad \xi_{12} = \xi_{11} = \xi_{22}, \tag{6.22}$$

$$\sigma_{12} = \sigma_{11} = \sigma_{22} = 1. \tag{6.23}$$

Variance and scale parameter constraints were applied as previously suggested in relevant literature (Apanasovich and Genton, 2010). This parsimonious formulation ensures cross-covariance symmetry and nonnegative definiteness of the BiLGP.

In the remainder of this chapter, we assume a bivariate structure with all cross-loadings being equal to zero as no item is allowed to inform more than one factor (latent process). As shown in Chapter 3, this assumption is reasonable for model implementation using the ageing domain data from the ELSA. However, structural misspecification becomes a major issue if the latent factor structure is not clear, i.e., items loading to more than one factors (latent processes); fitting a bivariate model to a process where there is really only a single univariate latent process; or, having a static factor instead of a time-varying latent process. Methods for proper assessment have not yet being developped, but looking at factor (latent process) loadings and latent process mean function parameter estimates can provide some insight on this matter. For example, low factor loadings after fitting the BiLGP model could suggest that the bivariate process structure might be inappropriate and a univariate model may be used instead. Proper investigation before and after model fitting is essential to avoid lack of model fit and erroneous conclusions being drawn.

6.3.3 Computational Complexity

As shown in previous work on Gaussian processes, this methodology suffers from high-dimensionality problems (Vandenberg-Rodes and Shahbaba, 2015). In particular, computational cost depends on multiple factors which include the sample size N, the number of repeated measurements per individual J_n, the number of items I and the number of thresholds to be estimated. All of these factors can make inference quite complex and computationally cumbersome even for the univariate case.

Looking at the bivariate extension, inference necessitates even more computational power since dimensionality increases vastly. Previous literature in multivariate Matérn processes, proposes a two step approach for model implementation where each univariate model is fitted separately and parameter estimates are, then, used as fixed to facilitate estimation of the remaining parameters of the multivariate covariance. In particular, for the purposes of the ELSA application, we first get parameter estimates by fitting the ULGP model to each domain data separately and, then, proceed with estimating the cross-correlation parameter for the joint cross-covariance function of the bivariate model. This technique has been previously implemented and was proven to provide valid results under a similar framework (Vandenberg-Rodes and Shahbaba, 2015). Notice that, since after the univariate application some $\xi_{d_1} \neq \xi_{d_2}$, for the bivariate model we assume that $\xi_{11} = \xi_{22} = \xi_{12} = \frac{\xi_{d_1}+\xi_{d_2}}{2}$. This is not expected to vastly affect inference results, since for large values and small changes of the scale parameter, Gaussian function smoothness minorly changes, as shown in Figure 6.2.

R version 3.6.3 was used to perform all analyses within this chapter. Similar to Chapter 4, both the simulation study and the data application were performed using The High End Computing Cluster at Lancaster University. R code was developed to implement the ULGP and BiLGP models. The Rcpp R package was used to improve computational time. Parameter estimation of each ULGP model implemented using ageing domain data (10000 iterations) ranged from 24 to 190 hours approximately and differed according to the number of parameters within each individual model.

6.3.4 Parameter Tuning

Determining the value for the number of interior knots for the B-spline basis in Eq. 6.17 is essential to accurately approximate the mean function of the LGP structure. Common methods for selecting the number of interior knots for the B-spline basis include cross-validation or, more sophisticated, Bayesian algorithms (DiMatteo et al., 2001). However, due to the computational complexity of the proposed algorithm such methods could be computationally cumbersome or, even, infeasible in practice. To identify the ideal number of knots for data application purposes, we initially used simulated data and developed several different criteria based on which we compared performance under varying scenarios for the number of interior knots. We build a BIC-like criterion following the proposed

algorithm of Ibrahim et al. (2008) and we also used the "elbow" method (Thorndike, 1953), where determination of the number of interior knots to be used is based on the location of the "elbow" of the model performance curve, i.e., performance improves in a slower rate considering the trade-off between the number of added parameters in the model and the performance benefits. We validated any results obtained by exploring both methods' properties using simulated data (Section 6.4).

6.4 Simulation Studies

In this section, we explore the properties of the BiLGP model with parameter estimation and individual trajectory estimation accuracy, as discussed in Section 6.3.1. We do so, after having investigated the performance of the ULGP model for the assumed type of data. Indicatively, algorithm convergence is also assessed for some of the examined scenarios.

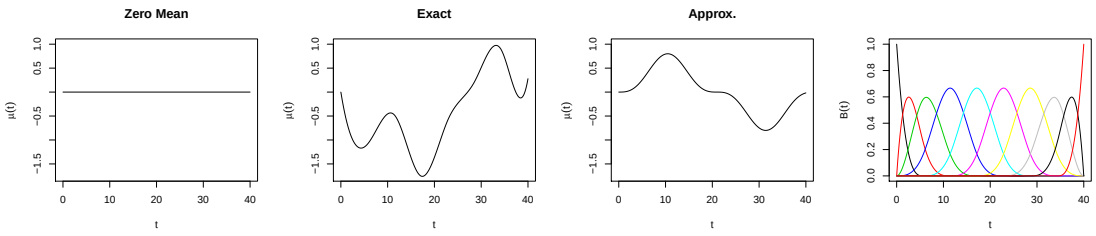


Figure 6.3: Mean curves of the LGPs assumed generated for the simulation studies and B-splines basis functions used for their approximation (6 interior knots).

First, we simulated time points $t_{nj} \in [0, 40]$ and randomly sampled $J_n \in \{2, 3, \ldots, 8\}$, for all $n \in \{1, 2, \ldots, N\}$ to mimic the data structure observed in ELSA. Our first aim was to assess parameter estimation and latent curve estimation accuracy under varing sample size, latent mean curve and covariance parameter scenarios. Both identification constraint options where considered to ensure that they do not affect parameter estimation and algorithm convergence. Responses were generated based on the simulation regimes shown in Table 6.1. Notice that for the BiLGP model ξ_{dd} is the corresponding scale parameter common across all constituent processes, i.e., $\xi_{11} = \xi_{22}$ (parsimonious Matérn model of Genton and Kleiber (2015)). We additionally assume that $\nu = 5/2$ for all constituent processes. We considered three scenarios for the LGP mean function: $\mu(t) = 0$ (zero mean); $\mu(t)$ generated using prespecified coefficients for the cubic

ULGP Model				
Scenario	**Mean Function**	**Ident. Constraints**	**Varying**	**Correlation**
1A	$\mu(t) = 0$	$\gamma_0 = 0$ and $\sigma = 1$	$N \in \{500, 1000, 2000\}$, $\xi = 10$	-
1B	$\mu(t)$ generated using prespecified coefficients for the cubic B-splines	$\gamma_0 = 0$ and $\sigma = 1$	$N \in \{500, 1000, 2000\}$, $\xi = 10$	-
1C	$\mu(t) = 0.8\sin^3(0.15t)$	$\gamma_0 = 0$ and $\sigma = 1$	$N \in \{500, 1000, 2000\}$, $\xi = 10$	-
1D	$\mu(t) = 0.8\sin^3(0.15t)$	$\alpha_1 = 1$ and $b_{1,1} = 0$	$N \in \{500, 1000, 2000\}$, $\xi = 10$	-
1E	$\mu(t) = 0.8\sin^3(0.15t)$	$\alpha_1 = 1$ and $b_{1,1} = 0$	$\xi \in \{5, 25, 35\}$, $N = 250$	-
BiLGP Model				
2A	$\mu_1(t) = 0$ and $\mu_2(t) = 0$	$\gamma_{10} = 0$, $\gamma_{20} = 0$, $\sigma_1 = 1$, $\sigma_2 = 1$	$N \in \{100, 500, 1000\}$, $\xi_{11} = \xi_{22} = 10$	$\beta_{12} \in \{-0.2, -0.8, 0, 0.2, 0.8\}$
2B	$\mu_1(t) = 0$ and $\mu_2(t)$ generated using prespecified coefficients for the cubic B-splines	$\gamma_{10} = 0$, $\gamma_{20} = 0$, $\sigma_1 = 1$, $\sigma_2 = 1$	$N \in \{100, 500, 1000\}$, $\xi_{11} = \xi_{22} = 10$	$\beta_{12} \in \{-0.2, -0.8, 0, 0.2, 0.8\}$
2C	$\mu_1(t) = 0$ and $\mu_2(t) = 0.8\sin^3(0.15t)$	$\gamma_{10} = 0$, $\gamma_{20} = 0$, $\sigma_1 = 1$, $\sigma_2 = 1$	$N \in \{100, 500, 1000\}$, $\xi_{11} = \xi_{22} = 10$	$\beta_{12} \in \{-0.2, -0.8, 0, 0.2, 0.8\}$
2D	$\mu_1(t)$ generated using prespecified coefficients for the cubic B-splines and $\mu_2(t) = 0.8\sin^3(0.15t)$	$\gamma_{10} = 0$, $\gamma_{20} = 0$, $\sigma_1 = 1$, $\sigma_2 = 1$	$N \in \{100, 500, 1000\}$, $\xi_{11} = \xi_{22} = 10$	$\beta_{12} \in \{-0.2, -0.8, 0, 0.2, 0.8\}$
2E	$\mu_1(t) = 0$ and $\mu_2(t) = 0.8\sin^3(0.15t)$	$\alpha_1 = 1$, $\alpha_5 = 1$, $b_{5,1} = 0$, $b_{1,1} = 0$	$N \in \{100, 500, 1000\}$, $\xi_{11} = \xi_{22} = 10$	$\beta_{12} = -0.5$
2F	$\mu_1(t) = 0$ and $\mu_2(t) = 0.8\sin^3(0.15t)$	$\alpha_1 = 1$, $\alpha_5 = 1$, $b_{5,1} = 0$, $b_{1,1} = 0$	$\xi_{11}, \xi_{22} \in \{5, 25, 35\}$, $N = 250$	$\beta_{12} = -0.5$

Table 6.1: Simulation study regimes for the ULGP and BiLGP modelling frameworks.

B-splines (exact), in particular $\mu(t) = \boldsymbol{B}(t)\boldsymbol{\gamma}^{\mathrm{T}}$ where $\boldsymbol{B}(t)$ is the B-splines basis function matrix and $\boldsymbol{\gamma}^{\mathrm{T}} = [-1.15, -1.45, 0.31, -2.66, -0.19, -0.10, 1.87, -0.78, 0.28]$; and, $\mu(t) = 0.8\sin^3(0.15t)$ (Figure 6.3). Multiple scenarios were also considered for the samples sizes, scale parameters and cross-correlation parameter of the bivariate Matérn covariance. For all scenarios, cubic B-splines with 6 interior knots were used (Figure 6.3).

Parameter estimates were computed based on 5000 repetitions for each of the R=100 data sets generated per scenario. Performance was measured using the MAE, the relative error (RE) and the root mean integrated squared error (RMISE) across all data replications R when appropriate. We define the RMISE as

$$\frac{1}{R}\sum_{R}^{r=1}\sqrt{\int_0^8 (\mu_d(s) - m_d^{(r)}(s))^2 ds} \tag{6.24}$$

where $m_d^{(r)}(s)$ is the estimate of $\mu_d(s)$ in the r^{th} replication. Algorithm convergence was also assessed using the Rubin-Gelman $\hat{R}$ (total chain variance over within chain variance; Gelman, 1996) which was less than 1.2 and ≈ 1 under all N for all parameters. Figures B.17 to B.21 show the GR statistics using 5 chains and varying parameter initialisation scenarios. We noticed that a burn-in of 10% of the chain length should be sufficient to ensure convergence under all simulation scenarios.

Table 6.2 shows parameter estimation accuracy results of the ULGP model (scenarios

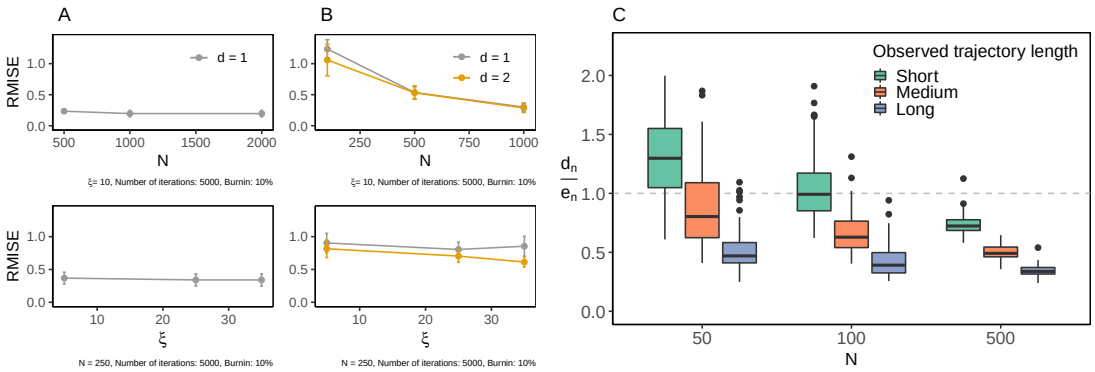


Figure 6.4: RMISE for the ULGP model; results shown correspond to simulation regimes 1D and 1E (A). RMISE for the BiLGP model; results shown correspond to simulation regimes 2E and 2F (B); and, individual curve estimation accuracy for the BiLGP model under multiple scenarios for the length of the observed trajectory of each individual (C).

1A to 1D) whereas Figure 6.6 illustrates the corresponding performance results for the BiLGP model for selected correlation scenarios (2A to 2E). We observed that parameter estimates are relatively robust and high accuracy is achieved even for small sample sizes. Notice that the MAE of ξ (scale parameter) is high when we look at the MAE scores but that is due to its large true value (for all scenarios $\xi_{11} = \xi_{22} = 10$) since small changes in its value only slightly affect process roughness (Figure 6.2). We confirmed this by looking at the corresponding relative error for this particular parameter which was less than 1 for all simulation regimes. Accuracy is not affected by varying ξ_{dd} values which indicates that process smoothness does not affect algorithm convergence and estimation (Figures 6.4A and 6.4B). The BiLGP method showed high performance in approximating both latent mean functions of the bivariate Gaussian process (Figures 6.4B and 6.5). As expected, larger sample size promoted parameter and latent mean curve estimation accuracy.

Individual trajectory estimation accuracy was also explored under varying sample size and generated data scenarios. In particular, three J_n scenarios: short ($2 \leq J_n \leq 8$) medium ($8 \leq J_n \leq 14$) and long trajectories ($14 \leq J_n \leq 20$). The sample sizes examined were 50, 100 and 500 subjects. The number of samples generated for each scenario combination was 100. The Expected a Posteriori (EAP) estimates were used. An overall accuracy score was obtained as follows:

$$o_n = \frac{d_n}{e_n} \text{ where}$$

$$d_n = \sqrt{\int_0^T (\theta_n(t) - \hat{\theta}_{n,EAP})^2 dt} \text{ and } e_n = \sqrt{\int_0^T (\theta_n(t) - \hat{m}(t))^2 dt}.$$

This showed if individual trajectory estimates approximate the true individual curves better compared to the mean function estimate. The o_n scores are provided in Figure 6.4. Even though the method does not perform well for short trajectories and $N = 50$, increasing sample size and longer observed trajectories both improve individual curve estimation accuracy. This is reasonable since when the length of the observed individual trajectories relative to the time interval of interest is small, the interpolation and extrapolation done to get the EAP estimates as well as the $\hat{\theta}_{nd}(\cdot)$ have larger deviations from the true values outside each individual trajectory range.

	N	α_1	α_2	α_3	α_4	b_1	b_2	b_3	b_4	ξ	γ_0	γ_1	γ_2	γ_3	γ_4	γ_5	γ_6	γ_7	γ_8	γ_9	σ
True value		1	1.20	0.90	1	0	-0.50	-0.20	-0.40	10	0	0	0	0	0	0	0	0	0	0	1
MAE (zero)	500	4.44e-02	5.88e-02	4.38e-02	3.90e-02	1.38e-01	1.62e-01	1.30e-01	1.39e-01	7.49e-01	-	2.32e-01	1.56e-01	1.84e-01	1.66e-01	1.74e-01	1.66e-01	1.90e-01	1.98e-01	1.93e-01	-
	1000	3.47e-02	3.26e-02	2.81e-02	3.25e-02	1.31e-01	1.54e-01	1.14e-01	1.32e-01	5.26e-01	-	2.15e-01	1.18e-01	1.67e-01	1.45e-01	1.50e-01	1.47e-01	1.58e-01	1.62e-01	1.65e-01	-
	2000	2.08e-02	3.06e-02	2.18e-02	2.17e-02	6.56e-02	7.94e-02	6.20e-02	7.05e-02	2.59e-01	-	1.28e-01	7.73e-02	8.67e-02	8.02e-02	9.48e-02	8.47e-02	1.01e-01	9.51e-02	9.77e-02	-
True value		1	1.20	0.90	1	0	-0.50	-0.20	-0.40	10	0	-1.15	-1.45	0.31	-2.66	-0.19	-0.10	1.87	-0.78	0.28	1
MAE (exact)	500	4.88e-02	5.61e-02	4.56e-02	4.60e-02	1.67e-01	1.85e-01	1.48e-01	1.63e-01	8.89e-01	-	3.20e-01	2.34e-01	2.24e-01	2.32e-01	2.00e-01	2.04e-01	2.13e-01	2.24e-01	2.26e-01	-
	1000	3.45e-02	3.79e-02	2.80e-02	3.17e-02	1.09e-01	1.32e-01	9.79e-02	1.13e-01	5.23e-01	-	2.12e-01	1.49e-01	1.71e-01	1.48e-01	1.28e-01	1.35e-01	1.30e-01	1.59e-01	1.40e-01	-
	2000	2.15e-02	3.12e-02	1.68e-02	2.37e-02	7.08e-02	8.58e-02	6.65e-02	7.30e-02	3.26e-01	-	1.49e-01	1.01e-01	1.16e-01	9.95e-02	9.71e-02	9.92e-02	9.72e-02	1.12e-01	9.62e-02	-
True value		1	1.20	0.90	1	0	-0.50	-0.20	-0.40	10	0	-0.09	0.31	1.16	-0.19	0.22	-0.72	-0.91	0.08	-0.07	1
MAE (approx.)	500	4.87e-02	5.73e-02	3.78e-02	3.83e-02	1.46e-01	1.75e-01	1.33e-01	1.46e-01	7.88e-01	-	2.54e-01	1.73e-01	2.01e-01	1.69e-01	1.85e-01	1.82e-01	2.23e-01	2.12e-01	2.09e-01	-
	1000	3.32e-02	3.37e-02	2.87e-02	2.87e-02	1.28e-01	1.51e-01	1.09e-01	1.29e-01	5.01e-01	-	2.18e-01	1.28e-01	1.73e-01	1.47e-01	1.45e-01	1.62e-01	1.55e-01	1.76e-01	1.59e-01	-
	2000	1.98e-02	2.68e-02	2.14e-02	1.99e-02	6.54e-02	7.73e-02	6.22e-02	6.90e-02	2.60e-01	-	1.14e-01	8.64e-02	9.25e-02	8.38e-02	9.29e-02	9.31e-02	1.07e-01	1.03e-01	9.10e-02	-
True value		1.00	1.20	0.90	2.20	0.00	-0.50	-0.20	0.90	10.00	0.01	-0.08	0.31	1.16	-0.19	0.23	-0.73	-0.90	0.07	-0.04	1.00
MAE (approx.)	500	-	5.77e-02	5.00e-02	1.17e-01	-	3.03e-02	3.26e-02	5.49e-02	7.99e-01	8.27e-02	1.33e-01	9.99e-02	9.72e-02	8.71e-02	1.36e-01	1.58e-01	1.42e-01	9.52e-02	8.61e-02	3.69e-02
	1000	-	5.57e-02	4.87e-02	1.24e-01	-	3.60e-02	2.21e-02	4.41e-02	2.55e-01	7.46e-02	7.08e-02	6.90e-02	1.00e-01	4.69e-02	6.32e-02	1.21e-01	1.19e-01	1.03e-01	7.02e-02	4.75e-02
	2000	-	1.99e-02	2.67e-02	8.38e-02	-	1.94e-02	1.48e-02	3.31e-02	1.64e-01	6.39e-02	6.65e-02	6.16e-02	3.98e-02	3.37e-02	4.24e-02	4.14e-02	6.50e-02	4.50e-02	4.32e-02	1.92e-02

Table 6.2: Parameter estimation accuracy as measured using the Mean Absolute Error for all simulated data scenarios examined for the ULGP model. Results shown correspond to simulation regimes 1A to 1D.

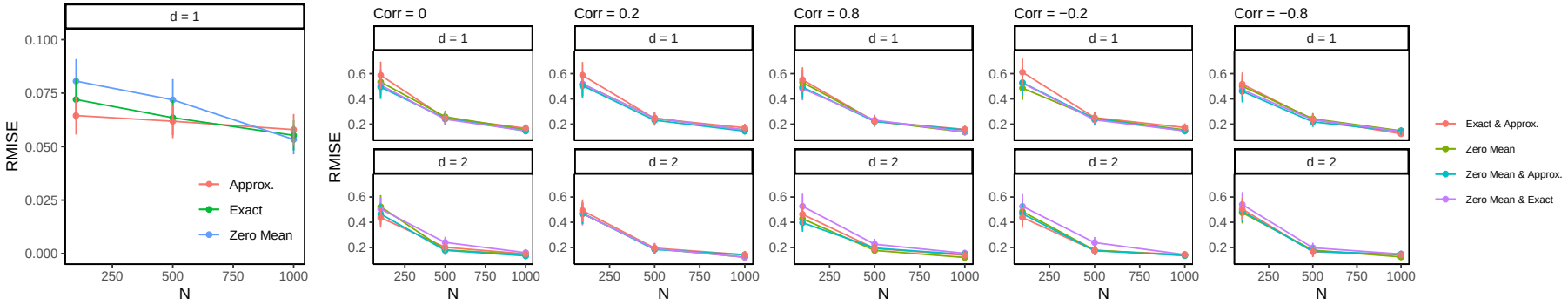


Figure 6.5: RMISE for the ULGP model; results shown correspond to simulation regimes 1A to 1C (left). RMISE for the BiLGP model; results shown correspond to simulation regimes 2A to 2D.

Parameter tuning

As discussed in previous section, there is currently no standard and well established method for tuning the number of interior knots used for the B-spline basis under the proposed framework apart from cross-validation. However, this would be extremely computationally demanding here. So in this paragraph, we explore two alternative methods that could potentially provide some insights on the value range that could improve model fit and avoid underfitting or overfitting. For the "elbow" method, model performance was measured using a Weighted F-score which was averaged across all items. In particular, for each item i, $i = 1, \ldots, I$

$$\text{Weighted } F\text{-score}_i = 2 \times \frac{\text{Precision}_i \times \text{Recall}_i}{\text{Precision}_i + \text{Recall}_i} \tag{6.25}$$

where

$$\text{Precision}_i = \sum_{c=0}^{C_i} N_c \frac{\text{TP}_c}{\text{TP}_c + \text{FP}_c} \text{ and } \text{Recall}_i = \sum_{c=0}^{C_i} N_c \frac{\text{TP}_c}{\text{TP}_c + \text{FN}_c}. \tag{6.26}$$

In contrast to a binary classifier, here TP_c is the number of item results that were correctly classified to level c at prediction, FP_c is the number of item results that were incorrectly classified as c and FN_c is the number of item results that were incorrectly classified into another item level rather than c. Notice that N_c are weights which correspond to the true number of responses from level c. The closer to 1 this constructed F-score is the better the prediction accuracy. For this testing procedure, 100 data samples were generated based on scenario 1D in Table 6.1 and mean function approximation was made by assuming a varying number of interior knots. Sample size was assumed to be equal to 250 individuals and inference was made based on 10000 iterations. For mean function parameters, initial values were assumed to be equal to zero and for the rest, inital values were fixed to the true parameter values. After plotting the number of interior knots against the $(1 - F\text{-score})$, we see that even though accuracy saturates when the number of knots increases, after 6-7 interior knots returns are considerably diminishing taking into account the increase in computational cost (Figure 6.7a).

Next, we built the BIC-like criterion of Ibrahim et al. (2008), let us call it IC. To check its performance under current model assumptions, we generated 100 data samples

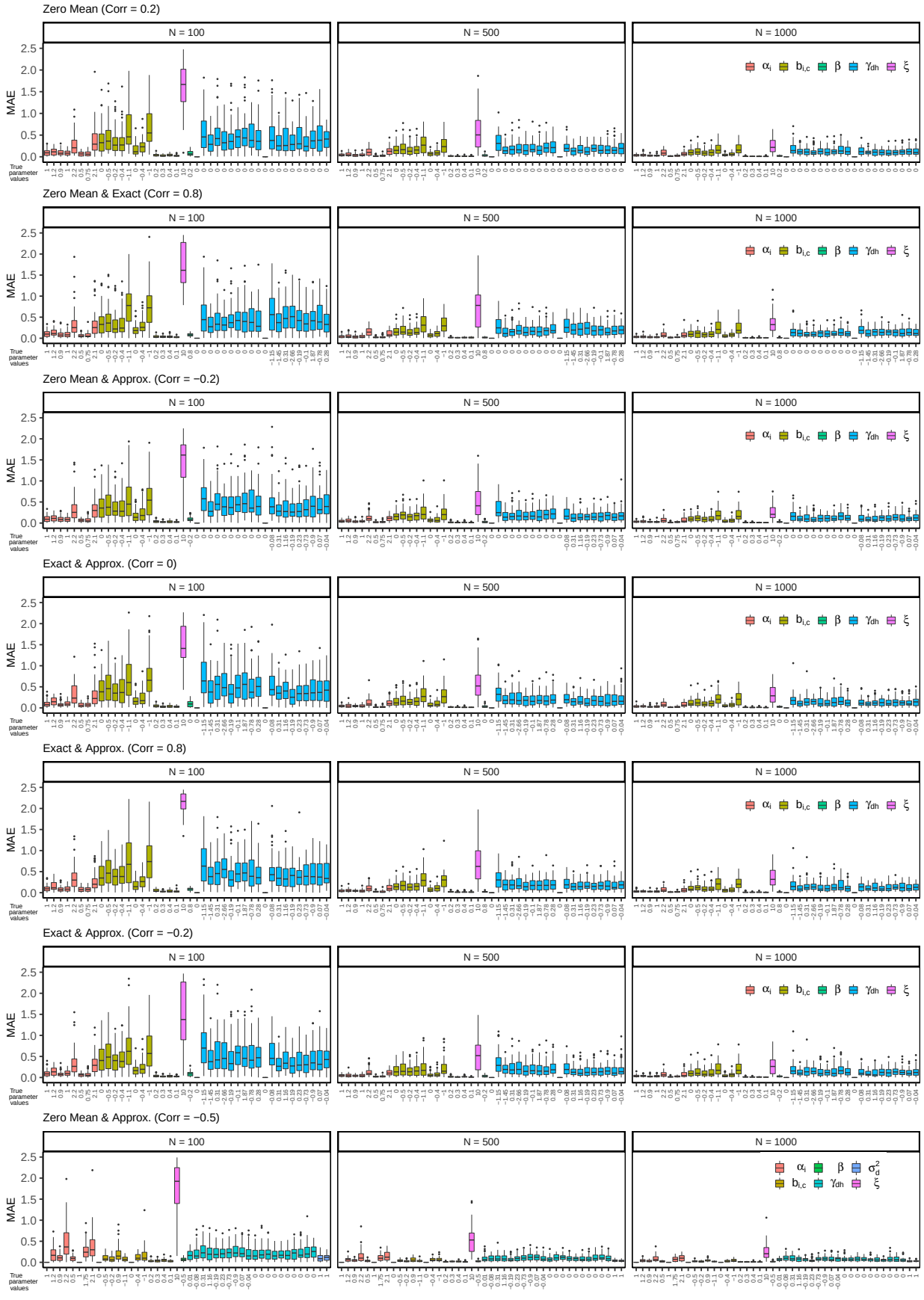
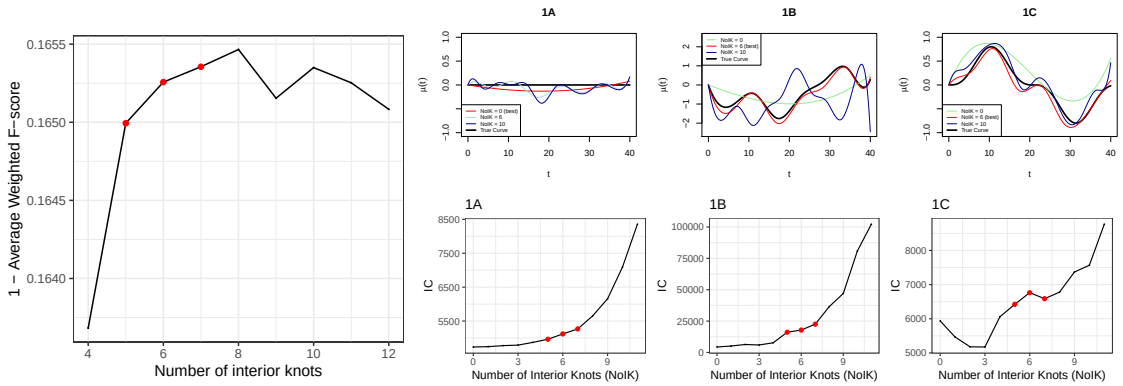


Figure 6.6: Parameter estimates MAE for the BiLGP model across all simulation replications. Results shown correspond to simulation regimes 2A to 2E.



(a) (b)

Figure 6.7: (a) One minus the weighted F-score as averaged across all items, for varying number of interior knots used for the B-spline basis approximation of the mean curve for the ULGP model. The red points highlight the ideal value range. (b) Examples of mean function approximation results under varying number of interior knots (top). IC scores for varying number of interior knots used for the B-spline basis approximation of the mean curve for the ULGP model. The red points highlight the ideal value range as determined using the "elbow" method.

based on 1A-C scenarios of Table 6.1. Figure 6.7b illustrates examples on mean function approximation and measured performance. Based on this measure, the suggested value of interior knots should provide a low IC value. We observed that this measure seems to be sensitive to overfitting but is tolerant to underfitting. However, we observe that a range of good interior knots candidates is given by looking at the point were IC score starts to accelarate. In real data, we will use both the above methods to get a suggested interior knots range but we will also critically examine the obtained results under all proposed interior knots scenarios. Current algortihms showed insufficient performance, therefore, an automatic algorithm to obtain the number of interior knots under this framework could be investigated in the future.

6.5 Analysis of ageing domains using data from the ELSA

In Chapter 3, we performed exploratory analysis of the ELSA data covering a large collection of variables measuring five key ageing domains: physical ability, psychological and social well-being, cognitive function and metabolic health. In the current chapter, we analyse data from four of the above domains to explore trends over time and between domain relationships. The final sample included 8,945 individuals aged 50 to 90 years old from which 53.64% are females, 75.19% cohabiting or married and 60.63% have attended

at least some college and high school. Wealth and social status tertiles were used to distinguish individuals of low (Routine), middle (Intermediate) and upper (Managerial) class (Table 3.1). Unfortunately, metabolic health data were not included in further analysis since the poor measurement of outcomes, low variability, large measurement errors and small number of individuals responding resulted in poor convergence of the algorithm.

As discussed in previous chapters, ageing is a multidimensional construct and, as it has been shown in past studies, there are multiple demographic factors that shape different ageing domain trajectories over age (Tampubolon, 2016a). For this study, we assumed that gender (female and male as reference), threefold social class (top, middle and bottom as reference), wealth tertiles (top, middle and bottom as reference), educational attainments (college and high school or less as reference) and marital status (married or cohabiting and other as reference) have an additive effect on each mean latent curve. This assumption was checked and confirmed using a small simulation study as shown in Supplementary Text A.4 of the Appendix. Even though results suggested that the additive effect assumption is reasonable for most of the explored domains, we further checked for a gender interaction by running the ULGP method for males and females separately. Results indicated that adding a gender interaction could be reasonable for the cognitive function and social well-being domains since crossing between the male and female estimated curves indicated a different rate of change of the mean domain curve over time, an effect which cannot be captured solely by the additive gender effect on the mean domain curve (Figure B.16b).

6.5.1 Univariate analysis of the ageing domains

Before we proceed with the bivariate analysis, we performed univariate analysis of the data for each domain separately. Here, instead of analysing the data looking at the first order latent structure (this would mean analysing data from each ageing subdomain as shown in Figures 3.11 and 3.12), we performed our analysis assuming the higher level second order structure (ageing domains).

Data from each domain were analysed separately. Convergence was reached after 5000 repetitions and confirmed by calculating the Gelman-Rubin statistic for five random initialisation scenarios (Figure B.26 of the Appendix). Mean function estimates were

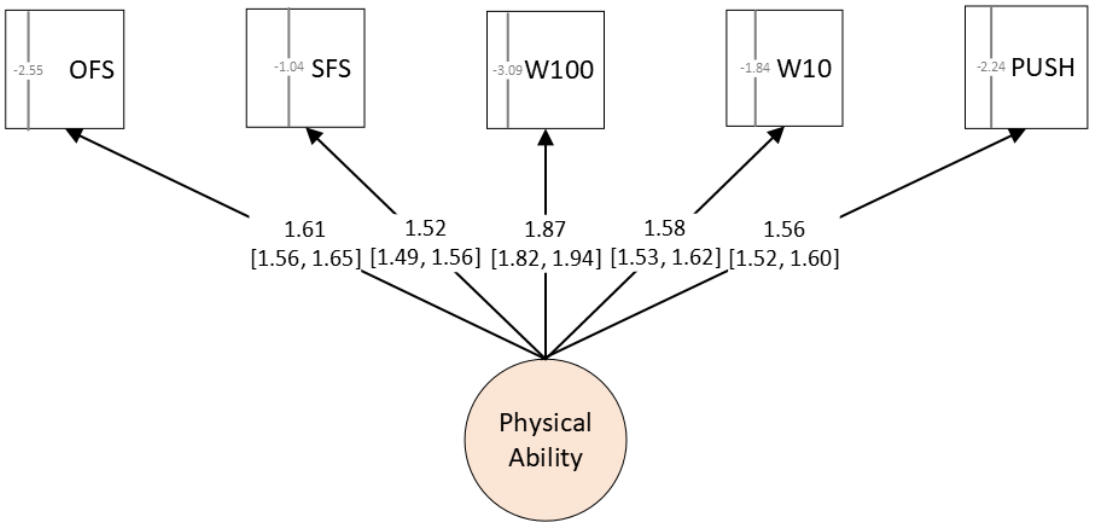


Figure 6.8: Item parameter estimates, 95% confidence intervals and threshold parameter estimates after ULGP model implementation on the physical ability ELSA data. The variables shown include information on: OFS-difficulty in climbing one flight of stairs without resting; SFS-difficulty in climbing several flights of stairs wihout resting; W100-difficulty in walking 100 yards; W10-difficulty in lifting or carrying weights over 10 pounds; PUSH-difficulty in pulling or pushing large objects.

robust and relatively invariant to the number of knots selected (Figure B.28). The chosen number of interior knots for the cubic B-splines used for basis approximation of the $\gamma(\cdot)$ functions were 1 for the CF and 2 for the PA, SW and PW domains respectively following results after implementing the "elbow" and IC score methods from Section 6.3.4 (Figures B.29 and B.30). The results shown were obtained based on the following initialisation scenario: $\alpha_i = 1$, $b_i = 0$ and steps for $b_{ic} - b_{ic-1} = 0.1$, $\xi = 10$, $\gamma_k = 0$. For identifiability reasons $\gamma_0 = 0$ and $\sigma = 1$. The obtained chains are shown in Figures B.22 to B.25 of the Appendix whereas item parameter estimates are shown in Figures 6.8 to 6.11.

We initially fitted the above models using all five covariates, however, the effect of marital status on cognitive function and the effect of education on psychological well-being were approximately equal to zero so they were omitted from the univariate analysis of the corresponding ageing domains. Figure 6.12 shows the results after analysing data from the four domains. The estimated mean function for each ageing domain along with the parameter estimates for the covariates that have been included into the model are illustrated. Approximate 95% confidence intervals were obtained by assuming independence of the produced parameter estimates within each chain and computing the 95% confidence interval of the parameter estimate samples obtained after implementing

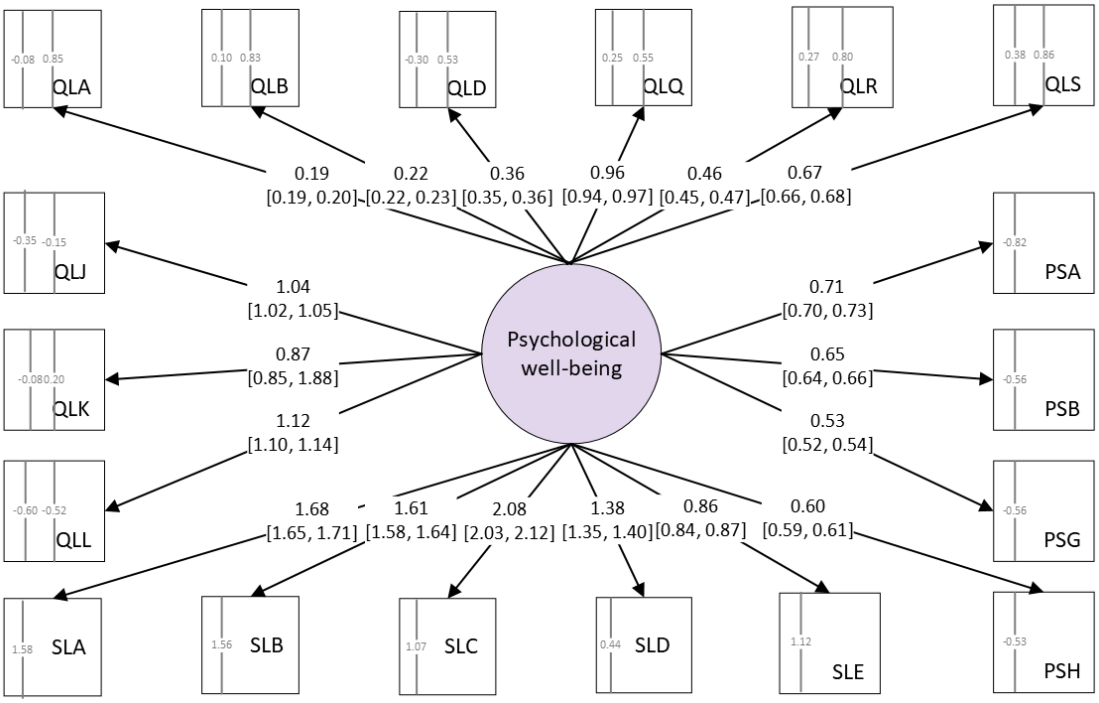


Figure 6.9: Item parameter estimates, 95% confidence intervals and threshold parameter estimates after ULGP model implementation on the psychological well-being ELSA data. The variables shown include information on: QLJ-how often individuals look forward to each day; QLK-how often individuals feel that their life has meaning; QLL-how often individuals enjoy the things they do; QLA-how often individuals feel age prevents them from doing things they like; QLB-how often individuals feel what happens to them is out of their control; QLD-how often individuals feel left out of things; QLQ-how often individuls feel satisfied with the way their life has turned out; QLR-how often individuals feel that life is full of opportunities; QLS-how often individuls feel the future looks good to them; SLA-whether they believe that in most ways their life is close to their ideal; SLB-whether they believe that the conditions of their life are excellent; SLC-whether they are satisfied with their life; SLD-whether, so far, they have got the important things they want in life; SLE-whether they would not change anything, if they could live their life again; PSA-whether individuals felt depressed much of the time during past week; PSB-whether individuals felt that everything they did was an effort; PSG-whether individuals felt sad during past week; PSH-whether individuals could not get going much of the time during past week.

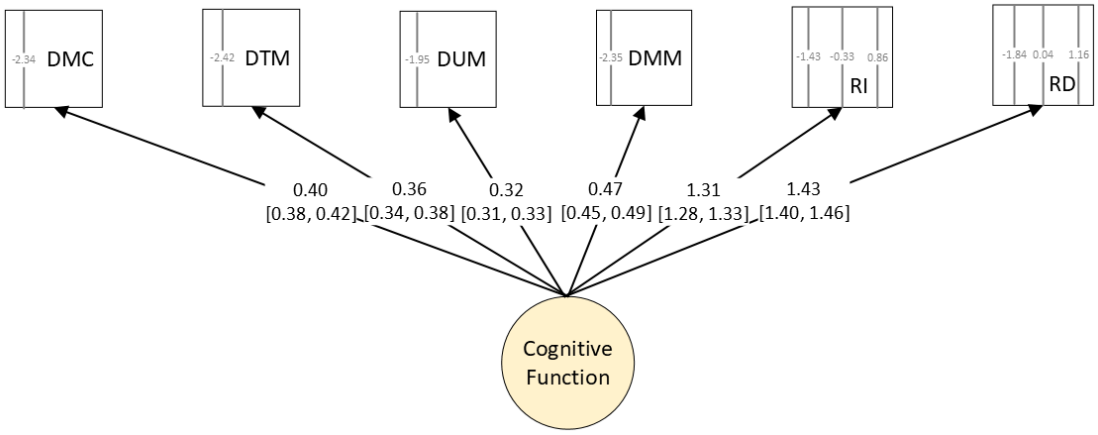


Figure 6.10: Item parameter estimates, 95% confidence intervals and threshold parameter estimates after ULGP model implementation on the cognitive function ELSA data. The variables shown include information on: DMC-difficulty in making phone calls; DTM-difficulty in taking medications; DUM-dIfficulty in using a map; DMM-difficulty in managing money; RI-ability to recall words immediately; RD-ability to recall words after delay.

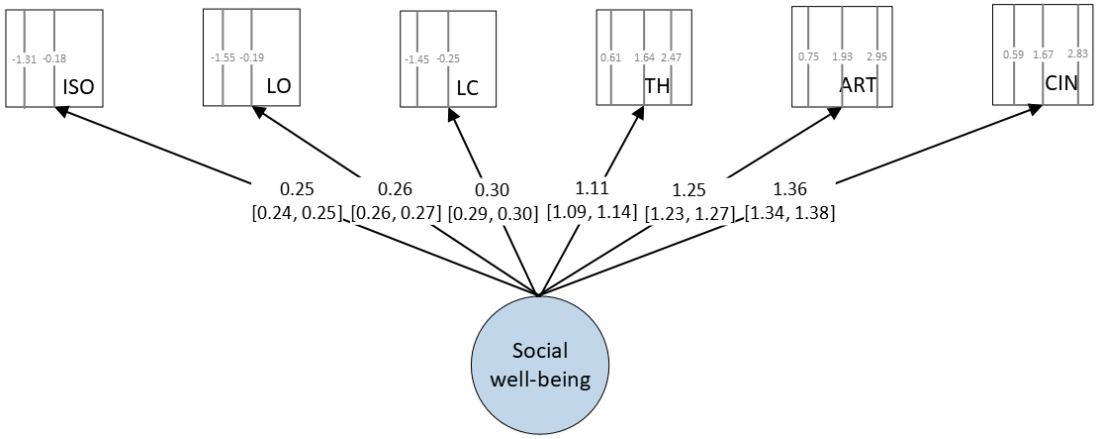


Figure 6.11: Item parameter estimates, 95% confidence intervals and threshold parameter estimates after ULGP model implementation on the social well-being ELSA data. The variables shown include information on: ISO-whether individual feels isolated; LO-whether individual feels left out; LC-whether individual feels lack of companionship; TH-frequency of going to the theatre, a concert or the opera; ART-frequency of going to art gallery or museum; CIN-frequency of going to the cinema.

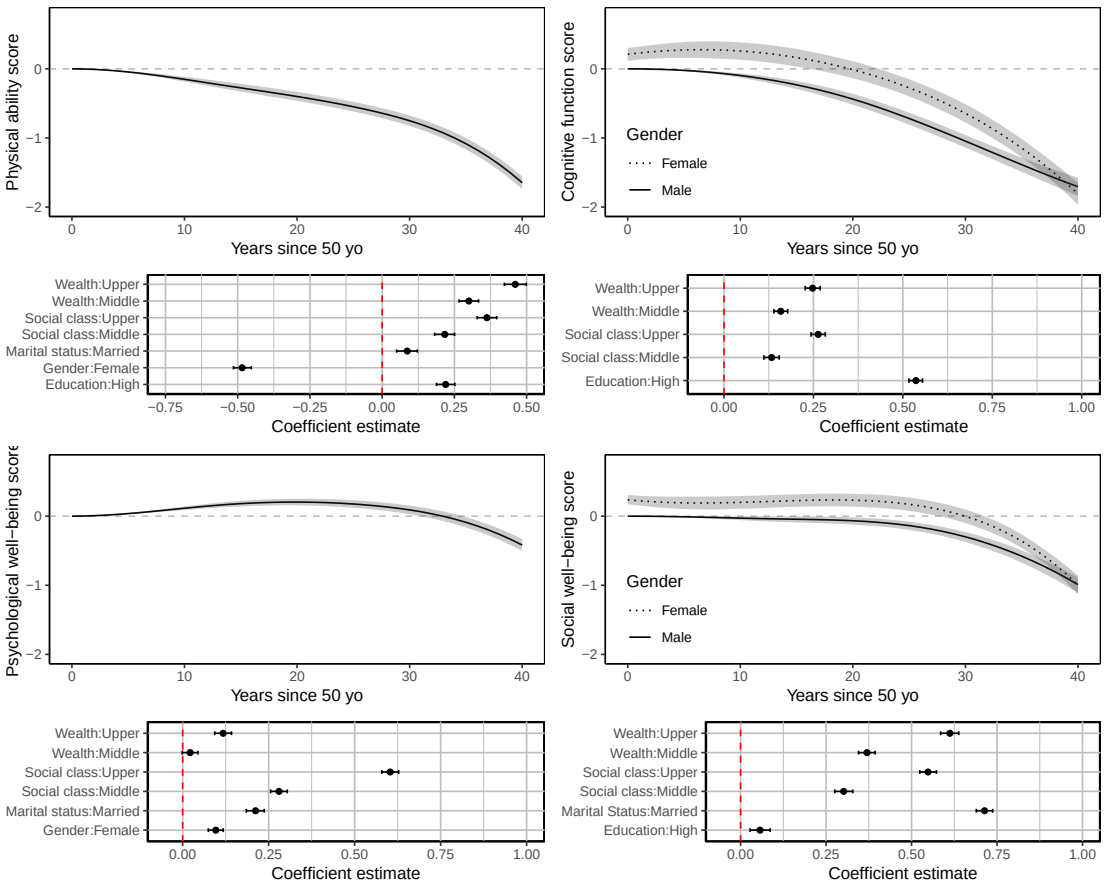


Figure 6.12: Physical ability (PA), cognitive function (CF), psychological well-being (PW) and social well-being (SW) latent mean curve estimates. Covariate effect estimates are depicted for each corresponding domain below the mean curve plot. In both plots approximate 95% confidence intervals are shown.

the StEM-algorithm. On top of the additive gender effect an interaction term was additionally explored separately for all domains and added when appropriate. Gender interaction was considered for the cognitive function and social well-being analysis.

A decrease on PA, CF and SW was observed as people are getting older. Interestingly, PW was the only domain which improves over time and starts to decay only later in life (30 years after the 50 years starting point, i.e people aged over 80 years old). SW also remains stable until 20 years after 50 which coincides to approximately the retirement age of the UK and is currently 66 years old. On the other hand, the PA decay is initiated from the age of 50 years and shows a constant decline until the early 80s. As for CF, results suggest that even though males have an overall lower score, female experience a more intense CF decay as they grow older. In addition, we observe that females live healthier for longer. The only exception is PA which probably has to do with overall muscle growth and strength, both often higher in males. At the same time, being in

the upper class, married and wealthy improved the overall ageing domain score. People with high educational attainment also have better cognitive function and social well-being. Interestingly, even though marital status has a positive effect on social well-being, education and socioeconomic position were found to be more important.

6.5.2 Pairwise bivariate analysis of the ageing domains

After having explored the latent trajectories of the four ageing domains we proceeded with pairwise bivariate analysis of the data. Figure 6.13 shows the estimated correlation structure for the four domains under study. Results indicate an overall positive correlation between most of the domains with stronger correlations between physical ability and psychological well-being. This supports previous findings which link physical strength and exercise with positive mental health outcomes (Stephens, 1988; Delle Fave et al., 2018). Another association was also found between social well-being and physical ability which could also be explained by the fact that people with good physical health tend to socialise more due to their abitlity to go for walks and being outside of their homes, but this was rather small. Regarding the observed dependence between responses from the same domain over time, larger correlations were found for social well-being data, physical ability and psychological well-being, whereas weaker correlations were found between the cognitive function responses, showing that poor cognitive function outcomes at older age cannot necessarily be justified by poor cognitive function outcomes at earlier time points. However, this conclusion needs further investigation as it can be caused due to poor measurement during the data collection or larger measurement errors in this specific part of the data.

The predictive power of the model was, finally, assessed using 10-fold cross-validation where we randomly took segments of individual observations to create the training and test sets for every iteration (Figure B.27). Results showed that high accuracy is achieved for CF, PA and SW data, whereas lower accuracy with large variations between the results is observed for PW data. This decrease in accuracy could potentially be explained by the high variation of responses and the large number of subdomains included in the construction of the second-level PW domain as shown in Figure 3.11. An in depth investigation of these results through sensitivity analysis is recommended, since underlying assumptions might have affected overall model fit and predictive power.

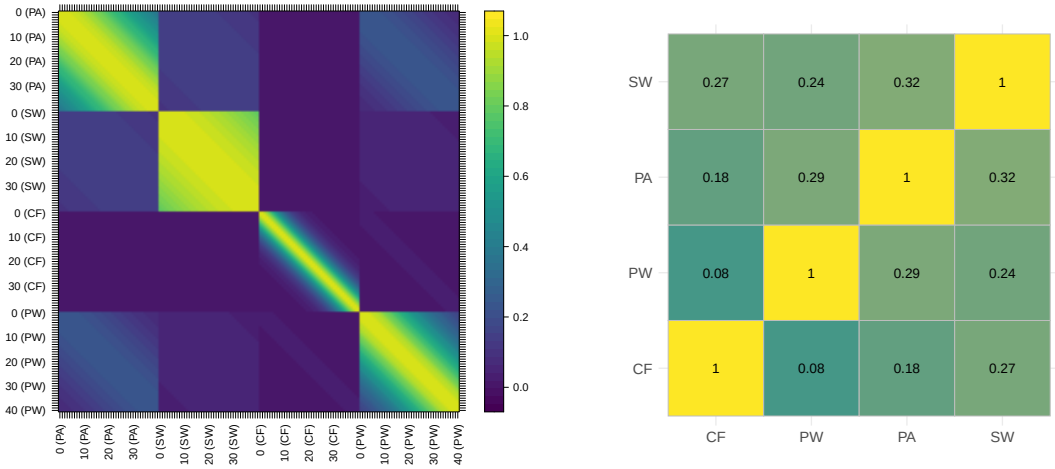


Figure 6.13: (Left) Multivariate Matérn covariance visualisation built using the estimates obtained from the ULGP and BiLGP model fits. Numbers on the horizontal and vertical axis represent years since the individual has reached 50 years and within the parenthesis are the initials of the domain to which each covariance block corresponds. (Right) Estimated correlation structure for all domains as estimated from the pairwise BiLGP model fits. CF, PW, PA and SW correspond to cognitive function, psychological well-being, physical ability and social well-being respectively.

6.5.3 Exploring cohort effects

We explored the presence and intensity of cohort effects into the individual responses to assess how these could potentially affect our latent trait trajectory estimates. We cross-sectionally grouped the individuals into 4 age groups (50-59, 60-69, 70-79, 80+) and compared the proportion of ordinal responses for the same age-group across the 8 data collection waves (Figure B.16a). The average difference ranged between 0 and 5% for most variables, except for some of the variables where larger differences have been observed, indicating stronger cohort effects for these questions/biomarkers. The exact proportions and the trend over the ELSA waves are plotted in Figures B.12 to B.15. We observed consistent patterns for groups of variables measuring similar traits, however, cohort effects were not present and of similar intensity for all variables and latent traits suggesting that adjusting for cohort effects into the structural model could possibly further complicate and bias the estimates out-weighting the possible benefits of controlling for them into the model. We further obtained the EAP individual estimates and plotted the average trajectory grouped by cohort (Figure B.16a). Similar to exploratory results, cohort effects were present for SW but weaker for PA, PW and CF. A possible solution could be to

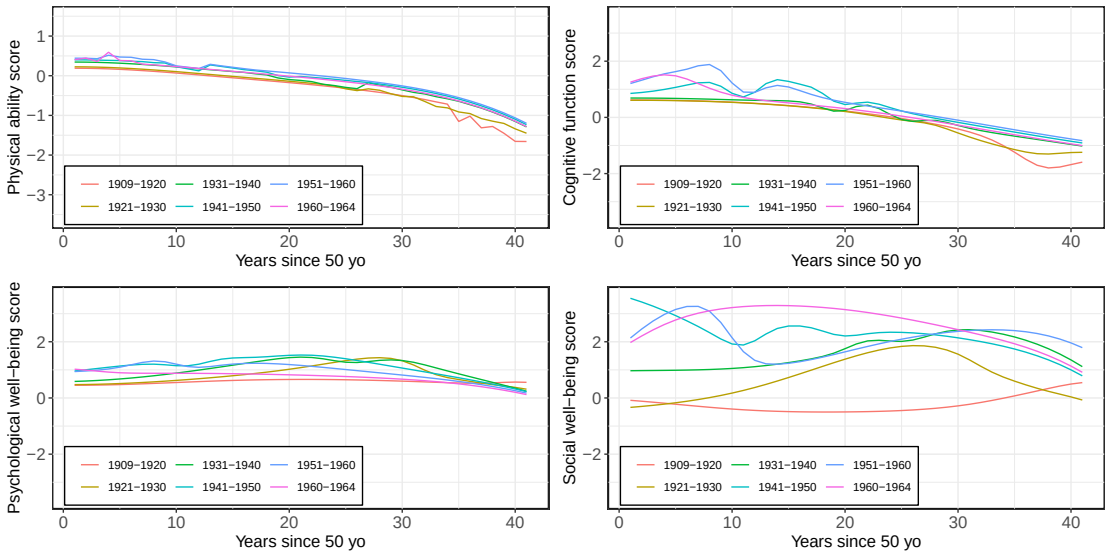


Figure 6.14: Physical ability (PA), cognitive function (CF), psychological well-being (PW) and social well-being (SW) average expected a posteriori (EAP) individual estimates grouped by cohort.

include such effects into the measurement model allowing for the item parameters to vary according to the cohort. However, this could further complicate the model and we leave this for further work.

6.6 Discussion

The study of the ageing process is a rather complex task. Concept complexity and multidimensionallity, as well as, the nature of the collected data, create multiple statistical challenges. Examples include: a large number of outcome variables since there is not a single metric to measure ageing; mixed types of variables; measurement errors; missing data; non-linear underlying trends of unknown functional form. In this chapter, we built upon the work presented across previous chapters and develop a BiLGP framework for modelling longitudinal survey data. In particular, we studied the progression of four major ageing domains (physical ability, cognitive function, social and psychological well-being) and the interrelationships between them after adjusting for common covariates, such as gender, social class, educational level, marital status and wealth. We analysed a large collection of ordinal data from the ELSA and determined path diagrams following previous analysis results (Chapter 3). We assessed model performance and estimation robustness under multiple data scenarios and conditions. Our implementation of the ULGP and BiLGP models revealed interesting findings for the progression of the four

ageing domains with some of the results confirmed in previous relevant literature. We also quantified relationships between some of the ageing domains which could be investigated further in future studies to examine and develop joint strategies towards supporting old people. Cohort effects were also explored but they were insignificant for all domains.

The work within this chapter relies on the following assumptions. First, we considered a higher level latent structure and ignored the variable subgroups shown in Figures 3.11 and 3.12 within each ageing domain. We did that to simplify the already complex BiLGP model structure. Incorporating the two-level structure to the model [for instance in Pattinson (2018)] could improve model interpretability and provide further insights on ageing domains and subdomains and also provide a methodological framework to analyse complex processes from other scientific fields. However, further research is needed to develop a model of this complexity. Second, we assumed that drop-outs were results of a Missing at Random (MAR) mechanism. This ensures that the missingness is ignorable, i.e., the full joint likelihood for the observed data and missingness indicators factorizes into two terms, only one of which depends on the parameters generating the underlying responses. Still, selection bias could potentially introduce some error in our estimates as it has been shown by Banack et al. (2019) regardless of the large sample size. Adjusting for longitudinal weights could possibly improve results and reduce bias but proper investigation of whether and how selection and survival bias affect inference will be investigated in future studies. The End of Life data available in ELSA, even though limited, could potentially provide some additional insights on this. Third, we presented results for a number of interior knots selection, however further research is needed to automatically select the ideal number of knots in an efficient way since to the best of our knowledge such an algortithm has not yet being developed and the IC measure could get some clear improvement. Finally, we assumed a bivariate Matérn covariance for the LGP since its construction and nice properties facilitated model fit. Nevertheless, research on multi-output Gaussian processes offers multiple covariance structure alternatives which could also facilitate extension to higher dimensions (Apanasovich and Genton, 2010; Vandenberg-Rodes and Shahbaba, 2015; Gonçalves and Gamerman, 2018). This project remains a work in progress and investigation on how the proposed framework could be extended to a multivariate framework which is computationally feasible and efficient remains one of our main future objectives. Reparametrisation of the cross-correlation

matrix in order to estimate the multivariate Matérn covariance structure is one of the key tasks that are essential for the extension of the model. We propose the use of Cholesky decomposition; for readers' reference we included the details in Supplementary Text A.5 of the Appendix. Nevertheless, initial experiments revealed convergence issues when this method is extended to more than two dimensions and further investigation is needed to conclude how the model should be adjusted to overcome this issue.

Regardless of its limitations, this work has major strengths. For the first time in the literature, we simultaneously analysed multiple ageing domains and model domain trajectories over time without making any assumptions for the functional form of the underlying trends. At the same time, we managed to model the overall relationships between the examined ageing domains. Possible extensions could look at modelling this dependence over time to explore how relationships between the examined domains evolve over time. This research could also bring major benefits since it provides a clear and flexible framework which allows an integrated view of the ageing process as an evolving mulidimensional construct. This promotes ageing understanding and facilitates adopting targeted and integrated strategies towards supporting old individuals. In addition, the current model creates new potentials in analysing longitudinal survey data by allowing modelling of multidimensional concepts measured using domain specific questionnaires. Through our simulation studies, we extensively explored the method's performance for irreguarly collected longitudinal data and provided an overview of model performance and potential utility under multiple data frameworks. We did so to assess its strengths and promote its utilisation from other scientific fields with similar research objectives.

6.7 Summary

In this chapter, we present a bivariate framework for analysing multivariate longitudinal survey data of ordinal nature. The proposed framework was implemented on data from the ELSA with the specific variabes in use being presented in detail in Chapter 3. Our findings provide an integrated view of the ageing process domains over time and their interrelationships which could support decision making to promote growing older whilst maintaining health and well-being. In the following chapter, we conclude this thesis with some additional concluding remarks and propose future directions.

Chapter 7

Concluding Remarks and Future Directions

The contributions of this thesis focus on developing flexible methodologies to handle complex longitudinal data and allow an in depth understanding of anti-cancer drug response and ageing using real-world evidence. In our first study, we explore dose-varying associations between high-dimensional genetic factors and drug response; here, the longitudinal responses are continuous and regularly collected across the dose points and complexity comes from modelling the non-stationary and dose-varying effect of high-dimensional covariates on drug response. The conclusions drawn provide a unique view of the association between genetic factors and drug dose response whilst offering high accuracy of drug response prediction. In our second study, we explore frailty progression and its distributional changes over time using an innovative quantile regression method; here, longitudinal responses are, again, continuous but irregularly collected. In fact, individual scores are obtained through aggregating multivariate questionnaire responses and complexity arises due to the large variations between individual frailty trajectories, substantial amounts of missing data and measurement error. In our third study, we explore four ageing domains and their interrelationships as individuals are getting older. The data handled are categorical, multivariate longitudinal and collected using multiple survey questionnaires. These data are also irregularly sampled across age expanding the already high problem complexity. The proposed methods of these last two studies really push gerontology forward enabling a more comprehensive view of the ageing process and the factors promoting positive health outcomes. Overall, we showed that modelling

longitudinal data becomes very complex as the data availability and databases grow. In practice, this can have large implications on the conclusions drawn for the populations of interest and the time-varying concepts examined. Motivated by our study findings, below, we highlight some of the key limitations of these studies and directions for possible future research.

In Chapter 4, we propose a model that reveals which genetic factors are associated to anti-cancer drug response and how their expression in the cell genome affects it over multiple doses. Specifically, the proposed algorithm focuses on feature screening of dose-invariant covariates under a dose-varying coefficient modelling framework. Previous studies have found that investigating the change in gene expression profiles may be informative for some biological processes taking place at cell level including cancer evolution and response to treatment (Boehm et al., 2021; Somel et al., 2020). Extending the current methodological framework to handle and screen ultra-high dimensional time-varying covariates could possibly offer important insights in several medicinal areas.

In Chapter 5, we dynamically model irregularly collected longitudinal data. The employed CQ model offers a unique opportunity to estimate frailty distribution dynamics in the presence of missing data with results interpretation offering an opportunity to better understand the ageing process and explore its dynamic change. On top of that, predicting individual outcomes in the presence of missing data may support decision making and the development of aid strategies towards vulnerable population subgroups. However, similar to Chapter 4, our method is limited in handling time-invariant covariates. Incoprporating time-varying information such as financial and social circumstances' change over time or emotional changes may provide further insights into the dynamics of frailty. Of course, this would create further computational challenges to be addressed.

In Chapter 6, our contributions are multifold. To the best of our knowledge, the bivariate extension of the latent Gaussian process framework is developed for the first time to analyse survey ageing data. We, also, offer the opportunity to obtain new insights on ageing progression by modelling not only the mean trajectories of four ageing domains over time but also their pairwise associations. Developing a model which can handle data from all domains jointly while decreasing the computation intensity of the employed algorithms for inference is a natural next step to our study. At the same time, exploring the time-varying interrelationships between the latent domains of interest

through adopting a time-varying expression for the correlation structure between the latent construct of the bivariate Gaussian process model could be another direction of scientific interest.

In the studies of Chapter 5 and 6, a major assumption is that the missing data are MAR. However, we recognise that infering model parameters under this assumption can often introduce some bias, especially for ageing studies where participants might die due to old age or drop-out due to hospitalisation and serious illness. Quantifying the amount of bias introduced and extending our methodology to deal with these non-ignorable/non-informative missing data could be another direction for future research with many applications in multiple scientific fields. In previous years, there have been numerous efforts to address this type of problem through adopting joint modelling approaches (Rizopoulos, 2012; Davis-Plourde et al., 2022; He and Luo, 2016; Van den Hout and Muniz-Terrera, 2016; Li et al., 2012). Combining the strength of the proposed non-parametric methodologies and joint modelling could be very fruitful and create a tool to analyse many types of complex longitudinal data, as, for example, in You and Qiu (2021) where authors developed a local polynomial mixed-effects model and analysed multiple blood pressure longitudinal outcomes.

Finally, in alignment with the project in Chapter 4, suggestions that genetic variants could explain healthy ageing heterogeneity in the population have been presented across the precedent literature. For instance, common Single Nucleotide Polymorphisms (SNPs) have been found collectively to explain a decent percentage of the phenotypic variance of a social deprivation measure of socio-economic status (Marioni et al., 2014). Also, Mekli et al. (2015) have examined the genetic associations between frailty in old age and biomarkers, finding four genes that might be related with frailty in old age. It is possible that genetic factors can explain the observed phenotypic variance of the ageing domain trajectories in the elderly population and, to the best of our knowledge, extensive research has not yet been conducted in this area. High-dimensional feature screening algorithms could be employed for this purpose, including feature subset selection techniques or feature extraction techniques to reduce their dimensionality and detect potential biomarkers of ageing. In fact, the ELSA data are accompanied by Genome-Wide Association Study data which could be used for this purpose.

Appendix A

Supplementary text

A.1 The ELSA questions and variable coding

Biomarker and chronic disease data

Binary indicators were used for cardiovascular and chronic disease variables as well as for medication intake data. Reverse coding was used since disease occurrence of medication intake meant a worse health score. For biomarker and body composition data, ordinal indicators were used as determined using the following cut-off points based on previous literature Sanders et al. (2014), as follows:

- for blood glucose level (mmol/L) – fasting samples only: 6.1 and 7.0 mmol/L;
- for the mean systolic pressure: 130 and 140;
- for the mean diastolic pressure: 85 and 90;
- for the Body Mass Index (kg/m^2): better score for a BMI between 18.5 (including 18.5) and 25, worse score for a BMI less than 18.5 or between 25 (including 25) and 30−[under/over]weight−, and worst score for a BMI greater or equal to 30−obese−.

For C-reactive protein levels (mg/L), blood triglycerides levels (mmol/L), blood HDL levels (mmol/L), blood fibrinogen levels (g/L) and blood glycated haemoglobin levels (%), we used tertiles to create the ordinal indicators of metabolic malfunctioning.

Missing data rates were large in biomarker and NV data. Figure A.1 shows the amount of missing values included in a random selection of the NV data. Readers can observe the large amount of missingness in the biomarker data compare to the rest variables in the data set.

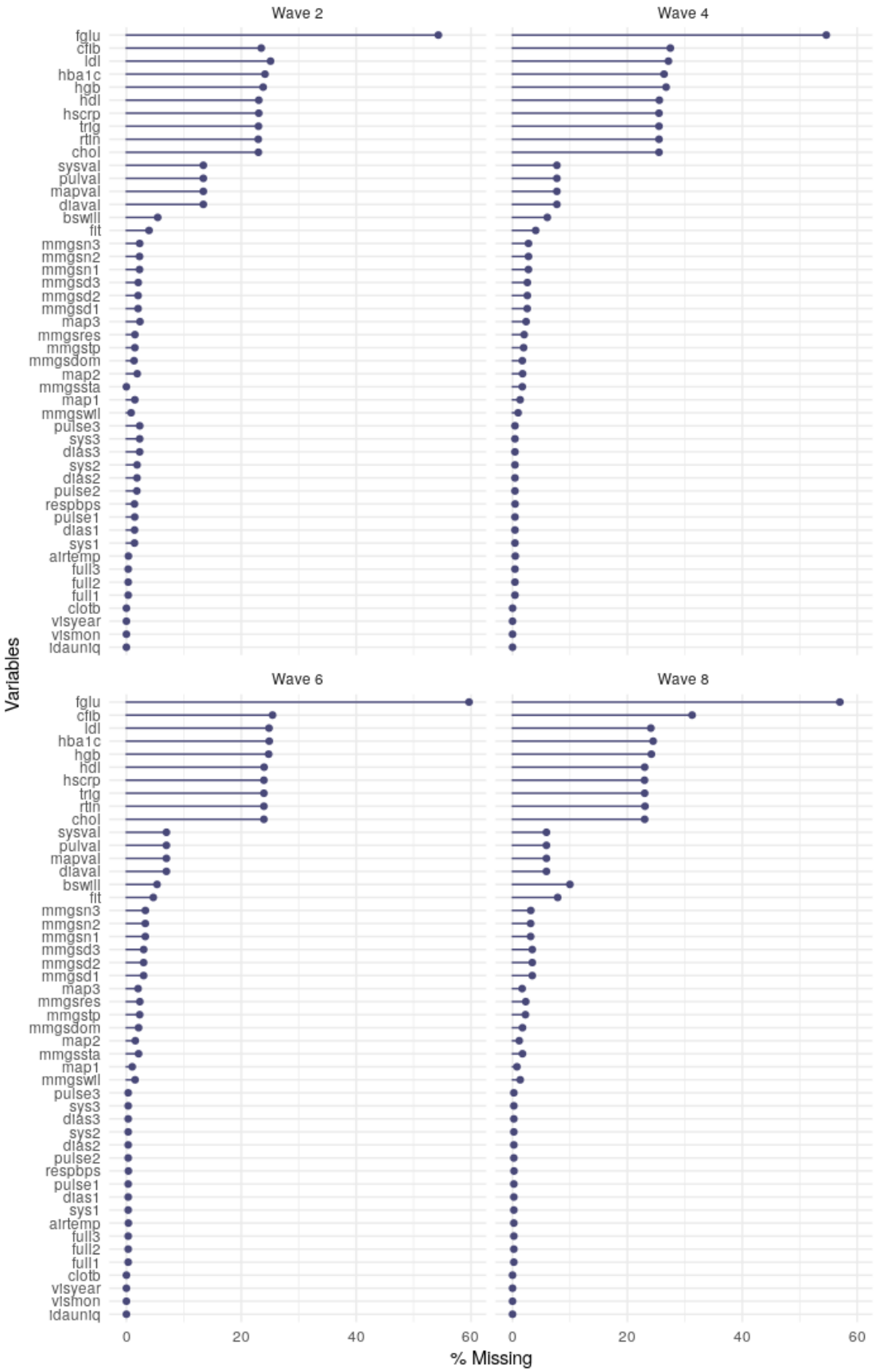


Figure A.1: Cross-sectional missing rates in the ELSA NV data. The first fourteen variables are some of the available biomarker data, the rest are other data and test results collected during the NVs.

Physical capability data

ADLs and IADLs questionnaires include difficulties with:

- Dressing, including putting on shoes and socks
- Walking across a room
- Bathing or showering
- Eating, such as cutting up [your/his/her] food
- Getting in or out of bed
- Using the toilet, including getting up or down
- Using a map to figure out how to get around in a strange place
- Recognising when you are in physical danger
- Preparing a hot meal
- Shopping for groceries
- Making telephone calls
- Communication (speech, hearing or eyesight)
- Taking medications
- Doing work around the house or garden
- Managing money, such as paying bills and keeping track of expenses

Mobility questionnaire includes difficulties with:

- Walking 100 yards
- Sitting for about two hours
- Getting up from a chair after sitting for long periods
- Climbing several flights of stairs without resting
- Climbing one flight of stairs without resting
- Stooping, kneeling, or crouching
- Reaching or extending [your / his / her] arms above shoulder level (either arm)
- Pulling or pushing large objects like a living room chair
- Lifting or carrying weights over 10 pounds, like a heavy bag of groceries
- Picking up a 5p coin from a table

Response and coding used: All responses to ADLs, IADLs and Mobility questions are binary: yes, if difficulty is present; and, no, if difficulty is absent. Since a positive answer indicates worse health, we coded 0 if "yes" and 1 if "no". From wave 4 onward two additional difficulties were added to the questionnaire: difficulty in recognising when in danger and difficulty in communicating (speech, hearing or eyesight).

Cognitive function data

For orientation in time questions score, we scaled responses between 0 and 2 to maximise the variability observed (0 for a score less than of equal to 2, 1 for a score equal to 3 and 2 for those who scored 4). For the rest variables tertiles were used.

Social well-being data

Social network questions: Participants were asked whether they have family, friends, children or partner; and, if yes :

1. How much do you feel they understand the way you feel about things?
2. How much can you rely on them if you have a serious problem?
3. How much can you open up to them if you need to talk about your worries?
4. How much do they criticise you?
5. How much do they let you down when you are counting on them?
6. How much do they get on your nerves?
7. On average, how often do they make demands on you?
8. On average, how often do you meet up with them?
9. On average, how often do you speak on the phone to them?
10. On average, how often do you write or e-mail them?
11. How often do you send or receive text messages from them?
12. How many of them would you say you have a close relationship?

Response and coding used: For questions 1-7 answers range from "Not at all" (1) to "A lot" (4) and reverse coding has been applied for those with negative meaning, i.e. questions 4-7. For questions 8-11 answers range from "Less than once a year or never" (1) to "Three or more times a week" (6). Finally, for the last question, responses range from "Not at all close" (1) to "Very close" (4). Since questions 8-11 referred to contact frequency with friends/family/children, we summarised their content into one variable by using their weighted sum. Higher weight has been applied for meeting up most frequently and lower for texting.

Civil and social participation questions:

1. Organisational membership: political party, trade union or environmental group.
2. Organisational membership: tenants or resident group or neighbourhood watch.

3. Organisational membership: member of a church or other religious group.
4. Organisational membership: member of a charitable association.
5. Organisational membership: an education, arts or music group or evening class.
6. Organisational membership: member of a social club.
7. Organisational membership: not a member of any organisation, club or society.
8. Organisational membership: member of a sports clubs, gym, or exercise class.
9. Organisational membership: member of any other organisations, clubs or societies.

and

10. Does respondent read a daily newspaper?
11. Does respondent have a hobby or pastime?
12. Does respondent have taken a holiday in the UK in the last 12 months?
13. Does respondent have taken a holiday abroad in the last 12 months?
14. Does respondent have gone on a daytrip or outing in the last 12 months?
15. Does respondent use the internet and/or email?
16. Does respondent own a mobile phone?

Response and coding used: Answers to questions 1-16 were either "yes"−for being a member of such organisation or doing as the statement suggests, coded as 1−or "no"−for not being a member of such organisation or not doing as the statement suggests, coded as 0−. As for questions 1-9, membership to different organisations has being summarised into one new variable which indicating whether participant was a member to any organisation by summing the number of organisations a participant was a member. A missing value has being assigned only if an individual has not answered any of the independent questions referring to membership to organisations. Subsequently, we used tertiles as cut-off points to distinguish between socially active and socially inactive participants in terms of organisations participation.

Cultural participation questions:

1. How often respondent goes to the cinema?
2. How often respondent eats out of the house?
3. How often respondent goes to an art gallery or museum?
4. How often respondent goes to the theatre, a concert or the opera?
5. Would respondent like to go to the cinema more often?

6. Would respondent like to eat out of the house more often?
7. Would respondent like to go to art galleries or museums more often?
8. Would respondent like to go to the theatre, concerts or the opera more often?

Response and coding used: For questions 1-4 possible responses range from "Never" (1) to "Twice a month or more" (6) whereas for questions 5-8 responses were dichotomous ("yes"−1 or "no"−0). Questions 5-8 have not included into the study though due to high missingness rates (more than 25% in most waves).

UCLA Loneliness Scale questions:

1. How often respondent feels they lack companionship?
2. How often respondent feels left out?
3. How often respondent feels isolated from others?
4. How often respondent feels in tune with the people around them?
5. How often respondent feels lonely?

Response and coding used: Answers to the UCLA Loneliness Scale questionnaire answers vary from "Often" (1) to "Some of the time" (2) and "Hardly ever or never" (3). Reverse coding has been used for item 4.

Psychological well-being data

Satisfaction with Life Scale, response and coding used: Satisfaction with Life Scale consists of five positive statements about individual satisfaction with life and nswers vary from 1 ("Strongly disagree") to 7 ("Strongly agree") with the mid-point being 4 ("Neither agree or disagree")−7-point Likert Scale. The statements presented were:

1. In most ways their life is close to ideal.
2. The conditions of their life are excellent.
3. The respondent is satisfied with their life.
4. So far, the respondent has got the important things they want in life.
5. If the respondent could live their life again, they would change almost nothing.

CASP-19, response and coding used: CASP-19 questionnaire included questions on enjoyment of life and positive affective well-being, with the main measured outcome being the quality of life. In particular, the 19-item questionnaire addresses four main domains: control, autonomy, self-realisation and pleasure. Possible answers were "Often", "Some

of the time" and "Hardly ever or never" and scores vary from 1 ("Never") to 4 ("Often"). Note that questions 1,2,4,5,8,9 have negative meaning, hence reverse coding has been used, i.e. 1 for "Often" to 4 for "Never". The specific questions were:

1. How often respondent feels age prevents [him/her] from doing things [he/she] like? (Control)
2. How often respondent feels what happens to [him/her] is out of [his/her] control? (Control)
3. How often respondent feels free to plan for the future? (Control)
4. How often respondent feels left out of things? (Control)
5. How often family responsibilities prevents respondent from doing things? (autonomy)
6. How often respondent can do the things [he/she] want to do? (autonomy)
7. How often respondent feels [he/she] can please [him/her]self what [he/she] does? (autonomy)
8. How often respondent feels [his/her] health stops [him/her] doing what [he/she] wants to do? (autonomy)
9. How often shortage of money stops [him/her] doing things? (autonomy)
10. How often respondent looks forward to each day? (pleasure)
11. How often respondent feels that [his/her] life has meaning? (pleasure)
12. How often respondent enjoys the things [he/she] do? (pleasure)
13. How often respondent enjoys being in the company of others? (pleasure)
14. How often respondent looks back on [his/her] life with a sense of happiness? (pleasure)
15. How often respondent feels full of energy these days? (self-realisation)
16. How often respondent chooses to do things [he/she] has never done before? (self-realisation)
17. How often respondent feels satisfied with the way [his/her] life has turned out? (self-realisation)
18. How often respondent feels that life is full of opportunities? (self-realisation)
19. How often respondent feels the future looks good to [him/her]? (self-realisation)

Negative affective well-being questions: Negative affective well-being has been assessed using the Center for Epidemiological Studies Depression (CES-D) Scale. The questionnaire

includes questions about how participants feel the previous to the interview week. Possible answers were either "yes" (0) or "no" (1) and the questions asked are presented below:

1. Has respondent felt depressed much of the time during the past week?
2. Has respondent felt everything they did during the past week was an effort?
3. Has respondent felt their sleep was restless during the past week?
4. Was respondent happy much of the time during the past week?
5. Has respondent felt lonely much of the time during the past week?
6. Has respondent enjoyed life much of the time during the past week?
7. Has respondent felt sad much of the time during the past week?
8. Could respondent not get going much of the time during the past week?

Response and coding used: Reverse coding has been used for questions 4 and 6, i.e. 1 for "yes" and 0 for "no".

Self-perceived social status

Subjective social status refers to self-perceptions of one's own social position. It is measured by respondents indicating on the rung of a ladder where they stand in society based on money, education and employment. The question asked was: "Think of this ladder as representing where people stand in our society. At the top of the ladder are the people who are the best off-those who have the most money, most education and best jobs. At the bottom are the people who are the worst off-who have the least money, least education, and the worst jobs or no jobs. The higher up you are on this ladder, the closer you are to the people at the very top and the lower you are, the closer you are to the people at the very bottom. Please mark a cross on the rung on the ladder where you would place yourself.". Respondents who had put their mark in between two rungs were assigned to the higher of these rungs. Scores varied from 10 to 100. We re-coded this variable into a three category variable (Lower, Middle and Upper class) based on variable tertiles.

Self-perceived relative deprivation

To get a measure of relative deprivation, scores were assigned to individuals based on the number of items reported in the following question: "Does having too little money stop

you from doing any of the following things [. . .]?". The possible responses were yes or no. The interviewer, then, coded all that apply from the following list of activities:

1. buy your first choices of food items;
2. have family and friends round for a drink or meal;
3. have an outfit to wear for social or family occasions;
4. keep your home in a reasonable state of decoration;
5. replace or repair broken electrical goods pay for fares or other transport costs to get to and from places you want to go;
6. buy presents for friends or family once a year;
7. take the sorts of holidays you want;
8. treat yourself from time to time; or,
9. none of these.

Since only a few individuals reported any items in this list, we recoded the variable into a binary variable with 0 denoting experiencing none of the difficulties in the list and 1 denoting that the individual reported at least one of the items in the list.

A.2 Variables used specifically for the construction of the FI

Below are the variables names, content or questions asked and cut-off points used for the FI calculation using the ELSA data waves 1 to 8. These variables were chosen based on the following criteria: 1. deficits should be associated with health; 2. deficits' prevalence should generally increase with age (except for deficits that show decreasing prevalence due to decreased survival); 3. they should not saturate too early, meaning that if, for example most 60-years-old people possess a particular deficit then this would be inadequate to be included into the construct; 4. deficits should cover a range of systems, and; 5. if the index is to be used for longitudinal data including the same people, the items that make up the FI should be the same from one wave to the next.

INFORMATION	VAR. NAME	CUT-OFF POINTS
Variables measuring activities of daily living, instrumental activities of daily living and mobility.		
Difficulty in walking 100 yards	hemobwa	Yes = 1, No = 0
Difficulty in sitting for about 2 hours	hemobsi	Yes = 1, No = 0
Difficulty in getting up from a chair after sitting for long periods	hemobch	Yes = 1, No = 0
Difficulty in climbing several flights of stairs without resting	hemobcs	Yes = 1, No = 0
Difficulty in climbing one flight of stairs without resting	hemobcl	Yes = 1, No = 0
Difficulty in stooping, kneeling, or crouching	hemobst	Yes = 1, No = 0
Difficulty in reaching or extending your arms above shoulder level	hemobre	Yes = 1, No = 0
Difficulty in pulling or pushing large objects like a living room chair	hemobpu	Yes = 1, No = 0
Difficulty in lifting or carrying weights over 10 pounds, like a heavy bag	hemobli	Yes = 1, No = 0
Difficulty in picking up a 5p coin from a table	hemobpi	Yes = 1, No = 0
Difficulty in communication (speech, hearing or eyesight)	headlco	Yes = 1, No = 0
Difficulty in dressing, including putting on shoes and socks	healddr	Yes = 1, No = 0
Difficulty in walking across a room	headlwa	Yes = 1, No = 0
Difficulty in bathing or showering	headlba	Yes = 1, No = 0
Difficulty in eating, such as cutting up your food	headlea	Yes = 1, No = 0
Difficulty in getting in or out of bed	headlbe	Yes = 1, No = 0
Difficulty in using the toilet, including getting up or down	headlwc	Yes = 1, No = 0
Difficulty in using a map to figure out how to get around in a strange place	headlma	Yes = 1, No = 0
Difficulty in preparing a hot meal	headlpr	Yes = 1, No = 0
Difficulty in shopping for groceries	headlsh	Yes = 1, No = 0
Difficulty in making telephone calls	headlph	Yes = 1, No = 0
Difficulty in recognising when in physical danger	headlda	Yes = 1, No = 0
Difficulty in taking medications	headlme	Yes = 1, No = 0
Difficulty in doing work around the house or garden	headlho	Yes = 1, No = 0
Difficulty in managing money, such as paying bills and keeping track of expenses	headlmo	Yes = 1, No = 0
Variables measuring cardiovascular disease occurrence		
Cardiovascular disease: High-blood Pressure or Hypertension	hediabp	Yes = 1, No = 0
Cardiovascular disease: Heart attack (including myocardial infraction or coronary thrombosis)	hediami	Yes = 1, No = 0
Cardiovascular disease: Stroke (cerebral vascular disease)	hediast	Yes = 1, No = 0
Cardiovascular disease: Angina	hediaan	Yes = 1, No = 0
Cardiovascular disease: A heart murmur	hediahm	Yes = 1, No = 0
Cardiavascular disease: Congestive Heart Failure	hediahf	Yes = 1, No = 0
Cardiovascular disease: Diabetes or high blood sugar	hediadi	Yes = 1, No = 0
Cardiovascular disease: An Abnormal Heart Rhythm	hediaar	Yes = 1, No = 0
Variables measuring chronic disease occurrence		
Chronic Lung disease such as chronic bronchitis or emphysema	hediblu	Yes = 1, No = 0
Asthma	hedibas	Yes = 1, No = 0
Arthritis (including osteoarthritis, or rheumatism)	hedibar	Yes = 1, No = 0
Osteoporosis, sometimes calles thin or little bones	hedibos	Yes = 1, No = 0
Cancer or a malignant tumour (excluding minor skin cancers)	hedibca	Yes = 1, No = 0
Parkinson's disease	hedibpd	Yes = 1, No = 0
Any emotional, nervous or psychiatric problems	hedibps	Yes = 1, No = 0
Dementia, organic brain syndrome, senility or other serious memory impairment	hedibde	Yes = 1, No = 0
Alzheimer's disease	hedibad	Yes = 1, No = 0
Variables measuring self-reported health and depressive symptoms		
Self-reported health: HRS/HSE version	srh_hse/ srh_hrs	Excellent/Very Good = 1, Good = 0.5, Fair/Poor/Bad/Very Bad = 0
Much of the time during the past week, have you felt depressed?	psceda	Yes = 1, No = 0
Much of the time during the past week, have you felt that everything you did was an effort?	pscedb	Yes = 1, No = 0
Much of the time during the past week, has your sleep been restless?	pscedc	Yes = 1, No = 0
Much of the time during the past week, were you happy?	pscedd	Yes = 0, No = 1
Much of the time during the past week, have you felt lonely?	pscede	Yes = 1, No = 0
Much of the time during the past week, have you enjoyed life?	pscedf	Yes = 0, No = 1
Much of the time during the past week, have you felt sad?	pscedg	Yes = 1, No = 0
Much of the time during the past week, could you not get going?	pscedh	Yes = 1, No = 0
Variables measuring cognitive function		
Please tell me today's date: day	cfdatd	Succeed = 0, Failed = 1
Please tell me today's date: month	cfdatm	Succeed = 0, Failed = 1
Please tell me today's date: year	cfdaty	Succeed = 0, Failed = 1
Please tell me what day of the week it is today.	cfday	Succeed = 0, Failed = 1
Please tell me the words that you can recall: enter number of words recalled immediately.	cflisen	≤lowest quintile = 1, >lowest quintile = 0
A little while ago, you were read a list of words and repeated the ones you could remember. Please tell me any words you can remember now.	cflisd	≤lowest quintile = 1, >lowest quintile = 0

A.3 The stochastic Expectation-Maximisation algorithm for parameter estimation of the BiLGP model

Below we provide a detailed description of the stochastic Expectation-Maximisation algorithm employed for parameter estimation of the BiLGP model presented in Chapter 6. Notice that definitions of parameters and notation used can be found in Chapter 6. As described previously, we iterate between a StE-step, where we sample from the conditional distribution of Θ using Gibbs Sampling, and a M-step where we maximise the joint log-likelihood as defined in Eq. 6.19. In particular, having defined the augmented joint log-likelihood as

$$\ell(\boldsymbol{\Psi}; \boldsymbol{y}, \boldsymbol{y}^*, \boldsymbol{\Theta}) = \sum_{n=1}^{N} \Bigg\{ \sum_{i=1}^{I} \sum_{j=1}^{J_n} \mathbf{1}_{b_{i,y_{ni}(t_{nj})} < y^*_{ni}(t_{nj}) \leq b_{i,y_{ni}(t_{nj})+1}} \times \log \Bigg[N(y^*_{ni}(t_{nj}); -\sum_{d=1}^{D} \alpha_{id} \theta_{dn}(t_{nj})) \Bigg] \Bigg\} + \log \Bigg[MVN_{DJ_n}(\boldsymbol{\theta}_n; \boldsymbol{\mu}_n, \Sigma) \Bigg], \tag{A.1}$$

we let $\Psi^{(0)}$ be the initial parameter values and $(\tilde{\boldsymbol{\theta}}_n^{(0)}(t_{n1}), \ldots, \tilde{\boldsymbol{\theta}}_n^{(0)}(t_{nJ_n}))$, $n = 1, \ldots, N$ be the initial values of individual parameter values. In each step s ($s \geq 1$) of the StEM algorithm, the following steps are repeatedly performed.

StE-step: Given $\Psi^{(s)}$, we implement Gibbs sampling as follows

Step 1. We, first, sample $y^*_{ni}(t_{nj})$ from a truncated Normal distribution defined between $b_{i,y_{ni}(t_{nj})}$ and $b_{i,y_{ni}(t_{nj})+1}$ with mean $-\sum_{d=1}^{D} \alpha_{id} \tilde{\theta}^{(s)}_{dn}(t_{nj})$ and standard deviation equal to 1 for $n = 1, \ldots, N$, $i = 1, \ldots, I$ and $j = 1, \ldots, J_n$. Note, that $\tilde{\theta}^{(s)}_{dn}(t_{nj})$ is some initial value of $\theta_{dn}(t_{nj})$.

Step 2. For $n = 1, \ldots, N$, given the sampled $y^*_{ij}(t_{nj})$, we then sample $\tilde{\boldsymbol{\theta}}_n^{(s+1)}$ from the conditional distribution of $\boldsymbol{\theta}_n | \boldsymbol{y}^*_n(t_{n1}), \boldsymbol{y}^*_n(t_{n2}), \ldots, \boldsymbol{y}^*_n(t_{nJ_n})$. That is, a multivariate Normal with variance-covariance matrix $\Sigma^{-1}_{\text{new}} = (\Sigma^{-1} + \sum_{i=1}^{I} A_i A_i^T)$ and mean equal to $-\Sigma_{\text{new}} \sum_{i=1}^{I} A_i \boldsymbol{y}^*_{ni}$, where A_i is a sparse $DJ_n \times J_n$ matrix with the corresponding item loadings.

M-step: Given the $\tilde{\boldsymbol{\theta}}_n^{(s+1)}$, we can then obtain the parameter estimates

$$\Psi^{(s+1)} = \arg_{\Psi} \max \ell(\boldsymbol{\Psi}; \boldsymbol{y}, \boldsymbol{\Theta})$$

where $\ell(\boldsymbol{\Psi}; \boldsymbol{y}, \boldsymbol{\Theta})$ is the logarithm of 6.19.

As shown in Nielsen (2000), the final estimate of Ψ is given by the average of $\Psi^{(s)}$ across the last M iterations, specifically,

$$\hat{\Psi} = \frac{1}{M} \sum_{s=M_0+1}^{M_0+M} \Psi^{(s)}$$

where M_0 is the pre-specified burn-in.

A.4 The exploration of covariate effects of Chapter 6

We confirmed the additive effect assumption for the covariate effects of the ULGP model by simulating multivariate binary response data assuming a latent unidimensional model with $\mu(t) = 0.8 \sin^3(0.15t)$ and $\xi = 10$ for N = 10,000 individuals. We considered three scenarios for the effect of the covariate on the latent mean process: one where the covariate has a multiplicative effect, i.e. $\mu(t) = 0.8 \sin^3(0.15t) e^{\mathbf{1}_{x=1}}$; one where the covariate has an additive effect, i.e. $\mu(t) = 0.8 \sin^3(0.15t) + \mathbf{1}_{x=1}$; and, one where the covariate had no effect on the latent mean curve. In Figure A.2, we plotted the proportion of the generated responses which were equal to 1 for the two groups over t and the corresponding theoretical probabilities of answering 1 for each covariate group. As expected, if there is no effect the observed proportions should coincide, if there is an additive effect the proportions should be not crossing whereas if there is a multiplicative effect the empirical probability estimates should cross. Comparing this example with the empirical probability estimates we observed for each variable in the data we observed that our assumption holds for some of covariates and domains, thus, findings suggest that assuming an additive effect on the mean trend over time is a reasonable assumption (Figure A.3).

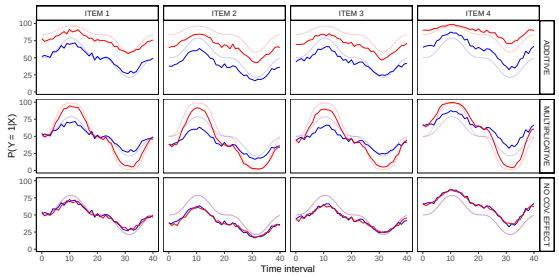


Figure A.2: Proportion of the generated responses which are equal to 1 for two groups (binary covariate) over t and the corresponding theoretica probabilities of answering 1 for each covariate group.

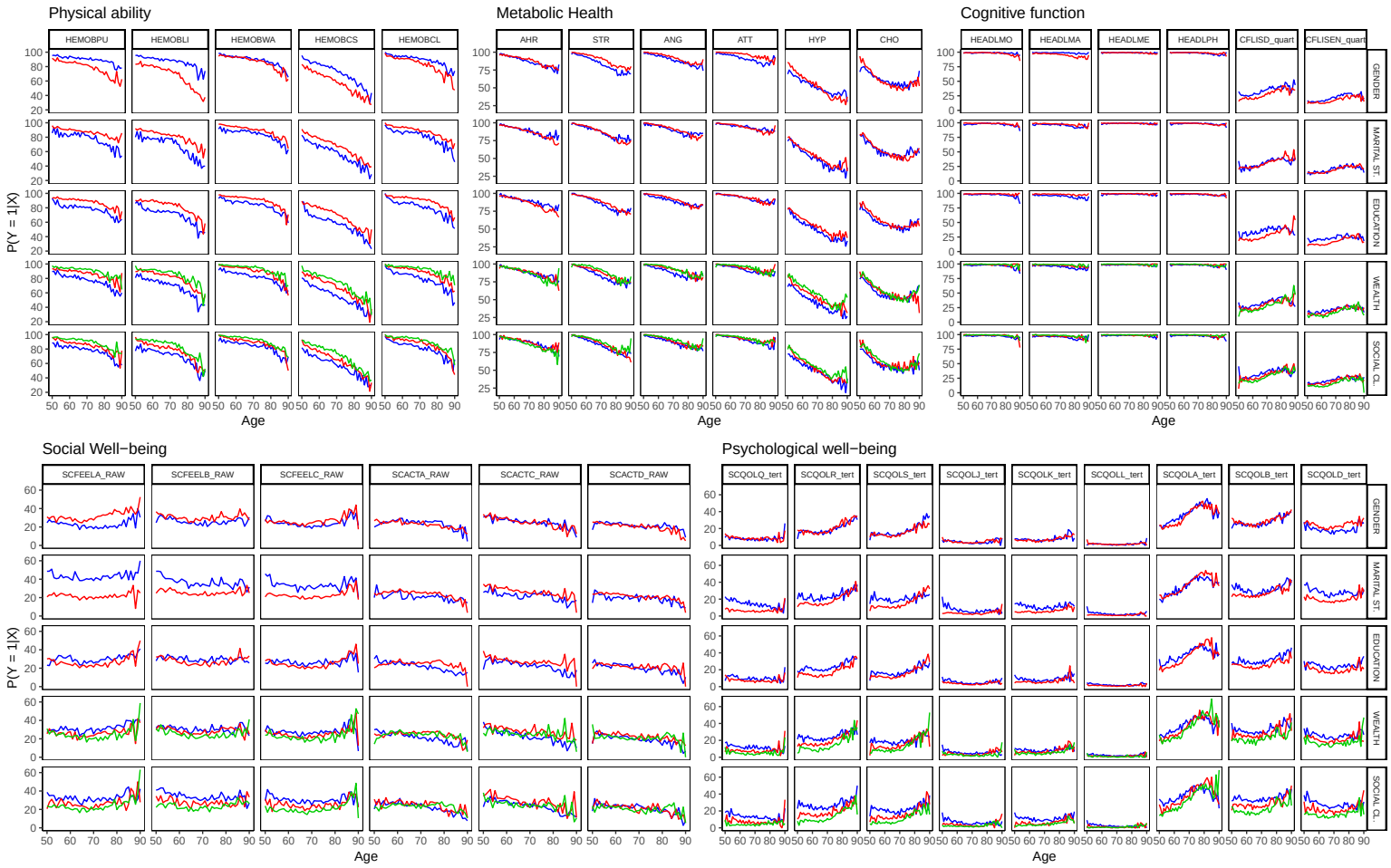


Figure A.3: Observed proportion of responses equal to one of the covariate categories for all covariates and all items in all five ageing domains.

A.5 Estimation of the multivariate Matérn cross-correlation matrix

Let B being a symmetric correlation matrix so that only the parameters in the upper (or lower) triangular part should be estimated. If $D = 2$ then the $B = \begin{bmatrix} 1 & \beta_{12} \\ \cdot & 1 \end{bmatrix}$ matrix is positive definite for all $\beta_{12} \in (-1, 1)$. However, for $D > 2$ additional constraints should be imposed so that $x^T Bx > 0$ for all nonzero real vectors x. To ensure correct estimation of the cross-correlation matrix of the multivariate Matérn covariance, we used a Cholesky-based parametrisation for B to ensure that the estimated $\hat{B}$ will always be valid correlation matrices (van Oest, 2021). The exact method is described below:

Let

$$B = \begin{bmatrix} \beta_{11} & \beta_{12} & \cdots & \beta_{1D} \\ \beta_{21} & \beta_{22} & \cdots & \beta_{2D} \\ \vdots & \vdots & \cdots & \vdots \\ \beta_{D1} & \beta_{D2} & \cdots & \beta_{DD} \end{bmatrix}$$

be a $D \times D$ positive definite correlation matrix, where $\beta_{dq} = \beta_{qd}$ (symmetry), $\beta_{dd} = 1 \; \forall d = 1, \ldots, D$, $\beta_{dq} \in (-1, 1)$ if $d \neq q$ and $d, q \in \{1, 2, \ldots, D\}$. Due to its symmetry and positive definiteness, B has a Cholesky decomposition: $B = L^T L$, where $L = (l_{dq})$ is a unique lower triangular matrix with all diagonal elements taking positive values. This L matrix has $D(D-1)/2$ free elements, in particular:

$$l_{dd} = \sqrt{1 - \sum_{k=1}^{d-1} l_{dq}^2} > 0 \text{ and } \sum_{k=1}^{d-1} l_{dq}^2 < 1, \; d = 2, \ldots, D.$$

Hence, if we let $\Omega = (\omega_{dq})$ be a lower triangular matrix with $\omega_{dd} = 0$ for all $d = 1, \ldots, D$ and $\theta_{dq} \in (0, \pi)$ for all $d = 2, \ldots, D$ and $q = 1, \ldots, d-1$, then we can re-parametrise L so that

$$l_{dq} = \cos(\omega_{dq}) \prod_{k=1}^{q-1} \sin(\omega_{dk})$$

where $d = 1, \ldots, D$ and $q = 1, \ldots, d$ (spherical parametrisation of the Cholesky decomposition).

Appendix B

Supplementary Figures

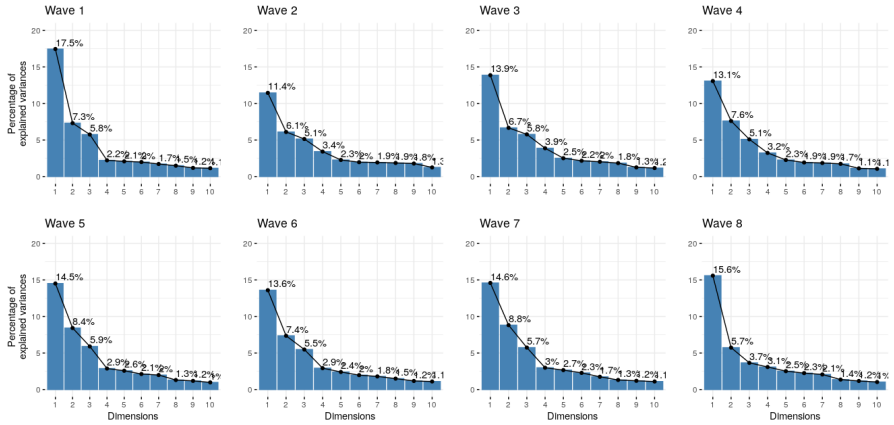


Figure B.1: Scree plot showing the percentage of inertia explained by the first 10 MCA dimensions (full data analysis).

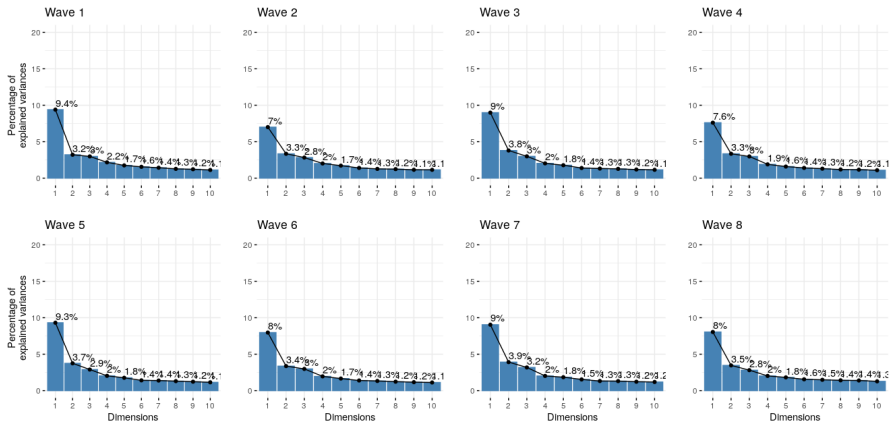


Figure B.2: Scree plot showing the percentage of inertia explained by the first 10 MCA dimensions (complete case analysis).

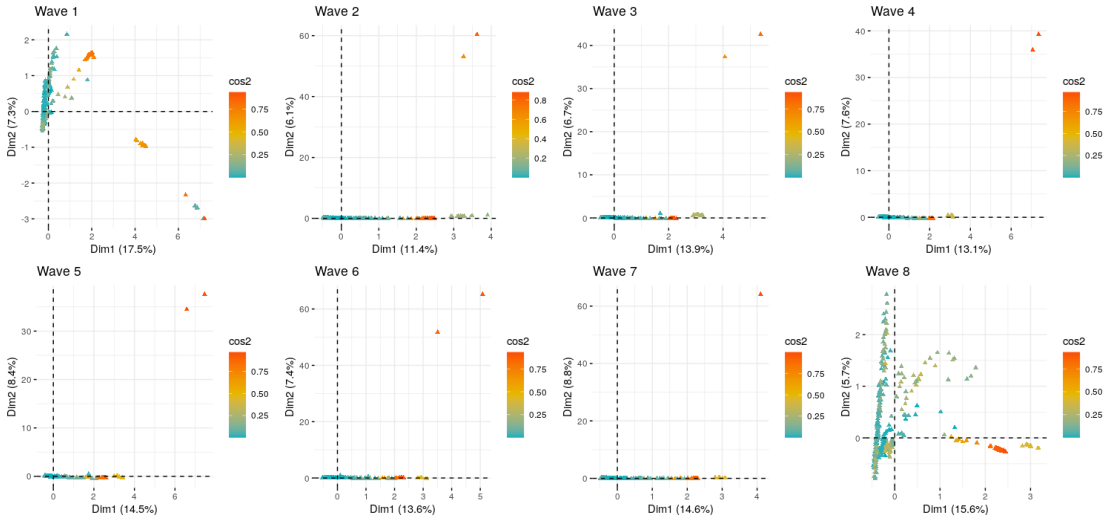


Figure B.3: Perceptual map of categories of variables investigated. Different colours were used to show the quality of each point for the indicated dimensions based on their squared cosine value. Analysis was performed to the full data.

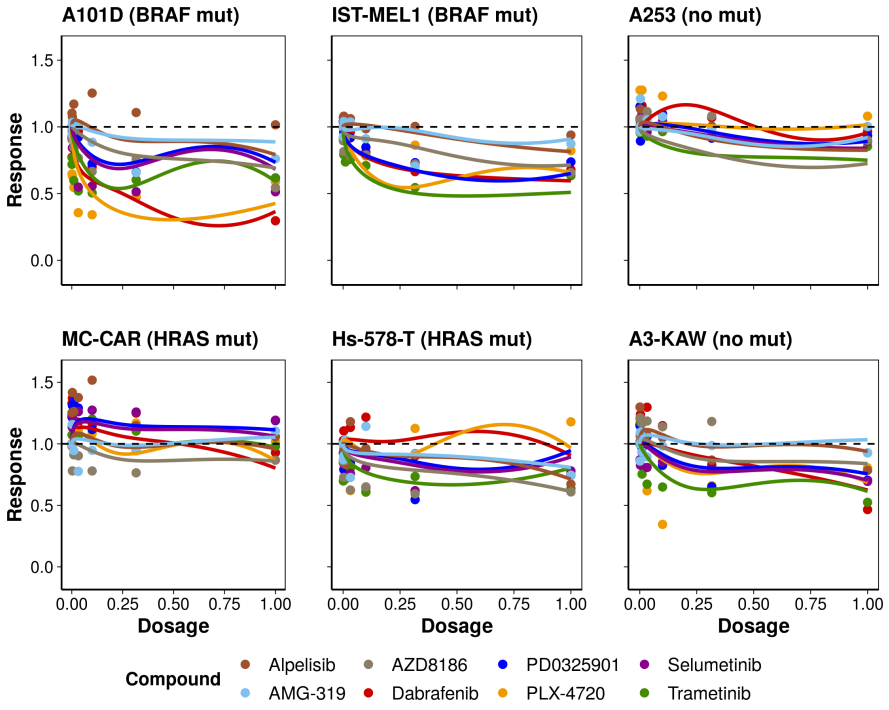


Figure B.4: Estimated mean drug response trajectories for *BRAF* and *HRAS* mutated and non-mutated cancer cell lines: analysis performed on GDSC2 data. Observed responses (points) and estimated mean trajectory (lines) of cell concentration for cancer cell lines with and without *BRAF* and *HRAS* mutations after treatment with the eight anticancer compounds examined using data from GDSC2.

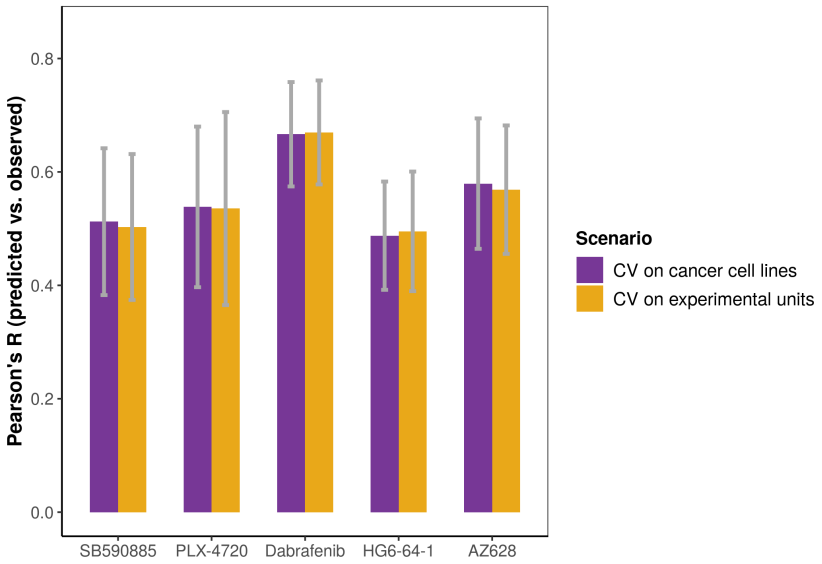


Figure B.5: Prediction accuracy for each different drug and scenario. Pearson correlation was estimated across observed and predicted AUC values. AUC values have been computed by calculating the area under the coefficient function curve (both observed and predicted). Training and test sets have been considered based on either the experimental units or on cancer cell lines only.

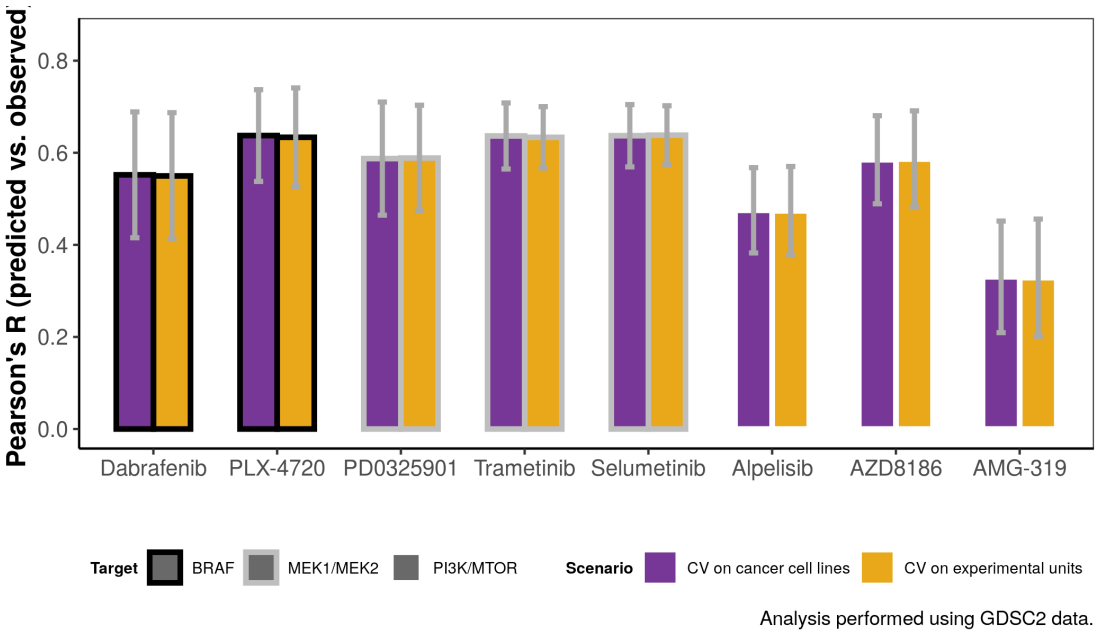


Figure B.6: Prediction accuracy for each different drug and scenario: analysis performed on GDSC2 data. Pearson correlation across observed and predicted AUC values. AUC values have been computed by calculating the area under the coefficient function curve (both observed and predicted) using the GDSC2 data. Training and test sets have been considered based on either the experimental units or on cancer cell lines only.

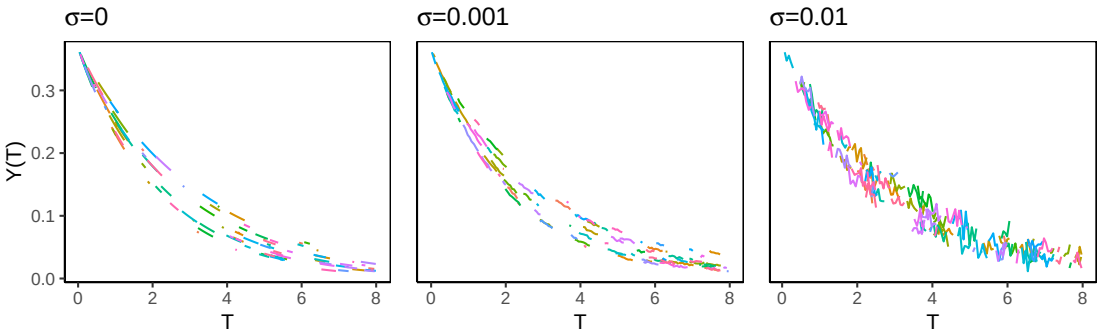


Figure B.7: Simulated trajectories from a 100 subjects under three different noise scenarios and when Δ_n varies across subjects.

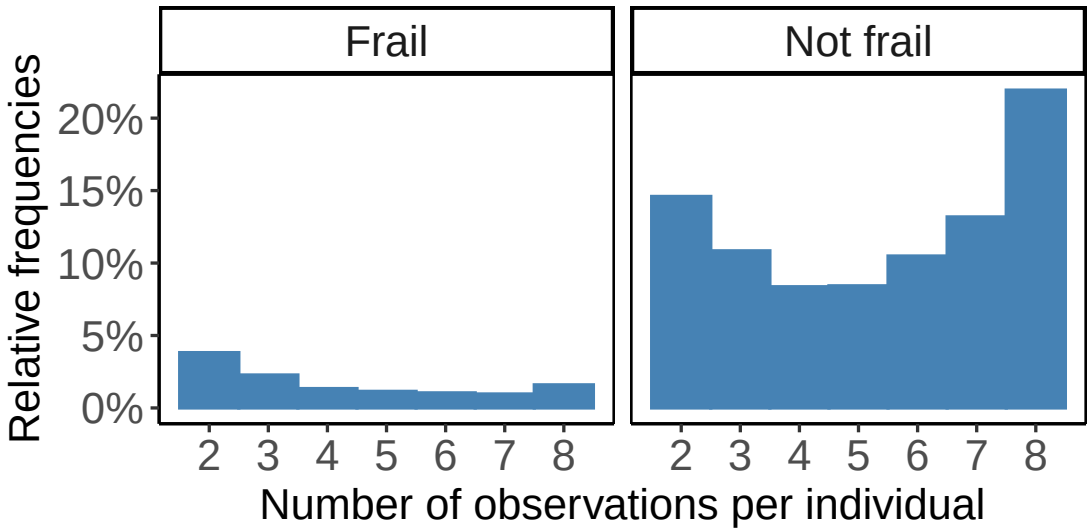


Figure B.8: Relative frequency of the number of longitudinal observations per individual. Individuals with a single observation have been excluded from the analysis. The total number of frail at baseline subjects is 947 compared to 6836 for the non-frail.

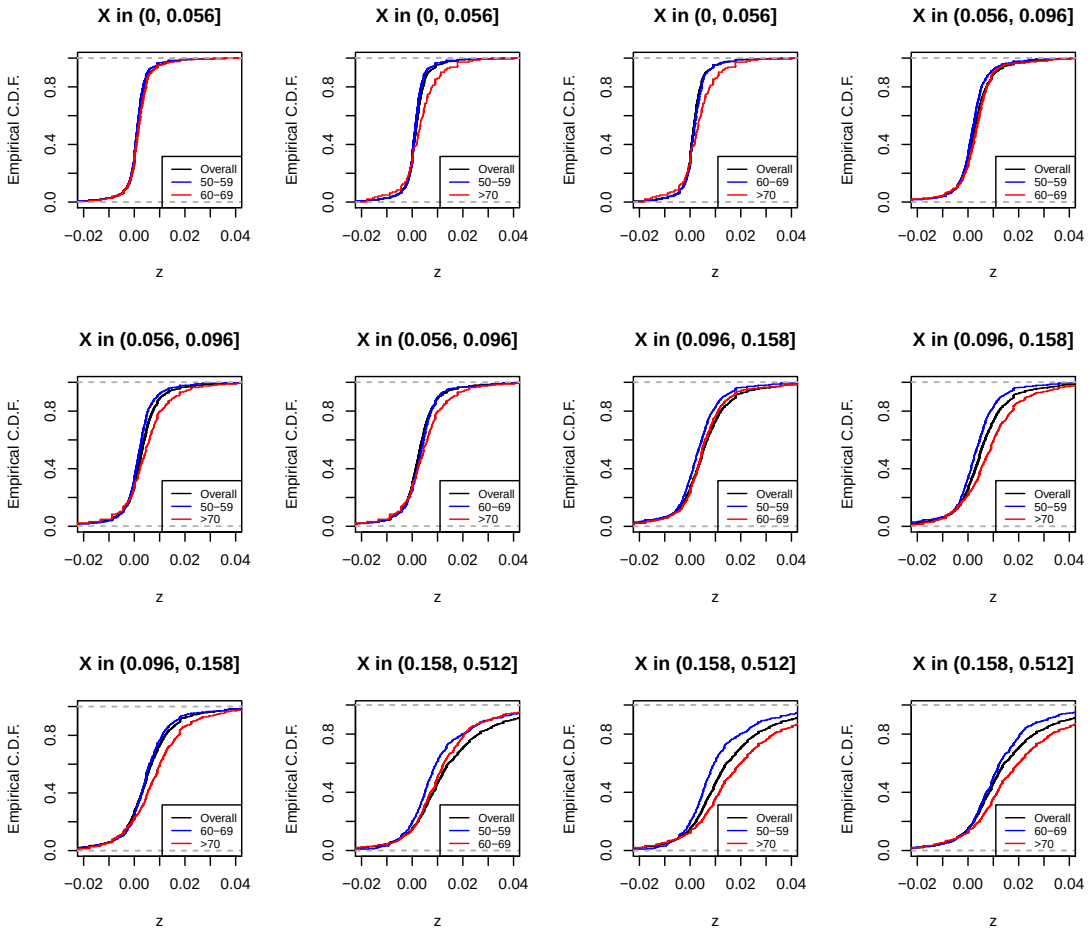


Figure B.9: Empirical c.d.f. estimates based on splitting the non-frail individuals into four level groups and further splitting them into 3 age groups. Level splits are based on quantiles. The figure demonstrates that the age dependence assumption is reasonable and should be included as a covariate into the final model.

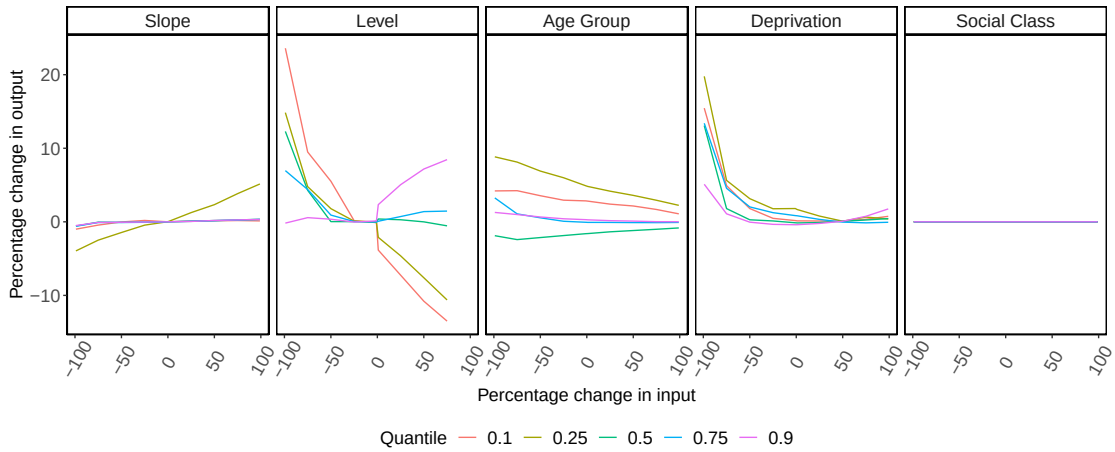


Figure B.10: Sensitivity analysis results for the effect of bandwidth selection on the estimated quantile trajectories. Overall undersmoothing leads to larger changes in the estimated quantile trajectories, with level bandwidth being the most crucial for model fit. The horizontal axis shows the percentage change in bandwidth which is plotted against the mean percentage change in the quantile estimates.

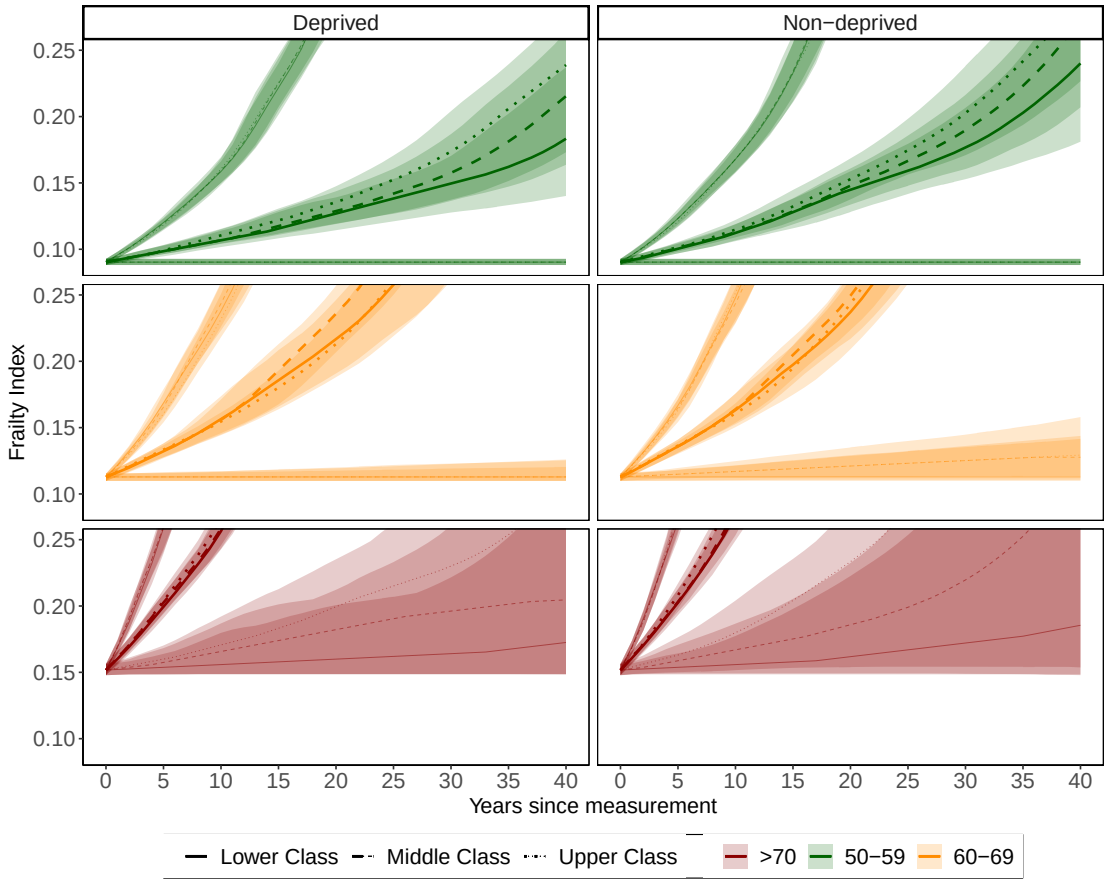


Figure B.11: Estimated quantile trajectories conditioned on the median per age group level along with 95% pointwise bootstrap confidence intervals (left) and difference in trajectories between non-deprived and deprived groups for all age-group and social class subgroups, along with 95% pointwise bootstrap confidence intervals for the difference (right). Apart from the marginal effects of age group, social class and deprivation we controlled for the interaction between age group, social class and deprivation and baseline. For the estimation, we used the cross-validation bandwidths: $h_H = 3.2 \times 10^{-4}$, $h_K = 1.7 \times 10^{-2}$ and $h_L = \{9.2 \times 10^{-2}, 4.7 \times 10^{-1}, 3.1 \times 10^{-1}\}$ for the age group, social status and deprivation variables. The bandwidths for the interaction terms *age group* $\times$ *social class* and *age group* $\times$ *deprivation* were 7.7×10^{-1} and 6.3×10^{-1} respectively. From top to bottom, α varies over 0.75, 0.50 (middle) and 0.25.

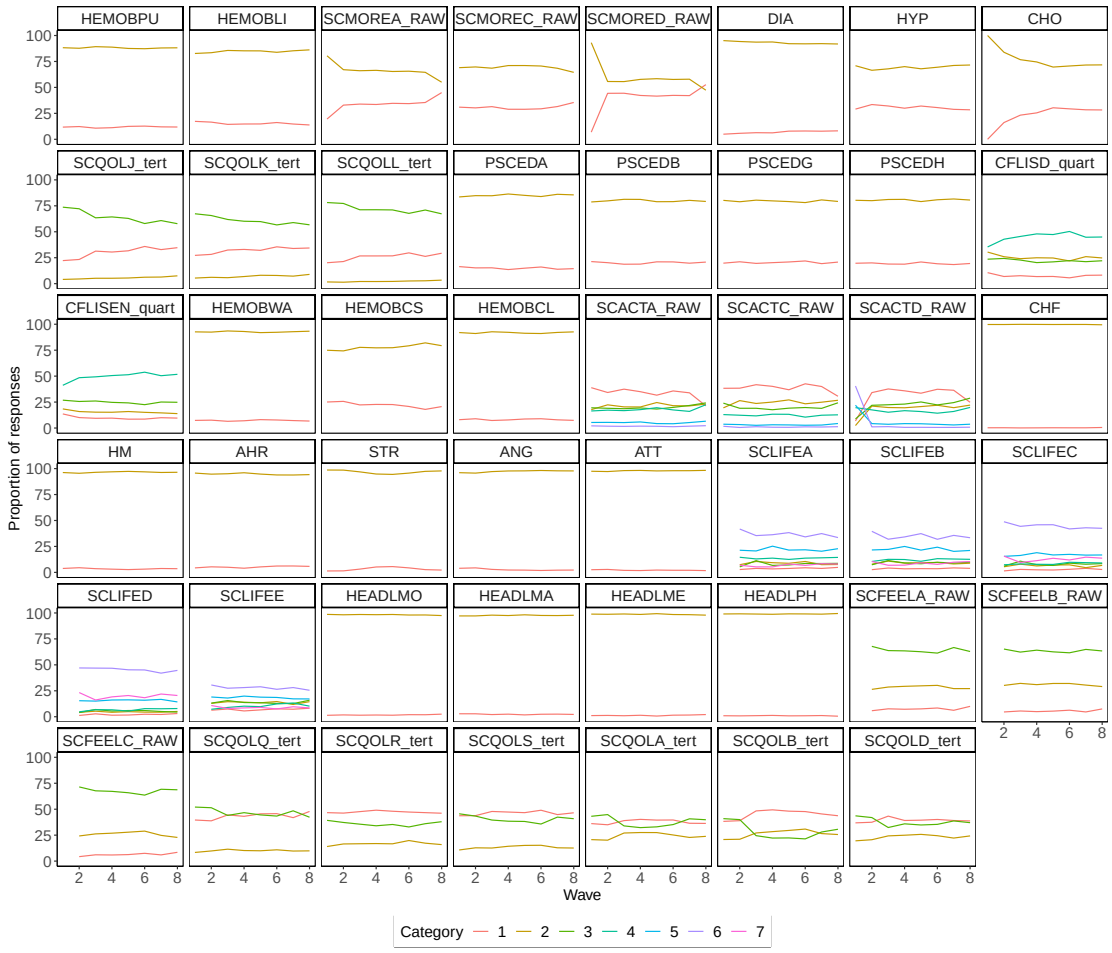


Figure B.12: Proportion of responses for each item level and variable for all domains over time for the 50s cohort.

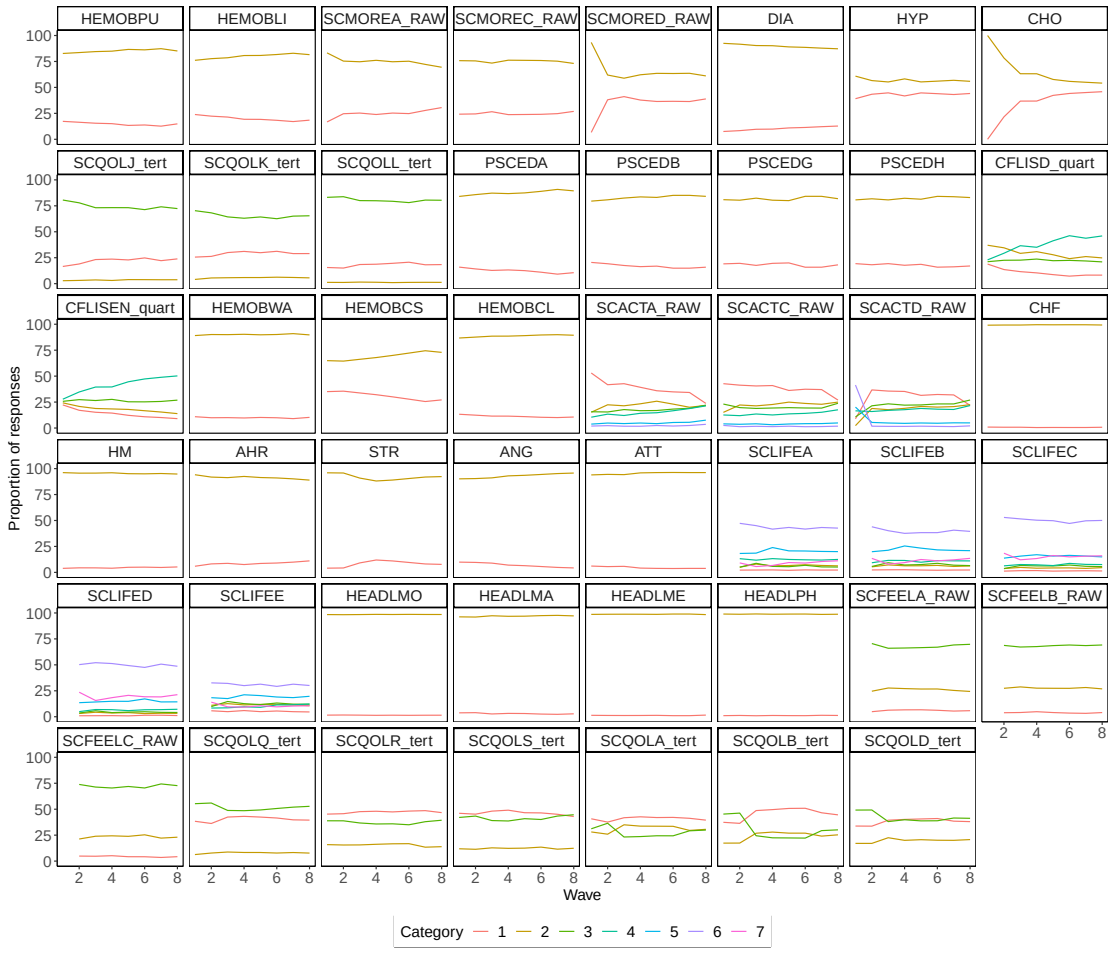


Figure B.13: Proportion of responses for each item level and variable for all domains over time for the 60s cohort.

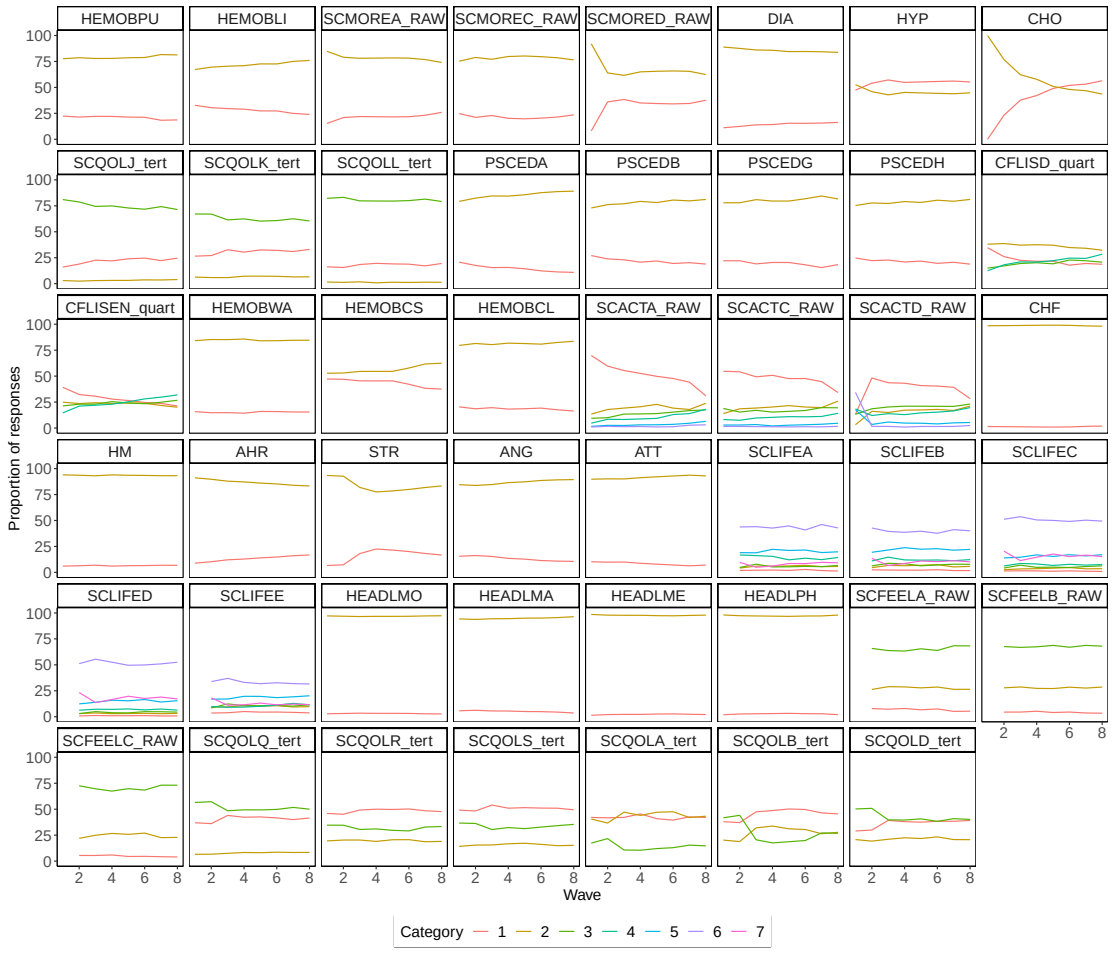


Figure B.14: Proportion of responses for each item level and variable for all domains over time for the 70s cohort.

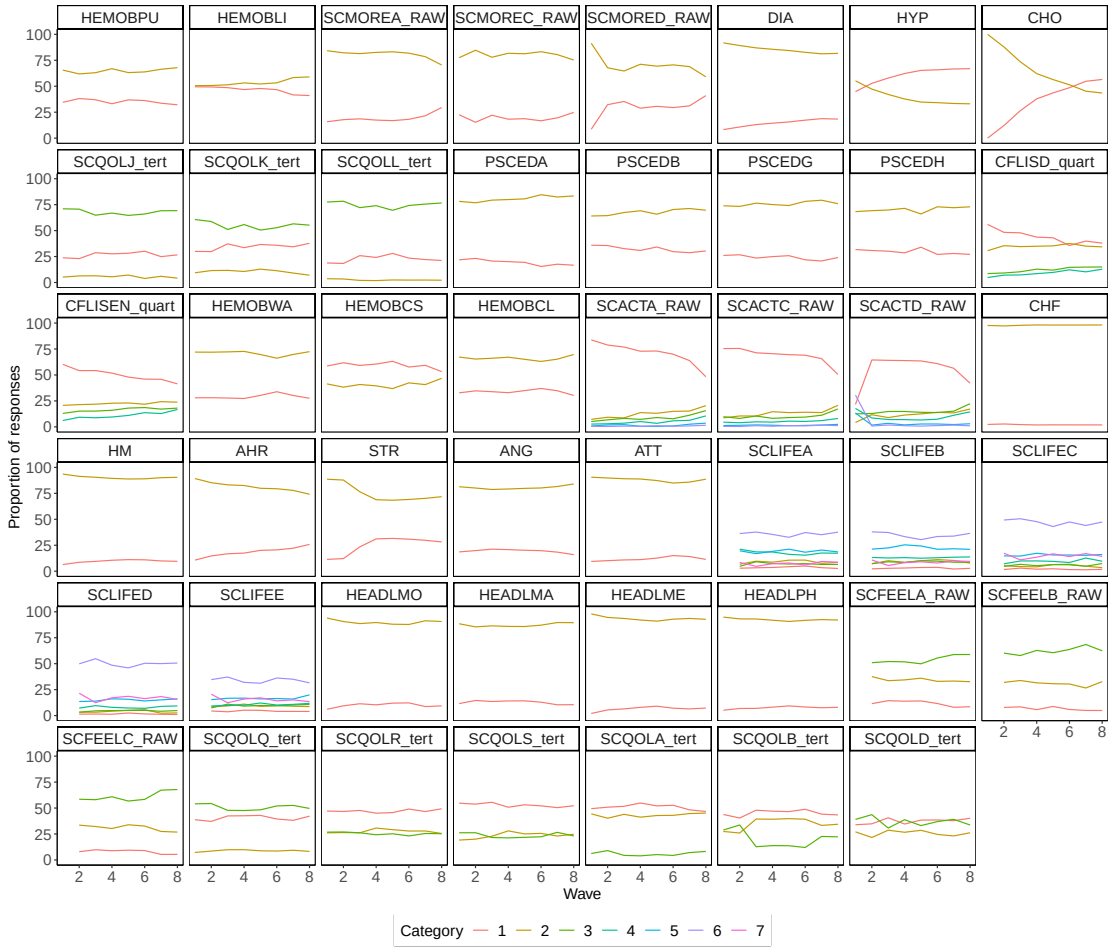
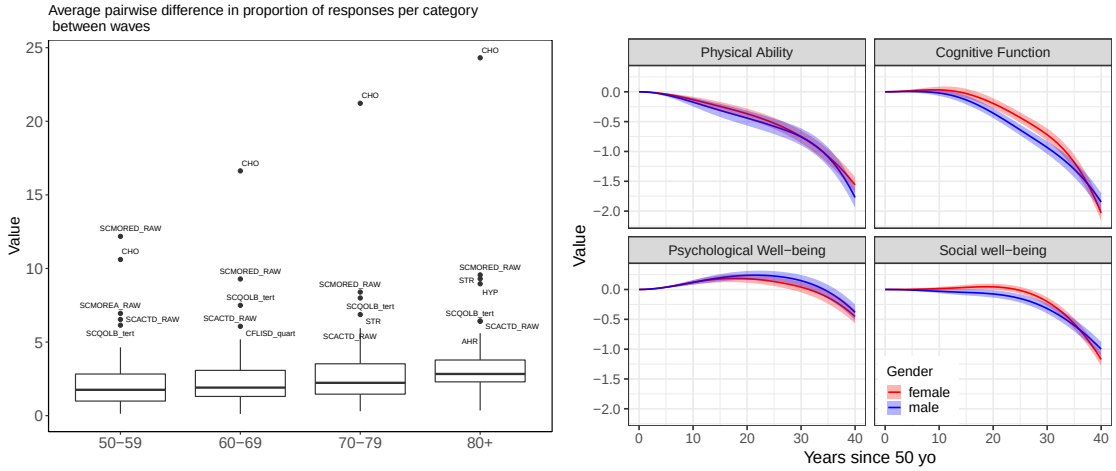


Figure B.15: Proportion of responses for each item level and variable for all domains over time for the 80s cohort.



(a) (b)

Figure B.16: (a) Average pairwise difference in proportion of responses per category between waves.(b) Latent mean curve estimated after fitting the univariate (ULGP) model for each gender data separately along with 95% confidence intervals.

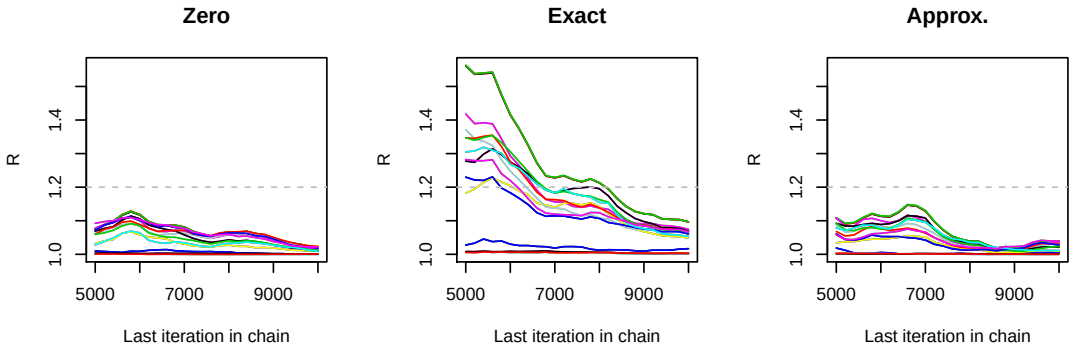


Figure B.17: Rubin statistics for the univariate laten Gaussian process model. Mean function defined as $\mu(t) = 0.8 \sin^3(0.15t)$.

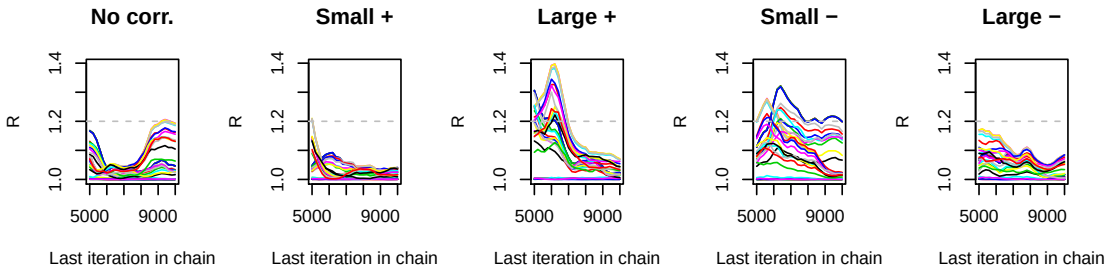


Figure B.18: Rubin statistics for the bivariate laten Gaussian process model. Mean functions where defined as $\mu_1(t) = 0$ and $\mu_2(t) = 0$.

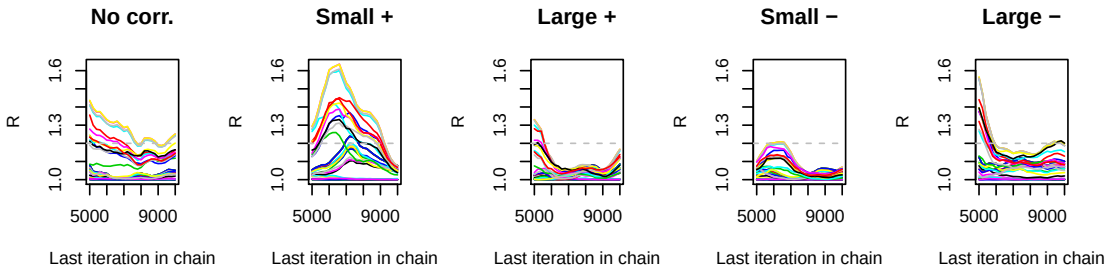


Figure B.19: Rubin statistics for the bivariate laten Gaussian process model. Mean functions where defined as $\mu_1(t) = 0$ and $\mu_2(t)$ generated using prespecified coefficients for the cubic B-splines.

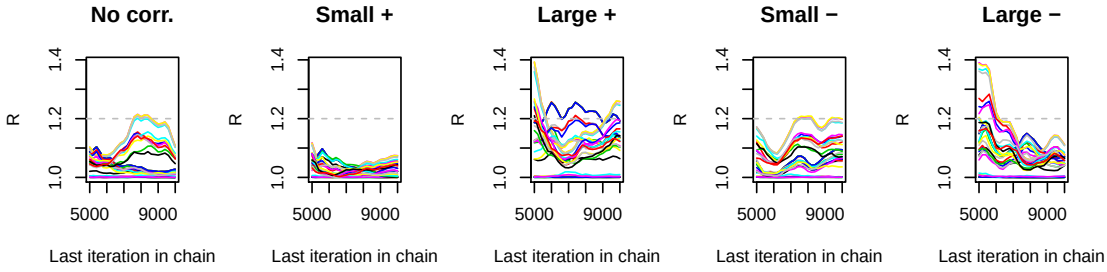


Figure B.20: Rubin statistics for the bivariate laten Gaussian process model. Mean functions where defined as $\mu_1(t) = 0$ and
$\mu_2(t) = 0.8 \sin^3(0.15t)$.

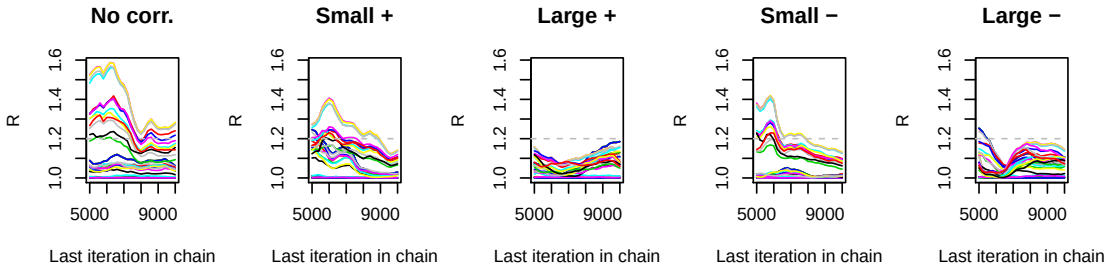


Figure B.21: Rubin statistics for the bivariate laten Gaussian process model. Mean functions where defined as $\mu_1(t) = 0.8 \sin^3(0.15t)$ and $\mu_2(t)$ generated using prespecified coefficients for the cubic B-splines.

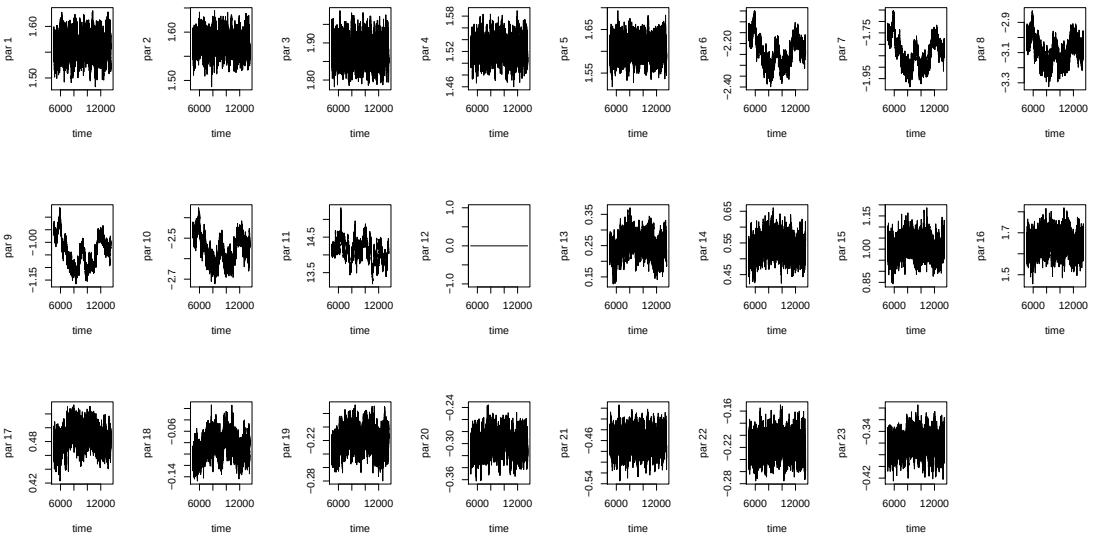


Figure B.22: Markov chains produced from the StEM algorithm of the ULGP model implementation on the physical ability ELSA data.

Figure B.23: Markov chains produced from the StEM algorithm of the ULGP model implementation on the psychological well-being ELSA data.

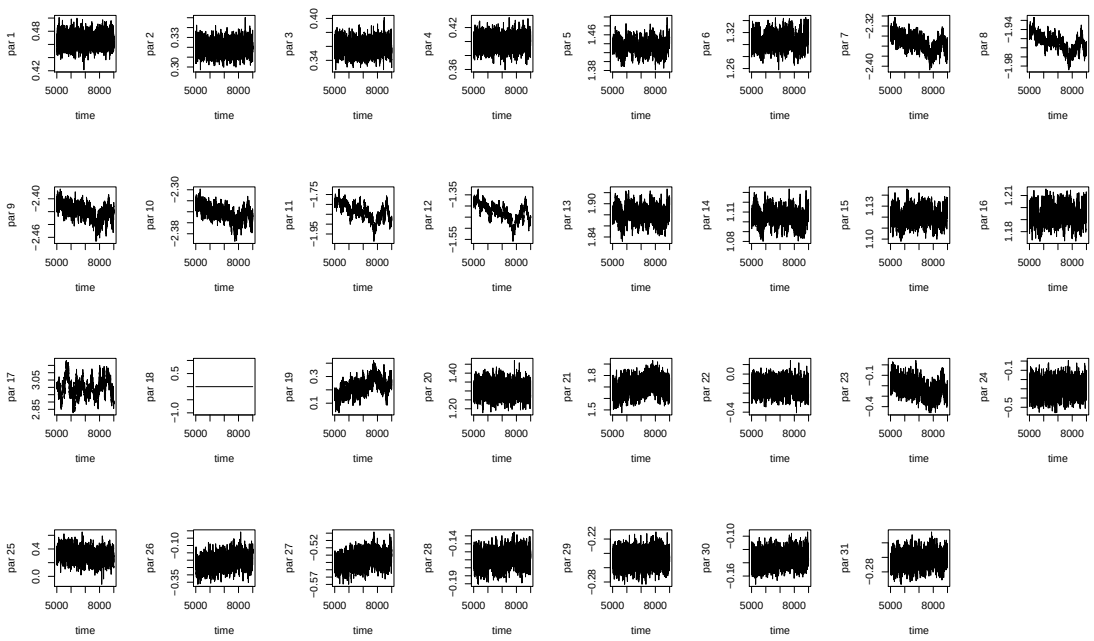


Figure B.24: Markov chains produced from the StEM algorithm of the ULGP model implementation on the cognitive function ELSA data.

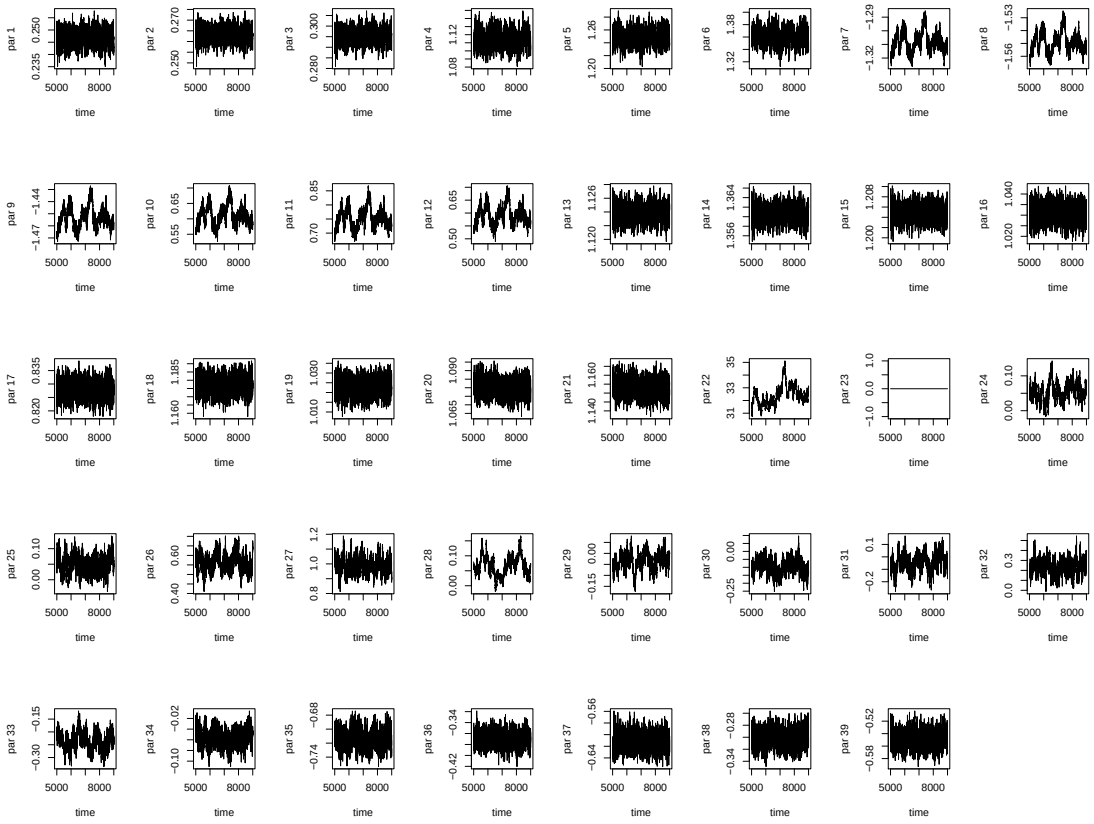


Figure B.25: Markov chains produced from the StEM algorithm of the ULGP model implementation on the social well-being ELSA data.

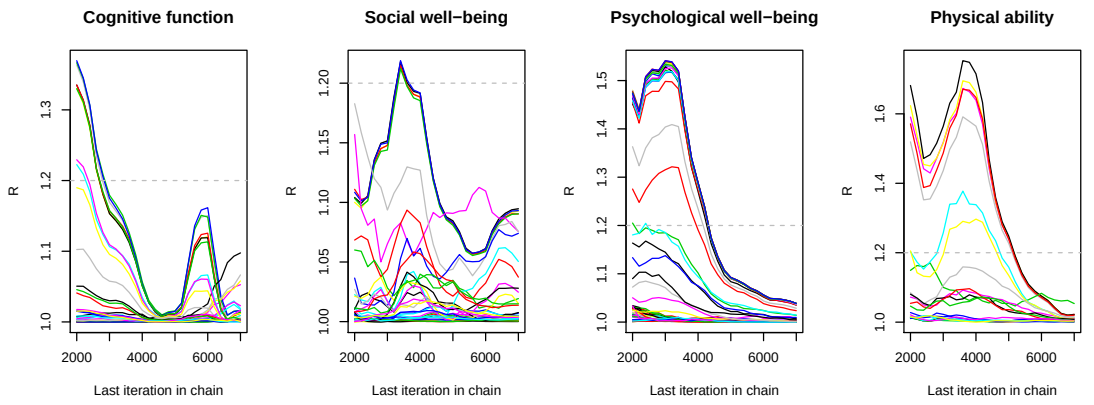


Figure B.26: Rubin statistics for the univariate laten Gaussian process model applied to the ELSA data.

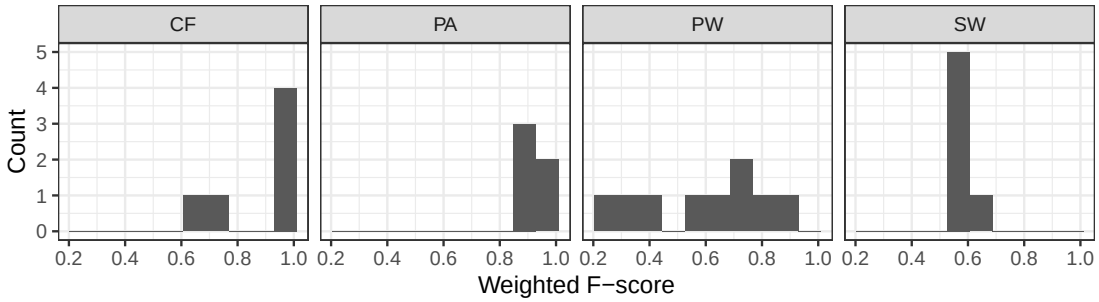
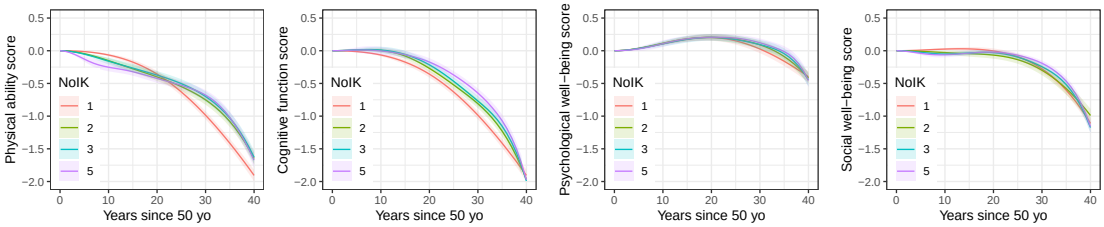


Figure B.27: Average prediction accuracy for each item based on the weighted F-score after performing 10-fold cross-validation using the LGP methodology implemented on the ELSA data. CF, PW, PA and SW correspond to cognitive function, psychological well-being, physical ability and social well-being respectively.



(a) (b) (c) (d)

Figure B.28: Physical ability (PA), cognitive function (CF), psychological well-being (PW) and social well-being (SW) latent mean curve estimates under varying number of interior knots (NoIK).

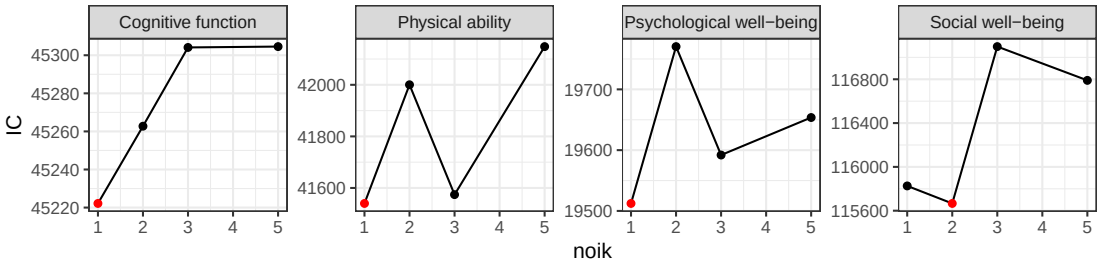


Figure B.29: Physical ability (PA), cognitive function (CF), psychological well-being (PW) and social well-being (SW) latent mean curve estimates under varying number of interior knots (NoIK).

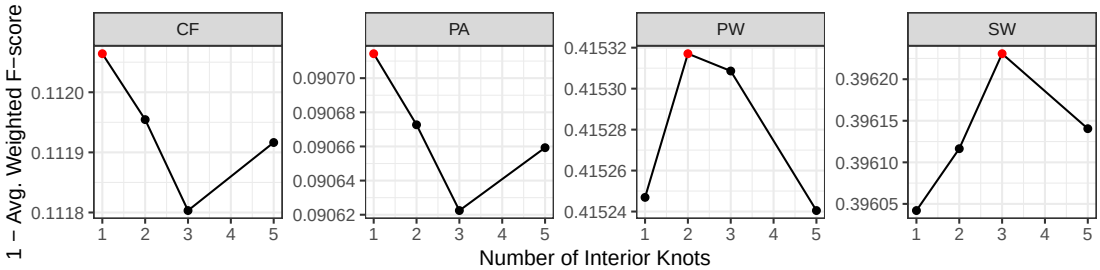


Figure B.30: Physical ability (PA), cognitive function (CF), psychological well-being (PW) and social well-being (SW) latent mean curve estimates under varying number of interior knots (NoIK).

Appendix C

Supplementary Tables

Ensembl Gene ID	Gene Name	Area	SD	Sign	Spearman Correlation	Mean fold change of BRAF mutation with respect to wild type	Protein-protein interaction network distance to BRAF
ENSG00000143621	KIR3DL1	0.370	0.107	-	-0.874	0.978	3
ENSG00000101057	CHST11	0.257	0.092	-	-0.817	0.899	NI
ENSG00000118113	APOC1P1	0.247	0.09	-	-0.918	1.190	NI
ENSG00000166173	PLEKHA6	0.239	0.086	-	-0.908	1.037	3
ENSG00000188290	PPM1F	0.223	0.068	+	0.910	0.883	3
ENSG00000139318	BFSP1	0.222	0.074	-	-0.800	1.217	NI
ENSG00000174099	PPP1R3A	0.217	0.082	+	0.774	1.078	3
ENSG00000143340	C16orf87	0.207	0.087	+	0.851	0.977	NI
ENSG00000143436	PARVA	0.203	0.081	+	0.890	0.984	2
ENSG00000116285	SLC39A13	0.202	0.079	-	-0.461	1.055	NI
ENSG00000136950	UCN2	0.198	0.07	-	-0.928	0.979	NI
ENSG00000102265	STMN3	0.198	0.087	+	0.834	1.201	2
ENSG00000106560	RNF130	0.197	0.083	-	-0.927	1.153	NI
ENSG00000196954	C3orf58	0.196	0.076	+	0.922	1.133	NI
ENSG00000198837	CXXC4	0.188	0.079	+	0.866	0.995	NI
ENSG00000197702	THBD	0.179	0.093	0	-0.967	1.231	4
ENSG00000173542	SIRT3	0.173	0.066	-	-0.760	1.013	3
ENSG00000105679	PLAT	0.172	0.092	-	-0.878	1.322	4
ENSG00000116580	MPPED1	0.168	0.066	+	0.430	0.978	NI
ENSG00000085274	INSL3	0.162	0.068	-	-0.973	0.965	NI
ENSG00000165915	FAM163A	0.159	0.078	-	-0.983	1.106	NI
ENSG00000197329	CNIH3	0.153	0.08	-	-0.918	0.938	NI
ENSG00000143793	GJA3	0.153	0.067	0	-0.940	0.933	NI
ENSG00000105355	BTG2	0.152	0.078	+	0.959	1.035	2
ENSG00000189325	DLX6	0.152	0.059	0	0.686	0.987	NI
ENSG00000198829	DLC1	0.151	0.053	-	-0.928	0.974	3
ENSG00000110031	GAPDHS	0.150	0.077	+	0.886	1.232	NI
ENSG00000147894	JAG2	0.149	0.069	-	-0.994	0.981	3
ENSG00000084731	SMOX	0.146	0.057	0	0.816	1.070	NI
ENSG00000106077	ZMYND8	0.145	0.091	+	0.907	1.020	3
ENSG00000117360	PDPK1	0.143	0.067	-	-0.867	0.950	3
ENSG00000035681	SYNJ1	0.143	0.048	+	0.948	1.037	4
ENSG00000186868	PHF23	0.140	0.086	+	0.926	0.942	NI
ENSG00000123560	KIF3C	0.140	0.138	0	0.801	1.588	NI
ENSG00000111145	DUSP16	0.138	0.075	+	0.937	1.138	2
ENSG00000143543	MYBL2	0.137	0.054	-	-0.443	0.979	2

ENSG00000134531	CARD19	0.137	0.075	0	0.999	1.273	NI
ENSG00000131732	SHROOM3	0.133	0.071	+	0.891	1.060	4
ENSG00000126088	CCDC120	0.132	0.068	+	0.741	1.045	NI
ENSG00000139116	LBX2	0.129	0.064	-	-0.965	0.942	NI
ENSG00000131773	SLC13A3	0.129	0.054	+	0.665	1.155	NI
ENSG00000183696		0.127	0.072	-	-0.795	1.195	NI
ENSG00000120685	GON4L	0.127	0.064	0	0.930	0.978	NI
ENSG00000126870	MACC1	0.125	0.06	-	-0.593	0.996	NI
ENSG00000167633	RAB32	0.124	0.058	-	-0.835	1.008	NI
ENSG00000143365	SPRY1	0.124	0.055	+	0.904	0.981	4
ENSG00000104365	HSPG2	0.122	0.061	-	-0.979	1.001	NI
ENSG00000006377	HES4	0.122	0.07	+	0.898	0.889	NI
ENSG00000186143	FAM43A	0.121	0.063	+	0.967	1.002	NI
ENSG00000145107	ICOS	0.120	0.069	-	-0.883	1.147	NI
ENSG00000135537	MAG	0.120	0.053	-	-0.932	1.028	4
ENSG00000155330	IL1R2	0.120	0.07	+	0.915	0.875	4
ENSG00000248099	AKAP5	0.119	0.039	-	-0.330	0.994	3
ENSG00000172209	RBSN	0.119	0.052	+	0.736	0.968	3
ENSG00000244439	PRR30	0.116	0.064	+	0.107	0.962	NI
ENSG00000075415	LRP5	0.116	0.061	0	-0.974	1.010	3
ENSG00000026025	DUSP8	0.115	0.094	+	0.927	1.152	NI
ENSG00000150760	C1orf35	0.114	0.065	-	-0.888	1.070	NI
ENSG00000125351	COMMD2	0.112	0.055	+	0.930	0.947	NI
ENSG00000109472	DCAF12L2	0.112	0.074	+	0.857	0.792	NI
ENSG00000175550	SEMA6D	0.112	0.06	+	0.619	1.037	NI
ENSG00000128829	MAPT	0.111	0.036	-	0.032	1.048	2
ENSG00000100034	RGS19	0.111	0.063	+	0.558	1.039	2
ENSG00000148672	NECTIN1	0.108	0.071	-	-0.973	0.993	3
ENSG00000143786	CAMK2B	0.108	0.074	+	0.736	1.267	3
ENSG00000131381	ATAD3C	0.107	0.036	+	0.370	0.999	NI
ENSG00000036672	DRAP1	0.107	0.038	+	0.532	0.968	NI
ENSG00000175877	GDF11	0.107	0.032	0	0.648	0.993	NI
ENSG00000161265	SDHAP3	0.106	0.041	0	0.670	0.944	NI
ENSG00000069424	RAB34	0.104	0.083	-	-0.816	1.079	4
ENSG00000182500	SEC31B	0.104	0.053	+	0.671	1.037	NI
ENSG00000109113	A2M	0.102	0.042	-	-0.663	1.065	5
ENSG00000159147	EMC9	0.101	0.055	+	0.988	0.976	NI
ENSG00000170382	PAIP2B	0.100	0.042	+	1.000	0.948	NI
ENSG00000152217	STPG3	0.099	0.051	-	-0.502	0.928	NI
ENSG00000101216	DENND4B	0.099	0.06	+	0.649	0.989	NI
ENSG00000163958	ST8SIA2	0.099	0.043	+	0.300	0.983	NI
ENSG00000140992	ATL1	0.098	0.034	+	0.703	0.987	NI
ENSG00000175899	SERPIND1	0.097	0.066	-	-0.871	1.337	4
ENSG00000138018	ITPRIP	0.097	0.062	0	0.655	0.978	NI
ENSG00000183935	ANTXR2	0.096	0.063	-	-0.997	1.058	NI
ENSG00000156206	ABHD11	0.096	0.055	-	-0.619	0.990	NI
ENSG00000132256	LUZP1	0.096	0.048	0	-0.299	1.031	NI
ENSG00000178935	ANGPT1	0.096	0.079	0	0.852	0.959	4
ENSG00000119138	RFLNB	0.096	0.074	0	-0.609	1.147	NI
ENSG00000187048	CORO1B	0.095	0.052	+	0.821	0.982	3
ENSG00000087258	GGT6	0.094	0.056	-	-0.341	0.988	NI
ENSG00000130703	PROSER1	0.094	0.074	+	0.891	1.006	NI
ENSG00000182541	C6orf222	0.094	0.071	0	0.538	1.007	NI
ENSG00000142227	DNMBP	0.093	0.079	-	-0.426	1.190	NI
ENSG00000058404	CFAP161	0.092	0.08	-	-0.986	0.930	NI
ENSG00000142798	IGF2R	0.091	0.051	-	-0.694	1.032	3
ENSG00000170955	ACP1	0.091	0.064	-	-0.704	1.290	3
ENSG00000011347	ARPC5L	0.091	0.045	-	-0.936	0.967	NI

ENSG00000186732	MAFB	0.091	0.046	+	0.835	0.984	3
ENSG00000172725	MARC1	0.091	0.057	0	-0.989	1.032	NI
ENSG00000068366	SYT7	0.091	0.055	0	-0.768	1.074	NI
ENSG00000163600	TM4SF19	0.090	0.051	-	-0.990	0.975	NI
ENSG00000133687	LMBRD2	0.090	0.053	+	0.988	1.064	NI
ENSG00000186205	UBTD1	0.090	0.046	0	0.620	0.838	NI
ENSG00000077152	ZNF552	0.090	0.056	0	0.055	0.951	NI
ENSG00000152804	TMEM270	0.089	0.047	+	0.434	1.073	NI
ENSG00000164741	FHL3	0.088	0.067	+	0.068	1.200	2
ENSG00000137575	BCL2A1	0.087	0.074	+	0.897	1.099	3
ENSG00000204103	NSMAF	0.086	0.056	+	0.903	0.932	3
ENSG00000168772	EMP3	0.086	0.077	+	0.698	0.884	NI
ENSG00000145040	TNFRSF10D	0.086	0.051	0	-0.292	1.228	3
ENSG00000154188	ORAI1	0.085	0.044	-	-0.067	1.069	4
ENSG00000179528	PTP4A2	0.085	0.061	-	-0.996	1.066	NI
ENSG00000197457	TMTC1	0.083	0.044	-	-0.623	0.982	NI
ENSG00000137872	PPP1R13B	0.083	0.048	+	1.000	1.007	NI
ENSG00000078795	ITGA10	0.082	0.057	-	-0.812	1.004	4
ENSG00000184007	KCNAB2	0.081	0.051	-	-0.856	1.007	3
ENSG00000171700	ZCCHC9	0.081	0.073	0	-0.789	1.009	NI
ENSG00000112851	CDH24	0.079	0.052	-	-0.930	1.017	NI
ENSG00000143319	HTR7P1	0.079	0.052	0	0.730	0.970	NI
ENSG00000088826	LHFPL3	0.078	0.034	-	-0.391	1.072	NI
ENSG00000131470	MSRB3	0.078	0.061	0	0.391	0.966	NI
ENSG00000185112	MRPL9	0.077	0.036	-	-0.297	1.129	NI
ENSG00000118508	PAPPA-AS1	0.077	0.048	+	-0.120	1.152	NI
ENSG00000164056	SPATA31D3	0.075	0.04	+	0.351	1.073	NI
ENSG00000099937	GMEB2	0.074	0.044	-	0.028	1.047	NI
ENSG00000105643	HHEX	0.073	0.06	0	-0.771	0.988	3
ENSG00000184916	EIF2AK4	0.073	0.065	0	-0.735	0.840	4
ENSG00000169116	PRPF3	0.072	0.048	0	-0.511	0.931	3
ENSG00000143850	SELENOI	0.071	0.053	0	0.324	0.935	NI
ENSG00000086758	PRSS16	0.070	0.061	+	0.920	0.969	NI
ENSG00000139880	KLF9	0.069	0.058	-	-0.533	0.957	NI
ENSG00000181744	FAM84A	0.069	0.054	0	-0.861	0.963	NI
ENSG00000162337	GPR22	0.069	0.048	0	0.537	0.966	NI
ENSG00000165233	ST5	0.069	0.043	0	-0.636	1.045	NI
ENSG00000102241	ASB9	0.068	0.042	0	-0.995	0.954	NI
ENSG00000187416	UROD	0.068	0.053	0	0.617	0.995	NI
ENSG00000183742	TMEM254	0.067	0.056	0	0.597	0.906	NI
ENSG00000169641	PRSS33	0.066	0.056	0	-0.520	1.017	NI
ENSG00000165886	ARRDC2	0.066	0.044	0	0.704	1.055	NI
ENSG00000171310	PELI1	0.065	0.045	-	-0.174	1.231	4
ENSG00000104368	UBE2Q2	0.065	0.052	+	0.146	1.410	NI
ENSG00000143127	DUSP6	0.064	0.042	-	-0.897	1.065	2
ENSG00000120457	SLFN12	0.064	0.051	-	-0.897	0.997	NI
ENSG00000198354	KHDRBS3	0.063	0.035	+	0.560	0.991	NI
ENSG00000115590	SCGB3A1	0.063	0.045	0	-0.297	1.009	NI
ENSG00000173221	CUBN	0.063	0.045	0	0.408	1.116	NI
ENSG00000152784	VIM	0.061	0.059	0	0.323	1.002	2
ENSG00000166444	CYP4A11	0.060	0.065	-	-1.000	1.077	NI
ENSG00000088808	GIMAP2	0.060	0.043	-	-0.903	0.962	NI
ENSG00000179841	VOPP1	0.060	0.039	+	0.012	0.977	NI
ENSG00000153551	MMP8	0.060	0.049	0	-0.766	1.198	NI
ENSG00000173530	LPXN	0.060	0.048	0	-0.569	1.053	3
ENSG00000206190	TRIM5	0.060	0.04	0	-0.492	1.013	NI
ENSG00000159082	PKD2L2	0.059	0.052	+	0.796	1.006	4
ENSG00000138771	SEC23B	0.058	0.052	+	-0.286	0.908	NI

ENSG00000069943	PRDM8	0.057	0.064	-	-1.000	1.050	NI
ENSG00000140379	PSMC3IP	0.057	0.049	0	0.716	1.665	NI
ENSG00000158296	JTB	0.056	0.041	-	0.238	1.041	NI
ENSG00000154415	SUCNR1	0.056	0.03	-	0.021	0.995	NI
ENSG00000163297	NRP2	0.056	0.068	0	0.076	1.110	4
ENSG00000183688	CPE	0.056	0.047	0	-0.461	1.120	NI
ENSG00000138443	CMTM7	0.056	0.051	0	-0.895	1.004	NI
ENSG00000110400	SDCBP	0.055	0.041	+	-0.310	0.931	4
ENSG00000186300	PARM1	0.054	0.044	-	0.202	0.978	NI
ENSG00000103355	SETBP1	0.054	0.034	-	0.248	1.184	NI
ENSG00000075826	SCNN1B	0.054	0.054	0	0.273	0.965	2
ENSG00000101310	LIMK2	0.053	0.038	+	0.995	1.021	3
ENSG00000164187	ABLIM1	0.053	0.045	+	0.971	0.943	2
ENSG00000114473	ZNF280A	0.053	0.048	0	0.722	1.014	NI
ENSG00000143727	WDR60	0.052	0.043	+	0.744	0.986	NI
ENSG00000145029	PMP22	0.051	0.042	0	-0.522	1.037	NI
ENSG00000100908	GLUD1	0.051	0.045	0	-0.094	0.960	4
ENSG00000143368	C9orf72	0.051	0.051	0	-0.670	1.005	NI
ENSG00000074660	SERPINA3	0.050	0.051	-	-0.716	1.033	4
ENSG00000126878	NICN1	0.050	0.04	+	0.866	0.978	NI
ENSG00000139438	USP2	0.050	0.052	0	-0.646	0.985	4
ENSG00000214855	RHOXF1	0.049	0.046	-	-0.891	0.992	NI
ENSG00000154978	UPP1	0.048	0.049	-	-0.883	0.967	NI
ENSG00000111266	SF3B4	0.047	0.037	-	-0.682	0.990	4
ENSG00000198513	RORC	0.046	0.051	+	0.463	0.917	NI
ENSG00000184545	UBE2T	0.046	0.038	0	-0.586	0.987	5
ENSG00000183386	CYP17A1	0.046	0.041	0	-0.675	1.048	3
ENSG00000112812	ABI2	0.046	0.045	0	-0.499	0.911	3
ENSG00000186788	ERBIN	0.045	0.03	0	0.433	1.007	3
ENSG00000148488	HTATSF1	0.044	0.062	0	-0.596	0.996	4
ENSG00000162981	ST8SIA6	0.044	0.03	0	0.735	0.987	NI
ENSG00000118557	IKBKB	0.043	0.034	+	0.003	0.985	3
ENSG00000168447	TIMP1	0.043	0.031	+	0.937	0.955	3
ENSG00000125864	DOCK1	0.043	0.039	0	-0.592	1.110	2
ENSG00000142082	AFG1L	0.042	0.049	-	-0.401	1.003	NI
ENSG00000148841	ZBP1	0.042	0.06	+	0.338	1.106	NI
ENSG00000135298	U2AF1L4	0.042	0.057	0	0.717	0.898	NI
ENSG00000058335	RASGRF1	0.041	0.043	-	-0.547	1.029	3
ENSG00000113269	PLP1	0.041	0.047	+	0.730	1.071	3
ENSG00000148795	GNAO1	0.041	0.056	0	0.439	0.994	3
ENSG00000140557	CAVIN3	0.040	0.052	-	-0.661	0.979	NI
ENSG00000161055	LARP6	0.040	0.031	0	-0.042	0.986	3
ENSG00000147144	OSBPL2	0.040	0.041	0	-0.016	0.976	NI
ENSG00000182890	GLUD2	0.038	0.06	0	-0.653	1.097	NI
ENSG00000196136	FAM222A	0.037	0.061	0	-0.372	1.249	NI
ENSG00000040633	MOB1B	0.037	0.045	0	-0.734	1.042	3
ENSG00000256040	HUWE1	0.037	0.049	0	-0.760	1.021	4
ENSG00000159388	PLIN3	0.036	0.046	0	-0.901	0.887	NI
ENSG00000114744	ZNF555	0.034	0.054	+	0.832	0.954	NI
ENSG00000107554	CASP4	0.034	0.05	0	0.424	1.050	NI
ENSG00000135414	PMFBP1	0.033	0.051	0	-0.205	1.096	NI
ENSG00000197768	SCARF1	0.032	0.047	+	0.793	0.989	5
ENSG00000107611	ZDHHC19	0.032	0.045	0	-0.219	1.051	NI
ENSG00000101883	ILF2	0.031	0.036	0	-0.175	1.001	NI
ENSG00000215915	MYNN	0.030	0.046	+	1.000	0.989	NI
ENSG00000169548	UPF3B	0.029	0.043	0	0.540	1.120	NI
ENSG00000068097	HEATR6	0.029	0.059	0	0.199	0.983	NI
ENSG00000172123	SLC25A3	0.028	0.056	0	0.538	1.118	NI

ENSG00000124256	KIF21A	0.028	0.034	0	-0.314	0.982	4
ENSG00000178726	ADGRB3	0.028	0.045	0	0.747	0.958	NI
ENSG00000109099	DONSON	0.027	0.06	0	-0.428	1.312	NI
ENSG00000121743	ELK3	0.026	0.04	0	-0.017	1.154	2
ENSG00000101040	ACP7	0.026	0.048	0	0.010	1.042	NI
ENSG00000133678	LRRN2	0.026	0.039	0	-0.528	0.967	NI
ENSG00000197081	C18orf25	0.026	0.056	0	0.311	1.037	NI
ENSG00000053918	IQCG	0.026	0.038	0	-0.045	0.994	NI
ENSG00000099204	ATP10A	0.025	0.041	0	-0.117	0.898	NI
ENSG00000167741	KCNQ1	0.024	0.049	0	0.918	0.979	3
ENSG00000118257	PIGB	0.024	0.046	0	0.771	1.157	NI
ENSG00000152242	KCNJ5	0.024	0.036	0	0.028	0.984	3
ENSG00000105497	ACSL4	0.023	0.048	0	-0.255	1.017	NI
ENSG00000183760	ISG20L2	0.023	0.039	0	-0.680	1.014	NI
ENSG00000197893	ZNF175	0.023	0.031	0	-0.354	1.004	NI
ENSG00000124374	GLRX	0.022	0.04	0	0.611	0.959	NI
ENSG00000102048	ERRFI1	0.022	0.042	0	0.538	1.176	3
ENSG00000185986	EMP1	0.022	0.038	0	-0.549	0.935	NI
ENSG00000105695	AIF1L	0.019	0.054	-	-0.688	0.992	NI
ENSG00000140367	NRAP	0.016	0.036	0	-0.069	1.020	NI

Table C.1: Gene rankings of all selected genes based on the magnitude of the genetic effect on drug response. A positive (+) sign translates to a positive effect on cells survival after drug administration, a negative (-) sign translates to a negative effect on cells survival and a mixed (0) effect translates to a varying effect on cells survival which depends on drug dosage. Spearman's correlation is calculated between drug dosage and gene estimated coefficient function values as an indicator of the magnitude change of the gene effect over the increasing dosage. Area corresponds to the area under the estimated coefficient curve and the SD corresponds to the standard deviation of the area based on bootstrapping. Mean fold change is calculated between the selected gene expression values of the cell lines carrying BRAF mutations with respect to wild type. Protein-protein interaction network distance is computed based on the shortest interaction path between the BRAF gene and each of the selected genes. Here, NI denotes absence of any interaction.

Pathway	**Number of genes**	**P-value**	**Adjusted p-value**
Glial Cell Differentiation WP2276	2	0	0.000
NLR Proteins WP288	2	0	0.000
Supression of HMGB1 mediated inflammation by THBD WP4479	2	0	0.000
Mammary gland development pathway - Puberty (Stage 2 of 4) WP2814	2	0	0.000
Apoptosis Modulation and Signaling WP1772	5	0.001	0.023
Primary Focal Segmental Glomerulosclerosis FSGS WP2572	4	0.001	0.023
NOTCH1 regulation of human endothelial cell calcification WP3413	2	0.001	0.023
Osteopontin Signaling WP1434	1	0.001	0.023
Sulindac Metabolic Pathway WP2542	1	0.002	0.041
Vitamin B12 Metabolism WP1533	3	0.003	0.050
Glucocorticoid and Mineralcorticoid Metabolism WP237	1	0.003	0.050
T-Cell antigen Receptor (TCR) Signaling Pathway WP69	4	0.004	0.052
Role Altered Glycolysation of MUC1 in Tumour Microenvironment WP4480	1	0.004	0.052
MAPK and NFkB Signalling Pathways Inhibited by Yersinia YopJ WP3849	1	0.004	0.052
Heme Biosynthesis WP561	1	0.005	0.057
The alternative pathway of fetal androgen synthesis WP4524	1	0.005	0.057
Complement and Coagulation Cascades WP558	3	0.007	0.058
Oligodendrocyte Specification and differentiation WP4304	2	0.007	0.058
Matrix Metalloproteinases WP129	2	0.006	0.058
Fibrin Complement Receptor 3 Signaling Pathway WP4136	2	0.006	0.058
TFs Regulate miRNAs related to cardiac hypertrophy WP1559	1	0.007	0.058

NAD metabolism, sirtuins and aging WP3630	1	0.007	0.058
Calcium Regulation in the Cardiac Cell WP536	5	0.009	0.069
Steroid Biosynthesis WP496	1	0.009	0.069
Selenium Micronutrient Network WP15	3	0.01	0.070
Photodynamic therapy-induced NF-kB survival signaling WP3617	2	0.01	0.070
MAPK Signaling Pathway WP382	7	0.013	0.072
Spinal Cord Injury WP2431	4	0.014	0.072
PPAR signaling pathway WP3942	3	0.011	0.072
Photodynamic therapy-induced HIF-1 survival signaling WP3614	2	0.012	0.072
Nucleotide-binding Oligomerization Domain (NOD) pathway WP1433	2	0.013	0.072
Genes targeted by miRNAs in adipocytes WP1992	1	0.014	0.072
Classical pathway of steroidogenesis, including diseases WP4523	1	0.015	0.072
Fatty Acid Omega Oxidation WP206	1	0.014	0.072
NAD+ metabolism WP3644	1	0.015	0.072
ID signaling pathway WP53	1	0.015	0.072
Serotonin and anxiety WP3947	1	0.012	0.072
Leptin Insulin Overlap WP3935	1	0.013	0.072
Notch Signaling WP268	2	0.017	0.080
Oxysterols derived from cholesterol WP4545	1	0.019	0.085
Mitochondrial LC-Fatty Acid Beta-Oxidation WP368	1	0.019	0.085
EGF/EGFR Signaling Pathway WP437	5	0.02	0.087
Interleukin-11 Signaling Pathway WP2332	2	0.021	0.089
MicroRNAs in cardiomyocyte hypertrophy WP1544	3	0.022	0.089
Vitamin B12 Disorders WP4271	1	0.022	0.089
TGF-B Signaling in Thyroid Cells for Epithelial-Mesenchymal Transition WP3859	1	0.026	0.095
Serotonin Receptor 2 and ELK-SRF/GATA4 signaling WP732	1	0.026	0.095
Urea cycle and metabolism of amino groups WP497	1	0.026	0.095
NAD+ biosynthetic pathways WP3645	1	0.025	0.095
Type II diabetes mellitus WP1584	1	0.025	0.095
B Cell Receptor Signaling Pathway WP23	3	0.03	0.096
Cardiac Hypertrophic Response WP2795	2	0.03	0.096
Regulation of Apoptosis by Parathyroid Hormone-related Protein WP3872	1	0.03	0.096
Fatty Acid Biosynthesis WP357	1	0.027	0.096
Vitamin D in inflammatory diseases WP4482	1	0.031	0.096
Blood Clotting Cascade WP272	1	0.031	0.096
EBV LMP1 signaling WP262	1	0.029	0.096
Estrogen signaling pathway WP712	1	0.03	0.096
Angiogenesis WP1539	1	0.029	0.096
Intraflagellar transport proteins binding to dynein WP4532	1	0.032	0.098
Brain-Derived Neurotrophic Factor (BDNF) signaling pathway WP2380	4	0.035	0.103
Integrin-mediated Cell Adhesion WP185	3	0.035	0.103
Focal Adhesion WP306	5	0.039	0.106
IL-1 signaling pathway WP195	2	0.041	0.106
Notch Signaling Pathway WP61	2	0.041	0.106
Oxidation by Cytochrome P450 WP43	2	0.041	0.106
T-Cell antigen Receptor (TCR) pathway WP3863	2	0.041	0.106
Endochondral Ossification WP474	2	0.037	0.106
Methionine De Novo and Salvage Pathway WP3580	1	0.037	0.106
PPAR Alpha Pathway WP2878	1	0.04	0.106
One Carbon Metabolism WP241	1	0.04	0.106
Wnt Signaling WP428	3	0.044	0.107
MECP2 and Associated Rett Syndrome WP3584	2	0.045	0.107
Folate Metabolism WP176	2	0.045	0.107
TLR4 Signaling and Tolerance WP3851	1	0.044	0.107
Toll-like Receptor Signaling WP3858	1	0.045	0.107
Constitutive Androstane Receptor Pathway WP2875	1	0.045	0.107
Canonical and Non-canonical Notch signaling WP3845	1	0.048	0.113
Wnt/beta-catenin Signaling Pathway in Leukemia WP3658	1	0.049	0.114

Gastric Cancer Network 1 WP2361	1	0.051	0.117
IL1 and megakaryocytes in obesity WP2865	1	0.052	0.117
Sterol Regulatory Element-Binding Proteins (SREBP) signalling WP1982	2	0.053	0.118
Gastric Cancer Network 2 WP2363	1	0.054	0.119
RAC1/PAK1/p38/MMP2 Pathway WP3303	2	0.061	0.125
Non-genomic actions of 1,25 dihydroxyvitamin D3 WP4341	2	0.061	0.125
IL17 signaling pathway WP2112	1	0.059	0.125
Resistin as a regulator of inflammation WP4481	1	0.06	0.125
Signal transduction through IL1R WP4496	1	0.06	0.125
p38 MAPK Signaling Pathway WP400	1	0.06	0.125
Fluoropyrimidine Activity WP1601	1	0.064	0.129
miRNAs involvement in the immune response in sepsis WP4329	1	0.064	0.129
Human Thyroid Stimulating Hormone (TSH) signaling pathway WP2032	2	0.065	0.129
Signaling of Hepatocyte Growth Factor Receptor WP313	1	0.067	0.130
Selenium Metabolism and Selenoproteins WP28	1	0.067	0.130
Fatty Acid Beta Oxidation WP143	1	0.069	0.133
Zinc homeostasis WP3529	1	0.072	0.137
Signaling Pathways in Glioblastoma WP2261	2	0.074	0.140
Apoptosis WP254	2	0.077	0.141
Nuclear Receptors in Lipid Metabolism and Toxicity WP299	1	0.076	0.141
Wnt Signaling in Kidney Disease WP4150	1	0.077	0.141
Striated Muscle Contraction Pathway WP383	1	0.083	0.149
Ferroptosis WP4313	1	0.083	0.149
ncRNAs involved in Wnt signaling in hepatocellular carcinoma WP4336	2	0.084	0.149
Nuclear Receptors WP170	1	0.09	0.158
TNF related weak inducer of apoptosis (TWEAK) Signaling Pathway WP2036	1	0.092	0.160
NO/cGMP/PKG mediated Neuroprotection WP4008	1	0.095	0.164
Circadian rhythm related genes WP3594	4	0.096	0.164
Pancreatic adenocarcinoma pathway WP4263	2	0.097	0.164
Androgen receptor signaling pathway WP138	2	0.102	0.171
Metabolic reprogramming in colon cancer WP4290	1	0.103	0.171
LncRNA involvement in canonical Wnt signaling and colorectal cancer WP4258	2	0.108	0.173
IL-5 Signaling Pathway WP127	1	0.107	0.173
Exercise-induced Circadian Regulation WP410	1	0.108	0.173
Histone Modifications WP2369	1	0.106	0.173
TNF alpha Signaling Pathway WP231	2	0.109	0.173
ATM Signaling Network in Development and Disease WP3878	1	0.112	0.177
Pathogenic Escherichia coli infection WP2272	1	0.114	0.178
IL-6 signaling pathway WP364	1	0.119	0.181
Regulation of Microtubule Cytoskeleton WP2038	1	0.119	0.181
Rett syndrome causing genes WP4312	1	0.119	0.181
Metapathway biotransformation Phase I and II WP702	3	0.12	0.181
Energy Metabolism WP1541	1	0.121	0.182
IL-2 Signaling Pathway WP49	1	0.125	0.186
G Protein Signaling Pathways WP35	2	0.131	0.193
Synaptic signaling pathways associated with autism spectrum disorder WP4539	1	0.133	0.195
Structural Pathway of Interleukin 1 (IL-1) WP2637	1	0.138	0.196
Copper homeostasis WP3286	1	0.136	0.196
Wnt Signaling Pathway WP363	1	0.137	0.196
TYROBP Causal Network WP3945	1	0.137	0.196
Epithelial to mesenchymal transition in colorectal cancer WP4239	3	0.142	0.200
Apoptosis-related network due to altered Notch3 in ovarian cancer WP2864	1	0.148	0.207
IL-4 Signaling Pathway WP395	1	0.152	0.211
MET in type 1 papillary renal cell carcinoma WP4205	1	0.158	0.216
Prader-Willi and Angelman Syndrome WP3998	1	0.159	0.216
Genotoxicity pathway WP4286	1	0.157	0.216
RANKL/RANK (Receptor activator of NFKB (ligand)) Signaling Pathway WP2018	1	0.163	0.218
RIG-I-like Receptor Signaling WP3865	1	0.162	0.218

Lung fibrosis WP3624	1	0.171	0.227
Insulin Signaling WP481	3	0.178	0.234
Kit receptor signaling pathway WP304	1	0.179	0.234
Hippo-Merlin Signaling Dysregulation WP4541	2	0.184	0.239
NRF2 pathway WP2884	2	0.196	0.250
Endometrial cancer WP4155	1	0.198	0.250
Pathways Affected in Adenoid Cystic Carcinoma WP3651	1	0.197	0.250
AGE/RAGE pathway WP2324	1	0.198	0.250
mRNA Processing WP411	2	0.216	0.271
Regulation of toll-like receptor signaling pathway WP1449	2	0.22	0.274
Leptin signaling pathway WP2034	1	0.225	0.278
Arrhythmogenic Right Ventricular Cardiomyopathy WP2118	1	0.23	0.282
Endoderm Differentiation WP2853	2	0.233	0.284
Mesodermal Commitment Pathway WP2857	2	0.25	0.303
Ectoderm Differentiation WP2858	2	0.265	0.317
Pyrimidine metabolism WP4022	1	0.265	0.317
Alzheimers Disease WP2059	1	0.283	0.336
Breast cancer pathway WP4262	2	0.291	0.344
Regulation of Actin Cytoskeleton WP51	2	0.297	0.348
Sudden Infant Death Syndrome (SIDS) Susceptibility Pathways WP706	2	0.299	0.349
ErbB Signaling Pathway WP673	1	0.305	0.353
Allograft Rejection WP2328	1	0.308	0.354
Focal Adhesion-PI3K-Akt-mTOR-signaling pathway WP3932	4	0.316	0.359
Nuclear Receptors Meta-Pathway WP2882	4	0.319	0.359
Myometrial Relaxation and Contraction Pathways WP289	2	0.317	0.359
Amino Acid metabolism WP3925	1	0.32	0.359
Corticotropin-releasing hormone signaling pathway WP2355	1	0.337	0.376
Genes related to primary cilium development (based on CRISPR) WP4536	1	0.347	0.385
Toll-like Receptor Signaling Pathway WP75	1	0.353	0.389
Wnt Signaling Pathway and Pluripotency WP399	1	0.358	0.392
VEGFA-VEGFR2 Signaling Pathway WP3888	3	0.368	0.396
Pathways Regulating Hippo Signaling WP4540	1	0.365	0.396
Senescence and Autophagy in Cancer WP615	1	0.366	0.396
Vitamin D Receptor Pathway WP2877	2	0.375	0.399
Neural Crest Differentiation WP2064	1	0.374	0.399
Ras Signaling WP4223	2	0.395	0.418
PI3K-Akt Signaling Pathway WP4172	4	0.403	0.421
Thermogenesis WP4321	1	0.402	0.421
Ciliary landscape WP4352	2	0.459	0.477
ESC Pluripotency Pathways WP3931	1	0.473	0.489
TGF-beta Signaling Pathway WP366	1	0.499	0.513
Angiopoietin Like Protein 8 Regulatory Pathway WP3915	1	0.51	0.521
Nonalcoholic fatty liver disease WP4396	1	0.518	0.527
Integrated Breast Cancer Pathway WP1984	1	0.57	0.576
GPCRs, Class A Rhodopsin-like WP455	2	0.578	0.581
Chemokine signaling pathway WP3929	1	0.607	0.607

Table C.2: Signalling pathways along with the adjusted p-values and the number of overlapping genes obtained after pathway enrichment analysis applied to the full scale analysis results.

Gene Set Name	# Genes in Gene Set (K)	Description	GiO	k/K	p-value	FDR
STK33_SKM_UP	275	Genes up-regulated in SKM-1 cells (AML) after knockdown of STK33 [Gene ID=65975] by RNAi.	12	0.0436	1.17E-07	1.62E-05
STK33_UP	285	Genes up-regulated in NOMO-1 and SKM-1 cells (AML) after knockdown of STK33 [Gene ID=65975] by RNAi.	12	0.0421	1.72E-07	1.62E-05
STK33_NOMO_UP	290	Genes up-regulated in NOMO-1 cells (AML) after knockdown of STK33 [Gene ID=65975] by RNAi.	10	0.0345	1.08E-05	0.000597
TGFB_UP.V1_DN	188	Genes down-regulated in a panel of epithelial cell lines by TGFB1 [Gene ID=7040].	8	0.0426	1.88E-05	0.000597
TGFB_UP.V1_UP	189	Genes up-regulated in a panel of epithelial cell lines by TGFB1 [Gene ID=7040].	8	0.0423	1.95E-05	0.000597
KRAS.DF.V1_UP	190	Genes up-regulated in epithelial lung cancer cell lines over-expressing an oncogenic form of KRAS [Gene ID=3845] gene.	8	0.0421	2.02E-05	0.000597
SIRNA_EIF4GI_UP	94	Genes up-regulated in MCF10A cells vs knockdown of EIF4G1 [Gene ID=1981] gene by RNAi.	6	0.0638	2.21E-05	0.000597
RAPA_EARLY_UP.V1_UP	166	Genes up-regulated in BJAB (lymphoma) cells by rapamycin (sirolimus) [PubChem = 6610346].	7	0.0422	6.64E-05	0.00157
P53_DN.V1_DN	193	Genes down-regulated in NCI-60 panel of cell lines with mutated TP53 [Gene ID=7157].	7	0.0363	0.00017	0.00321
RAF_UP.V1_UP	193	Genes up-regulated in MCF-7 cells positive for ESR1.	7	0.0363	0.00017	0.00321
KRAS.KIDNEY_UP.V1_UP	142	Genes up-regulated in epithelial kidney cancer cell lines over-expressing an oncogenic form of KRAS [Gene ID=3845] gene.	6	0.0423	0.000219	0.00377
KRAS.600_UP.V1_UP	276	Genes up-regulated in four lineages of epithelial cell lines over-expressing an oncogenic form of KRAS [Gene ID=3845] gene.	8	0.029	0.000272	0.00428
AKT_UP_MTOR_DN.V1_UP	178	Genes up-regulated by everolimus [PubChem = 6442177] in mouse prostate tissue transgenically expressing human AKT1.	6	0.0337	0.000729	0.0106
ESC_J1_UP_LATE.V1_UP	185	Genes up-regulated during late stages of differentiation of embryoid bodies from J1 embryonic stem cells.	6	0.0324	0.00089	0.0119
CYCLIN_D1_UP.V1_UP	187	Genes up-regulated in MCF-7 cells (breast cancer) over-expressing CCND1 [Gene ID=595] gene.	6	0.0321	0.000941	0.0119
EGFR_UP.V1_UP	192	Genes up-regulated in MCF-7 cells positive for ESR1 engineered to express ligand-activatable EGFR.	6	0.0312	0.00108	0.0119
RAF_UP.V1_DN	192	Genes down-regulated in MCF-7 cells (breast cancer) positive for ESR1 MCF-7 cells over-expressing active RAF1 gene.	6	0.0312	0.00108	0.0119
LEF1_UP.V1_UP	194	Genes up-regulated in DLD1 cells (colon carcinoma) over-expressing LEF1 [Gene ID=51176].	6	0.0309	0.00114	0.0119
PDGF_UP.V1_DN	134	Genes down-regulated in SH-SY5Y cells (neuroblastoma) in response to PDGF [Gene ID=] stimulation.	5	0.0373	0.00129	0.0128
PDGF_ERK_DN.V1_DN	148	Genes down-regulated in SH-SY5Y cells (neuroblastoma) in response to PDGF stimulation after pre-treatment with the ERK inhibitors.	5	0.0338	0.002	0.0189
ESC_V6.5_UP_EARLY.V1_DN	161	Genes down-regulated during early stages of differentiation of embryoid bodies from V6.5 embryonic stem cells.	5	0.0311	0.00288	0.0259
MTOR_UP.V1_DN	181	Genes down-regulated by everolimus [PubChem = 6442177] in prostate tissue.	5	0.0276	0.00473	0.0397
IL15_UP.V1_DN	182	Genes down-regulated in Sez-4 cells, first starved of IL2 and then stimulated with IL15.	5	0.0275	0.00484	0.0397
CYCLIN_D1_KE_.V1_UP	188	Genes up-regulated in MCF-7 cells (breast cancer) over-expressing a mutant K112E form of CCND1 [Gene ID=595] gene.	5	0.0266	0.00554	0.0397
PIGF_UP.V1_DN	188	Genes down-regulated in HUVEC cells (endothelium) by treatment with PIGF [Gene ID=5281].	5	0.0266	0.00554	0.0397
ATF2_S_UP.V1_UP	189	Genes up-regulated in myometrial cells over-expressing a shortened splice form of ATF2 [Gene ID=1386] gene.	5	0.0265	0.00567	0.0397
ERBB2_UP.V1_UP	189	Genes up-regulated in MCF-7 cells positive for ESR1 engineered to express ligand-activatable ERBB2.	5	0.0265	0.00567	0.0397
MEK_UP.V1_UP	194	Genes up-regulated in MCF-7 cells (breast cancer) positive for ESR1 [Gene ID=2099] MCF-7 cells over-expressing active MAP2K1 gene.	5	0.0258	0.00632	0.0426
RB_P130_DN.V1_UP	128	Genes up-regulated in primary keratinocytes from RB1 and RBL2 [Gene ID=5925, 5934] skin specific knockout mice.	4	0.0312	0.00739	0.0481

Table C.3: Overlaps have been detected using gene set enrichment analysis performed using a hypergeometric distribution. The false discovery rate analog of the hypergeometric p-value is displayed after correction for multiple hypothesis testing according to Benjamini and Hochberg. GiO stands for Genes in Overlap (k).

Pathway	NoG	P-value	Adj.P-value	Regul.
Axon guidance Homo sapiens R-HSA-422475	14	0.000	0.000	Up
Signaling by VEGF Homo sapiens R-HSA-194138	10	0.000	0.000	Up
CREB phosphorylation through the activation of Ras Homo sapiens R-HSA-442742	3	0.000	0.000	Down
Negative regulation of MAPK pathway Homo sapiens R-HSA-5675221	3	0.000	0.000	Down
TNF signaling Homo sapiens R-HSA-75893	3	0.000	0.000	Up
IRAK1 recruits IKK complex Homo sapiens R-HSA-937039	2	0.000	0.000	Down
IRAK1 recruits IKK complex upon TLR7/8 or 9 stimulation Homo sapiens R-HSA-975144	2	0.000	0.000	Down
TP53 Regulates Transcription of Death Receptors and Ligands Homo sapiens R-HSA-6803211	2	0.000	0.000	Down
DCC mediated attractive signaling Homo sapiens R-HSA-418885	2	0.000	0.000	Down
Ras activation uopn Ca2+ infux through NMDA receptor Homo sapiens R-HSA-442982	2	0.000	0.000	Down
RIP-mediated NFkB activation via ZBP1 Homo sapiens R-HSA-1810476	2	0.000	0.000	Down
Sodium-coupled sulphate, di- and tri-carboxylate transporters Homo sapiens R-HSA-433137	1	0.000	0.000	Down
Homo sapiens R-HSA-5339717	1	0.000	0.000	Down
Hemostasis Homo sapiens R-HSA-109582	14	0.001	0.018	Up
VEGFA-VEGFR2 Pathway Homo sapiens R-HSA-4420097	9	0.001	0.018	Up
RAF-independent MAPK1/3 activation Homo sapiens R-HSA-112409	3	0.001	0.018	Down
Lipoprotein metabolism Homo sapiens R-HSA-174824	3	0.001	0.018	Up
Post NMDA receptor activation events Homo sapiens R-HSA-438064	3	0.001	0.018	Down
Formation of Fibrin Clot (Clotting Cascade) Homo sapiens R-HSA-140877	3	0.001	0.018	Up
Death Receptor Signalling Homo sapiens R-HSA-73887	3	0.001	0.018	Up
N-Glycan antennae elongation Homo sapiens R-HSA-975577	2	0.001	0.018	Down
Intrinsic Pathway of Fibrin Clot Formation Homo sapiens R-HSA-140837	2	0.001	0.018	Down
N-glycan antennae elongation in the medial/trans-Golgi Homo sapiens R-HSA-975576	2	0.001	0.018	Down
RSK activation Homo sapiens R-HSA-444257	1	0.001	0.018	Down
NrCAM interactions Homo sapiens R-HSA-447038	1	0.001	0.018	Down
Activation of PUMA and translocation to mitochondria Homo sapiens R-HSA-139915	1	0.001	0.018	Down
Developmental Biology Homo sapiens R-HSA-1266738	17	0.002	0.027	Up
Interleukin-1 signaling Homo sapiens R-HSA-446652	3	0.002	0.027	Up
Caspase-mediated cleavage of cytoskeletal proteins Homo sapiens R-HSA-264870	2	0.002	0.027	Down
HDL-mediated lipid transport Homo sapiens R-HSA-194223	2	0.002	0.027	Down
ZBP1(DAI) mediated induction of type I IFNs Homo sapiens R-HSA-1606322	2	0.002	0.027	Down
Neurofascin interactions Homo sapiens R-HSA-447043	1	0.002	0.027	Down
IkBA variant leads to EDA-ID Homo sapiens R-HSA-5603029	1	0.002	0.027	Down
IKBKG deficiency causes EDA-ID via TLR Homo sapiens R-HSA-5603027	1	0.002	0.027	Down
Synthesis of (16-20)-hydroxyeicosatetraenoic acids (HETE) Homo sapiens R-HSA-2142816	1	0.002	0.027	Down
Homo sapiens R-HSA-442755	3	0.003	0.033	Up
Common Pathway of Fibrin Clot Formation Homo sapiens R-HSA-140875	2	0.003	0.033	Down
PTK6 Regulates Proteins Involved in RNA Processing Homo sapiens R-HSA-8849468	1	0.003	0.033	Down
Scavenging by Class F Receptors Homo sapiens R-HSA-3000484	1	0.003	0.033	Down
IRF3 mediated activation of type 1 IFN Homo sapiens R-HSA-1606341	1	0.003	0.033	Down
TNFR1-mediated ceramide production Homo sapiens R-HSA-5626978	1	0.003	0.033	Down
Homo sapiens R-HSA-2660825	1	0.003	0.033	Down
Constitutive Signaling by NOTCH1; Homo sapiens R-HSA-2660826	1	0.003	0.033	Down
TNFR1-induced NFkappaB signaling pathway Homo sapiens R-HSA-5357956	2	0.004	0.041	Down
Axonal growth inhibition (RHOA activation) Homo sapiens R-HSA-193634	1	0.004	0.041	Down
Glucocorticoid biosynthesis Homo sapiens R-HSA-194002	1	0.004	0.041	Down
ERKs are inactivated Homo sapiens R-HSA-202670	1	0.004	0.041	Down
Amino acid synthesis and interconversion (transamination) Homo sapiens R-HSA-70614	2	0.005	0.042	Down
Regulation of TNFR1 signaling Homo sapiens R-HSA-5357905	2	0.005	0.042	Down
Activation of Matrix Metalloproteinases Homo sapiens R-HSA-1592389	2	0.005	0.042	Down
Homo sapiens R-HSA-5368598	1	0.005	0.042	Down
Protein repair Homo sapiens R-HSA-5676934	1	0.005	0.042	Down
Phase 3 - rapid repolarisation Homo sapiens R-HSA-5576890	1	0.005	0.042	Down
Homo sapiens R-HSA-446388	1	0.005	0.042	Down
Relaxin receptors Homo sapiens R-HSA-444821	1	0.005	0.042	Down
p75NTR regulates axonogenesis Homo sapiens R-HSA-193697	1	0.005	0.042	Down
Dissolution of Fibrin Clot Homo sapiens R-HSA-75205	1	0.005	0.042	Down

VEGFR2 mediated cell proliferation Homo sapiens R-HSA-5218921	7	0.006	0.043	Up
NOD1/2 Signaling Pathway Homo sapiens R-HSA-168638	2	0.006	0.043	Down
R-HSA-5340588	1	0.006	0.043	Down
CHL1 interactions Homo sapiens R-HSA-447041	1	0.006	0.043	Down
Homo sapiens R-HSA-2892245	1	0.006	0.043	Down
Synthesis of IP2, IP, and Ins in the cytosol Homo sapiens R-HSA-1855183	1	0.006	0.043	Down
Signaling by NOTCH3 Homo sapiens R-HSA-1980148	1	0.006	0.043	Down
Signaling by NOTCH4 Homo sapiens R-HSA-1980150	1	0.006	0.043	Down
Miscellaneous substrates Homo sapiens R-HSA-211958	1	0.006	0.043	Down
Eicosanoids Homo sapiens R-HSA-211979	1	0.006	0.043	Down
Trafficking of AMPA receptors Homo sapiens R-HSA-399719	2	0.007	0.044	Down
Homo sapiens R-HSA-399721	2	0.007	0.044	Down
Sialic acid metabolism Homo sapiens R-HSA-4085001	2	0.007	0.044	Down
Regulation of TP53 Degradation Homo sapiens R-HSA-6804757	2	0.007	0.044	Down
Vitamin D (calciferol) metabolism Homo sapiens R-HSA-196791	1	0.007	0.044	Down
Pyrimidine salvage reactions Homo sapiens R-HSA-73614	1	0.007	0.044	Down
NF-kB activation; Homo sapiens R-HSA-933543	1	0.007	0.044	Down
Elevation of cytosolic Ca2+ levels Homo sapiens R-HSA-139853	1	0.007	0.044	Down
Homo Sapiens R-HSA-112314	5	0.008	0.047	Up
Degradation of the extracellular matrix Homo sapiens R-HSA-1474228	4	0.008	0.047	Up
Regulation of TP53 Expression and Degradation Homo sapiens R-HSA-6806003	2	0.008	0.047	Down
Androgen biosynthesis Homo sapiens R-HSA-193048	1	0.008	0.047	Down
Homo sapiens R-HSA-6804759	1	0.008	0.047	Down
Glutathione synthesis and recycling Homo sapiens R-HSA-174403	1	0.008	0.047	Down
Steroid hormones Homo sapiens R-HSA-209943	2	0.009	0.050	Down
Heme biosynthesis Homo sapiens R-HSA-189451	1	0.009	0.050	Down
p75NTR recruits signalling complexes Homo sapiens R-HSA-209543	1	0.009	0.050	Down
Bile salt and organic anion SLC transporters Homo sapiens R-HSA-425471	1	0.009	0.050	Down
Chylomicron-mediated lipid transport Homo sapiens R-HSA-174800	1	0.010	0.055	Down
Signaling by Activin Homo sapiens R-HSA-1502540	1	0.011	0.058	Down
Homo sapiens R-HSA-156988	1	0.011	0.058	Down
Homo sapiens R-HSA-442729	1	0.011	0.058	Down
Fatty acids Homo sapiens R-HSA-211935	1	0.011	0.058	Down
Apoptotic cleavage of cellular proteins Homo sapiens R-HSA-111465	2	0.012	0.060	Down
Pyrimidine catabolism Homo sapiens R-HSA-73621	1	0.012	0.060	Down
Platelet Adhesion to exposed collagen Homo sapiens R-HSA-75892	1	0.012	0.060	Down
Homo sapiens R-HSA-3772470	1	0.012	0.060	Down
Metabolism of porphyrins Homo sapiens R-HSA-189445	1	0.012	0.060	Down
Lipid digestion, mobilization, and transport Homo sapiens R-HSA-73923	3	0.013	0.063	Up
Netrin-1 signaling Homo sapiens R-HSA-373752	2	0.013	0.063	Down
NF-kB is activated and signals survival Homo sapiens R-HSA-209560	1	0.013	0.063	Down
TNFR1-induced proapoptotic signaling Homo sapiens R-HSA-5357786	1	0.013	0.063	Down
NLR signaling pathways; Homo sapiens R-HSA-168643	2	0.014	0.064	Down
Uptake and function of anthrax toxins Homo sapiens R-HSA-5210891	1	0.014	0.064	Down
IRF3-mediated induction of type I IFN Homo sapiens R-HSA-3270619	1	0.014	0.064	Down
TP53 regulating transcription of cell death genes; Homo sapiens R-HSA-6803205	1	0.014	0.064	Down
Polo-like kinase mediated events Homo sapiens R-HSA-156711	1	0.014	0.064	Down
NCAM signaling for neurite out-growth Homo sapiens R-HSA-375165	7	0.015	0.066	Up
Defective EXT2 causes exostoses 2 Homo sapiens R-HSA-3656237	1	0.015	0.066	Down
Defective EXT1 causes exostoses 1, TRPS2 and CHDS Homo sapiens R-HSA-3656253	1	0.015	0.066	Down
Constitutive Signaling by NOTCH1 HD Domain Mutants Homo sapiens R-HSA-2691232	1	0.015	0.066	Down
Signaling by NOTCH1 HD Domain Mutants in Cancer Homo sapiens R-HSA-2691230	1	0.015	0.066	Down
Intraflagellar transport Homo sapiens R-HSA-5620924	2	0.016	0.068	Down
MAP3K8 (TPL2)-dependent MAPK1/3 activation Homo sapiens R-HSA-5684264	1	0.016	0.068	Down
Homo sapiens R-HSA-6803204	1	0.016	0.068	Down
Retrograde transport at the Trans-Golgi-Network Homo sapiens R-HSA-6811440	2	0.017	0.070	Down
p75NTR signals via NF-kB Homo sapiens R-HSA-193639	1	0.017	0.070	Down
Homo sapiens R-HSA-438066	1	0.017	0.070	Down

TP53 Regulates Transcription of Cell Death Genes Homo sapiens R-HSA-5633008	2	0.018	0.071	Down
Ion homeostasis Homo sapiens R-HSA-5578775	2	0.018	0.071	Down
Defects in cobalamin (B12) metabolism Homo sapiens R-HSA-3296469	1	0.018	0.071	Down
Cell-extracellular matrix interactions Homo sapiens R-HSA-446353	1	0.018	0.071	Down
Synthesis of glycosylphosphatidylinositol (GPI) Homo sapiens R-HSA-162710	1	0.018	0.071	Down
Aflatoxin activation and detoxification Homo sapiens R-HSA-5423646	1	0.018	0.071	Down
Kinesins Homo sapiens R-HSA-983189	2	0.020	0.076	Down
Apoptotic execution phase Homo sapiens R-HSA-75153	2	0.020	0.076	Down
RHO GTPases Activate ROCKs Homo sapiens R-HSA-5627117	1	0.020	0.076	Down
Signaling by NODAL Homo sapiens R-HSA-1181150	1	0.020	0.076	Down
Signaling by Interleukins Homo sapiens R-HSA-449147	9	0.021	0.077	Up
IRS-mediated signalling Homo sapiens R-HSA-112399	7	0.021	0.077	Up
Insulin receptor signalling cascade Homo sapiens R-HSA-74751	7	0.022	0.077	Up
IGF1R signaling cascade Homo sapiens R-HSA-2428924	7	0.022	0.077	Up
Homo sapiens R-HSA-2404192	7	0.022	0.077	Up
IRS-related events triggered by IGF1R Homo sapiens R-HSA-2428928	7	0.022	0.077	Up
Toll Like Receptor 10 (TLR10) Cascade Homo sapiens R-HSA-168142	3	0.023	0.077	Up
Toll Like Receptor 5 (TLR5) Cascade Homo sapiens R-HSA-168176	3	0.023	0.077	Up
MyD88 cascade initiated on plasma membrane Homo sapiens R-HSA-975871	3	0.023	0.077	Up
Homo sapiens R-HSA-975138	3	0.023	0.077	Up
MyD88 dependent cascade initiated on endosome Homo sapiens R-HSA-975155	3	0.024	0.077	Up
Toll Like Receptor 7/8 (TLR7/8) Cascade Homo sapiens R-HSA-168181	3	0.024	0.077	Up
Phase 1 - Functionalization of compounds Homo sapiens R-HSA-211945	3	0.022	0.077	Up
Voltage gated Potassium channels Homo sapiens R-HSA-1296072	2	0.024	0.077	Down
Metabolism of fat-soluble vitamins Homo sapiens R-HSA-6806667	2	0.024	0.077	Down
Regulation of innate immune responses to cytosolic DNA Homo sapiens R-HSA-3134975	1	0.021	0.077	Down
STING mediated induction of host immune responses Homo sapiens R-HSA-1834941	1	0.024	0.077	Down
Other semaphorin interactions Homo sapiens R-HSA-416700	1	0.024	0.077	Down
Ephrin signaling Homo sapiens R-HSA-3928664	1	0.024	0.077	Down
Gap junction assembly Homo sapiens R-HSA-190861	1	0.022	0.077	Down
Synthesis of Leukotrienes (LT) and Eoxins (EX) Homo sapiens R-HSA-2142691	1	0.023	0.077	Down
Homo sapiens R-HSA-2979096	1	0.021	0.077	Down
ERK/MAPK targets Homo sapiens R-HSA-198753	1	0.024	0.077	Down
Defects in vitamin and cofactor metabolism Homo sapiens R-HSA-3296482	1	0.025	0.079	Down
Tie2 Signaling Homo sapiens R-HSA-210993	1	0.025	0.079	Down
Signaling by EGFR Homo sapiens R-HSA-177929	8	0.027	0.081	Up
GRB2 events in EGFR signaling Homo sapiens R-HSA-179812	6	0.029	0.081	Up
SHC1 events in EGFR signaling Homo sapiens R-HSA-180336	6	0.029	0.081	Up
SOS-mediated signalling Homo sapiens R-HSA-112412	6	0.029	0.081	Up
SHC1 events in ERBB4 signaling Homo sapiens R-HSA-1250347	6	0.029	0.081	Up
RAF/MAP kinase cascade Homo sapiens R-HSA-5673001	6	0.029	0.081	Up
FRS-mediated FGFR2 signaling Homo sapiens R-HSA-5654700	6	0.029	0.081	Up
FRS-mediated FGFR4 signaling Homo sapiens R-HSA-5654712	6	0.029	0.081	Up
FRS-mediated FGFR3 signaling Homo sapiens R-HSA-5654706	6	0.029	0.081	Up
FRS-mediated FGFR1 signaling Homo sapiens R-HSA-5654693	6	0.029	0.081	Up
Toll Like Receptor 9 (TLR9) Cascade Homo sapiens R-HSA-168138	3	0.029	0.081	Up
Defective B3GALT6 causes EDSP2 and SEMDJL1 Homo sapiens R-HSA-4420332	1	0.028	0.081	Down
Defective B4GALT7 causes EDS, progeroid type Homo sapiens R-HSA-3560783	1	0.028	0.081	Down
Defective B3GAT3 causes JDSSDHD Homo sapiens R-HSA-3560801	1	0.028	0.081	Down
Chondroitin sulfate biosynthesis Homo sapiens R-HSA-2022870	1	0.028	0.081	Down
TP53 regulating transcription of cell cycle genes; Homo sapiens R-HSA-6804115	1	0.028	0.081	Down
CTLA4 inhibitory signaling Homo sapiens R-HSA-389513	1	0.029	0.081	Down
Endogenous sterols Homo sapiens R-HSA-211976	1	0.028	0.081	Down
Homo sapiens R-HSA-198725	1	0.028	0.081	Down
IKK complex recruitment mediated by RIP1 Homo sapiens R-HSA-937041	1	0.027	0.081	Down
G0 and Early G1 Homo sapiens R-HSA-1538133	1	0.027	0.081	Down
Fc epsilon receptor (FCERI) signaling Homo sapiens R-HSA-2454202	8	0.031	0.082	Up
ARMS-mediated activation Homo sapiens R-HSA-170984	6	0.032	0.082	Up

Signalling to p38 via RIT and RIN Homo sapiens R-HSA-187706	6	0.034	0.082	Up
Frs2-mediated activation Homo sapiens R-HSA-170968	6	0.034	0.082	Up
MAPK1/MAPK3 signaling Homo sapiens R-HSA-5684996	6	0.033	0.082	Up
Prolonged ERK activation events Homo sapiens R-HSA-169893	6	0.035	0.082	Up
Signaling by Leptin Homo sapiens R-HSA-2586552	6	0.032	0.082	Up
Interleukin receptor SHC signaling Homo sapiens R-HSA-912526	6	0.034	0.082	Up
Signalling to RAS Homo sapiens R-HSA-167044	6	0.035	0.082	Up
Interleukin-2 signaling Homo sapiens R-HSA-451927	6	0.035	0.082	Up
MyD88:Mal cascade initiated on plasma membrane Homo sapiens R-HSA-166058	3	0.035	0.082	Up
Toll Like Receptor TLR1:TLR2 Cascade Homo sapiens R-HSA-168179	3	0.035	0.082	Up
Toll Like Receptor TLR6:TLR2 Cascade Homo sapiens R-HSA-168188	3	0.035	0.082	Up
Toll Like Receptor 2 (TLR2) Cascade Homo sapiens R-HSA-181438	3	0.035	0.082	Up
L1CAM interactions Homo sapiens R-HSA-373760	3	0.030	0.082	Up
Cell surface interactions at the vascular wall Homo sapiens R-HSA-202733	3	0.034	0.082	Up
Ion transport by P-type ATPases Homo sapiens R-HSA-936837	2	0.032	0.082	Down
Synthesis of substrates in N-glycan biosythesis Homo sapiens R-HSA-446219	2	0.034	0.082	Down
Signaling by Hippo Homo sapiens R-HSA-2028269	1	0.033	0.082	Down
Regulation of FZD by ubiquitination Homo sapiens R-HSA-4641263	1	0.035	0.082	Down
HS-GAG degradation Homo sapiens R-HSA-2024096	1	0.033	0.082	Down
Synthesis of very long-chain fatty acyl-CoAs Homo sapiens R-HSA-75876	1	0.033	0.082	Down
CD28 dependent PI3K/Akt signaling Homo sapiens R-HSA-389357	1	0.035	0.082	Down
Laminin interactions Homo sapiens R-HSA-3000157	1	0.031	0.082	Down
Insulin processing Homo sapiens R-HSA-264876	1	0.031	0.082	Down
Diseases associated with the TLR signaling cascade Homo sapiens R-HSA-5602358	1	0.033	0.082	Down
Diseases of Immune System Homo sapiens R-HSA-5260271	1	0.033	0.082	Down
TRAF6 mediated NF-kB activation Homo sapiens R-HSA-933542	1	0.035	0.082	Down
Homo sapiens R-HSA-163125	1	0.031	0.082	Down
Synthesis of IP3 and IP4 in the cytosol Homo sapiens R-HSA-1855204	1	0.030	0.082	Down
Homo sapiens R-HSA-499943	1	0.034	0.082	Down
Platelet calcium homeostasis Homo sapiens R-HSA-418360	1	0.034	0.082	Down
Homo sapiens R-HSA-416572	1	0.037	0.086	Down
Basigin interactions Homo sapiens R-HSA-210991	1	0.037	0.086	Down
Integrin alphaIIb beta3 signaling Homo sapiens R-HSA-354192	1	0.038	0.087	Down
G-protein activation Homo sapiens R-HSA-202040	1	0.038	0.087	Down
Signaling by Insulin receptor Homo sapiens R-HSA-74752	7	0.039	0.088	Up
Potassium Channels Homo sapiens R-HSA-1296071	3	0.039	0.088	Up
Homo sapiens R-HSA-211897	2	0.039	0.088	Down
Signaling by PDGF Homo sapiens R-HSA-186797	8	0.040	0.089	Up
Signaling by SCF-KIT Homo sapiens R-HSA-1433557	7	0.040	0.089	Up
FCERI mediated MAPK activation Homo sapiens R-HSA-2871796	6	0.040	0.089	Up
Chondroitin sulfate/dermatan sulfate metabolism Homo sapiens R-HSA-1793185	2	0.040	0.089	Down
Gap junction trafficking Homo sapiens R-HSA-190828	1	0.040	0.089	Down
Neuronal System Homo sapiens R-HSA-112316	7	0.041	0.090	Up
Constitutive Signaling by AKT1 E17K in Cancer Homo sapiens R-HSA-5674400	1	0.041	0.090	Down
VEGFR2 mediated vascular permeability Homo sapiens R-HSA-5218920	1	0.041	0.090	Down
Signaling by ERBB4 Homo sapiens R-HSA-1236394	7	0.043	0.090	Up
Signalling to ERKs Homo sapiens R-HSA-187687	6	0.042	0.090	Up
Interleukin-3, 5 and GM-CSF signaling Homo sapiens R-HSA-512988	6	0.043	0.090	Up
Platelet degranulation Homo sapiens R-HSA-114608	3	0.043	0.090	Up
Integrin cell surface interactions Homo sapiens R-HSA-216083	2	0.043	0.090	Down
Signaling by PTK6 Homo sapiens R-HSA-8848021	2	0.043	0.090	Down
Pyrimidine metabolism Homo sapiens R-HSA-73848	1	0.042	0.090	Down
Diseases associated with glycosaminoglycan metabolism Homo sapiens R-HSA-3560782	1	0.043	0.090	Down
Homo sapiens R-HSA-445989	1	0.042	0.090	Down
Metabolism of steroid hormones Homo sapiens R-HSA-196071	1	0.042	0.090	Down
Regulation of actin dynamics for phagocytic cup formation Homo sapiens R-HSA-2029482	2	0.044	0.090	Down
Cobalamin (Cbl, vitamin B12) transport and metabolism Homo sapiens R-HSA-196741	1	0.044	0.090	Down
Homo sapiens R-HSA-8849471	1	0.044	0.090	Down

Gap junction trafficking and regulation Homo sapiens R-HSA-157858	1	0.044	0.090	Down
Glycolysis Homo sapiens R-HSA-70171	1	0.044	0.090	Down
Downstream signaling of activated FGFR2 Homo sapiens R-HSA-5654696	7	0.047	0.091	Up
Downstream signaling of activated FGFR4 Homo sapiens R-HSA-5654716	7	0.047	0.091	Up
Downstream signaling of activated FGFR3 Homo sapiens R-HSA-5654708	7	0.047	0.091	Up
MAP kinase activation in TLR cascade Homo sapiens R-HSA-450294	2	0.045	0.091	Down
Semaphorin interactions Homo sapiens R-HSA-373755	2	0.046	0.091	Down
Activation of G protein gated Potassium channels Homo sapiens R-HSA-1296041	1	0.046	0.091	Down
Homo sapiens R-HSA-997272	1	0.046	0.091	Down
G protein gated Potassium channels Homo sapiens R-HSA-1296059	1	0.046	0.091	Down
Phase 2 - plateau phase Homo sapiens R-HSA-5576893	1	0.045	0.091	Down
Sema4D in semaphorin signaling Homo sapiens R-HSA-400685	1	0.047	0.091	Down
EGFR downregulation Homo sapiens R-HSA-182971	1	0.047	0.091	Down
MAPK targets/ Nuclear events mediated by MAP kinases Homo sapiens R-HSA-450282	1	0.047	0.091	Down
Glutathione conjugation Homo sapiens R-HSA-156590	1	0.046	0.091	Down
Activated NOTCH1 Transmits Signal to the Nucleus Homo sapiens R-HSA-2122948	1	0.048	0.092	Down
Cargo concentration in the ER Homo sapiens R-HSA-5694530	1	0.048	0.092	Down
Transmission across Chemical Synapses Homo sapiens R-HSA-112315	5	0.049	0.093	Up
Activation of BH3-only proteins Homo sapiens R-HSA-114452	1	0.049	0.093	Down
Signaling by FGFR4 Homo sapiens R-HSA-5654743	7	0.050	0.094	Up
Response to elevated platelet cytosolic Ca2+ Homo sapiens R-HSA-76005	3	0.050	0.094	Up
Signalling by NGF Homo sapiens R-HSA-166520	9	0.052	0.097	Up
Signaling by FGFR3 Homo sapiens R-HSA-5654741	7	0.052	0.097	Up
Intra-Golgi and retrograde Golgi-to-ER traffic Homo sapiens R-HSA-6811442	4	0.052	0.097	Up
FCERI mediated NF-kB activation Homo sapiens R-HSA-2871837	2	0.052	0.097	Down
Downstream signaling of activated FGFR1 Homo sapiens R-HSA-5654687	7	0.053	0.097	Up
Homo sapiens R-HSA-4641262	1	0.053	0.097	Down
Signaling by NOTCH2 Homo sapiens R-HSA-1980145	1	0.053	0.097	Down
Gluconeogenesis Homo sapiens R-HSA-70263	1	0.053	0.097	Down
Costimulation by the CD28 family Homo sapiens R-HSA-388841	2	0.054	0.098	Down
HS-GAG biosynthesis Homo sapiens R-HSA-2022928	1	0.054	0.098	Down
Homo sapiens R-HSA-1971475	1	0.055	0.099	Down
Adherens junctions interactions Homo sapiens R-HSA-418990	1	0.055	0.099	Down
Signaling by FGFR1 Homo sapiens R-HSA-5654736	7	0.056	0.101	Up
DAP12 signaling Homo sapiens R-HSA-2424491	7	0.057	0.101	Up
Activated TLR4 signalling Homo sapiens R-HSA-166054	3	0.057	0.101	Up
HSF1-dependent transactivation Homo sapiens R-HSA-3371571	1	0.057	0.101	Down
Downstream signal transduction Homo sapiens R-HSA-186763	7	0.058	0.102	Up
Cytosolic sensors of pathogen-associated DNA Homo sapiens R-HSA-1834949	2	0.058	0.102	Down
Uptake and actions of bacterial toxins Homo sapiens R-HSA-5339562	1	0.058	0.102	Down
Biosynthesis of the N-glycan precursor ; Homo sapiens R-HSA-446193	2	0.060	0.105	Down
Fanconi Anemia Pathway Homo sapiens R-HSA-6783310	1	0.060	0.105	Down
Biological oxidations Homo sapiens R-HSA-211859	4	0.061	0.106	Up
MAPK family signaling cascades Homo sapiens R-HSA-5683057	6	0.062	0.107	Up
Signaling by FGFR2 Homo sapiens R-HSA-5654738	7	0.063	0.108	Up
Signaling by WNT in cancer Homo sapiens R-HSA-4791275	1	0.063	0.108	Down
DAP12 interactions Homo sapiens R-HSA-2172127	7	0.064	0.109	Up
CD28 co-stimulation Homo sapiens R-HSA-389356	1	0.064	0.109	Down
TRAF6 Mediated Induction of proinflammatory cytokines Homo sapiens R-HSA-168180	2	0.065	0.111	Down
NCAM1 interactions Homo sapiens R-HSA-419037	1	0.066	0.112	Down
Striated Muscle Contraction Homo sapiens R-HSA-390522	1	0.068	0.115	Down
Inwardly rectifying K+ channels Homo sapiens R-HSA-1296065	1	0.070	0.117	Down
Antigen activating B Cell Receptor;Homo sapiens R-HSA-983695	1	0.070	0.117	Down
Signaling by FGFR Homo sapiens R-HSA-190236	7	0.071	0.117	Up
Toll Like Receptor 4 (TLR4) Cascade Homo sapiens R-HSA-166016	3	0.071	0.117	Up
COPI-dependent Golgi-to-ER retrograde traffic Homo sapiens R-HSA-6811434	2	0.071	0.117	Down
RHO GTPases Activate WASPs and WAVEs Homo sapiens R-HSA-5663213	1	0.071	0.117	Down
Platelet Aggregation (Plug Formation) Homo sapiens R-HSA-76009	1	0.072	0.119	Down

Innate Immune System Homo sapiens R-HSA-168249	13	0.078	0.128	Up
Homo sapiens R-HSA-983231	3	0.079	0.129	Up
Synthesis of PIPs at the plasma membrane Homo sapiens R-HSA-1660499	1	0.081	0.132	Down
Collagen degradation Homo sapiens R-HSA-1442490	1	0.081	0.132	Down
Intrinsic Pathway for Apoptosis Homo sapiens R-HSA-109606	1	0.083	0.134	Down
Meiotic recombination Homo sapiens R-HSA-912446	1	0.083	0.134	Down
Fcgamma receptor (FCGR) dependent phagocytosis Homo sapiens R-HSA-2029480	2	0.084	0.135	Down
Homo sapiens R-HSA-390471	1	0.084	0.135	Down
Homo sapiens R-HSA-6814122	1	0.085	0.136	Down
Cardiac conduction Homo sapiens R-HSA-5576891	3	0.086	0.137	Up
Transcriptional regulation of pluripotent stem cells Homo sapiens R-HSA-452723	1	0.087	0.138	Down
mRNA Splicing - Major Pathway Homo sapiens R-HSA-72163	3	0.088	0.139	Up
NGF signalling via TRKA from the plasma membrane Homo sapiens R-HSA-187037	7	0.089	0.140	Up
Non-integrin membrane-ECM interactions Homo sapiens R-HSA-3000171	1	0.090	0.141	Down
Immune System Homo sapiens R-HSA-168256	23	0.094	0.146	Up
Downstream TCR signaling Homo sapiens R-HSA-202424	2	0.094	0.146	Down
EPHB-mediated forward signaling Homo sapiens R-HSA-3928662	1	0.094	0.146	Down
Retinoid metabolism and transport Homo sapiens R-HSA-975634	1	0.094	0.146	Down
p75 NTR receptor-mediated signalling Homo sapiens R-HSA-193704	2	0.095	0.146	Up
Metabolism of nucleotides Homo sapiens R-HSA-15869	2	0.095	0.146	Up
GABA B receptor activation Homo sapiens R-HSA-977444	1	0.096	0.146	Down
Activation of GABAB receptors Homo sapiens R-HSA-991365	1	0.096	0.146	Down
Fatty Acyl-CoA Biosynthesis Homo sapiens R-HSA-75105	1	0.096	0.146	Down
Cell junction organization Homo sapiens R-HSA-446728	2	0.097	0.148	Down
mRNA Splicing Homo sapiens R-HSA-72172	3	0.099	0.150	Up
Inositol phosphate metabolism Homo sapiens R-HSA-1483249	1	0.101	0.152	Down
RHO GTPases activate PKNs Homo sapiens R-HSA-5625740	1	0.101	0.152	Down
Muscle contraction Homo sapiens R-HSA-397014	4	0.102	0.153	Up
Toll-Like Receptors Cascades Homo sapiens R-HSA-168898	3	0.105	0.157	Up
CLEC7A (Dectin-1) signaling Homo sapiens R-HSA-5607764	2	0.105	0.157	Up
Interferon gamma signaling Homo sapiens R-HSA-877300	2	0.106	0.157	Up
Transcriptional activation of mitochondrial biogenesis Homo sapiens R-HSA-2151201	1	0.106	0.157	Down
Binding and Uptake of Ligands by Scavenger Receptors Homo sapiens R-HSA-2173782	1	0.106	0.157	Down
Glucagon-like Peptide-1 (GLP1) regulates insulin secretion Homo sapiens R-HSA-381676	1	0.107	0.158	Down
Chaperonin-mediated protein folding Homo sapiens R-HSA-390466	2	0.108	0.159	Up
EPH-Ephrin signaling Homo sapiens R-HSA-2682334	2	0.109	0.160	Up
PLC beta mediated events Homo sapiens R-HSA-112043	1	0.110	0.160	Down
Diseases of metabolism Homo sapiens R-HSA-5668914	1	0.110	0.160	Down
G-protein mediated events Homo sapiens R-HSA-112040	1	0.112	0.163	Down
Phase 0 - rapid depolarisation Homo sapiens R-HSA-5576892	1	0.114	0.165	Down
mRNA Splicing - Minor Pathway Homo sapiens R-HSA-72165	1	0.115	0.166	Down
Transport to the Golgi and subsequent modification Homo sapiens R-HSA-948021	3	0.119	0.171	Up
Apoptosis Homo sapiens R-HSA-109581	3	0.120	0.171	Up
MyD88-independent TLR3/TLR4 cascade Homo sapiens R-HSA-166166	2	0.121	0.171	Up
Toll Like Receptor 3 (TLR3) Cascade Homo sapiens R-HSA-168164	2	0.121	0.171	Up
TRIF-mediated TLR3/TLR4 signaling Homo sapiens R-HSA-937061	2	0.121	0.171	Up
G alpha (z) signalling events Homo sapiens R-HSA-418597	1	0.120	0.171	Down
Programmed Cell Death Homo sapiens R-HSA-5357801	3	0.124	0.175	Up
Regulation of TP53 Activity Homo sapiens R-HSA-5633007	3	0.126	0.177	Up
Protein folding Homo sapiens R-HSA-391251	2	0.126	0.177	Up
Cytokine Signaling in Immune system Homo sapiens R-HSA-1280215	10	0.127	0.178	Up
TP53 Regulates Transcription of Cell Cycle Genes Homo sapiens R-HSA-6791312	1	0.128	0.179	Down
Golgi-to-ER retrograde transport Homo sapiens R-HSA-8856688	2	0.132	0.183	Up
GPVI-mediated activation cascade Homo sapiens R-HSA-114604	1	0.132	0.183	Down
G beta:gamma signalling through PI3Kgamma Homo sapiens R-HSA-392451	1	0.133	0.184	Down
mRNA 3'-end processing Homo sapiens R-HSA-72187	1	0.138	0.189	Down
Post-Elongation Processing of Intron-Containing pre-mRNA Homo sapiens R-HSA-112296	1	0.138	0.189	Down
Nuclear Receptor transcription pathway Homo sapiens R-HSA-383280	1	0.138	0.189	Down

Heparan sulfate/heparin (HS-GAG) metabolism Homo sapiens R-HSA-1638091	1	0.142	0.194	Down
TCR signaling Homo sapiens R-HSA-202403	2	0.147	0.200	Up
Mitochondrial biogenesis Homo sapiens R-HSA-1592230	1	0.147	0.200	Down
Amyloid fiber formation Homo sapiens R-HSA-977225	1	0.149	0.202	Down
G-protein beta:gamma signalling Homo sapiens R-HSA-397795	1	0.150	0.203	Down
Golgi Associated Vesicle Biogenesis Homo sapiens R-HSA-432722	1	0.156	0.210	Down
C-type lectin receptors (CLRs) Homo sapiens R-HSA-5621481	2	0.158	0.212	Up
GABA receptor activation Homo sapiens R-HSA-977443	1	0.165	0.221	Down
Activation of NF-kappaB in B cells Homo sapiens R-HSA-1169091	1	0.166	0.222	Down
Extracellular matrix organization Homo sapiens R-HSA-1474244	5	0.167	0.223	Up
Arachidonic acid metabolism Homo sapiens R-HSA-2142753	1	0.169	0.225	Down
Constitutive Signaling by NOTCH1 HD+PEST Domain Mutants Homo sapiens R-HSA-2894862	1	0.172	0.226	Down
Signaling by NOTCH1 in Cancer Homo sapiens R-HSA-2644603	1	0.172	0.226	Down
Signaling by NOTCH1 PEST Domain Mutants in Cancer Homo sapiens R-HSA-2644602	1	0.172	0.226	Down
Constitutive Signaling by NOTCH1 PEST Domain Mutants Homo sapiens R-HSA-2644606	1	0.172	0.226	Down
Signaling by NOTCH1 HD+PEST Domain Mutants in Cancer Homo sapiens R-HSA-2894858	1	0.172	0.226	Down
Cell-cell junction organization Homo sapiens R-HSA-421270	1	0.173	0.226	Down
Gastrin-CREB signalling pathway via PKC and MAPK Homo sapiens R-HSA-881907	7	0.174	0.227	Up
Cleavage of Growing Transcript in the Termination Region Homo sapiens R-HSA-109688	1	0.179	0.232	Down
RNA Polymerase II Transcription Termination Homo sapiens R-HSA-73856	1	0.179	0.232	Down
Post-Elongation Processing of the Transcript Homo sapiens R-HSA-76044	1	0.179	0.232	Down
Triglyceride Biosynthesis Homo sapiens R-HSA-75109	1	0.188	0.243	Down
PI Metabolism Homo sapiens R-HSA-1483255	1	0.192	0.247	Down
Ca2+ pathway Homo sapiens R-HSA-4086398	1	0.194	0.249	Down
Antigen processing: Ubiquitination & Proteasome degradation Homo sapiens R-HSA-983168	4	0.198	0.253	Up
Rho GTPase cycle Homo sapiens R-HSA-194840	2	0.202	0.257	Up
COPII (Coat Protein 2) Mediated Vesicle Transport Homo sapiens R-HSA-204005	1	0.202	0.257	Down
Homo sapiens R-HSA-159236	1	0.205	0.260	Down
Processing of Capped Intron-Containing Pre-mRNA Homo sapiens R-HSA-72203	3	0.206	0.261	Up
Platelet activation, signaling and aggregation Homo sapiens R-HSA-76002	4	0.207	0.261	Up
Homo sapiens R-HSA-5617472	1	0.210	0.264	Down
Activation of HOX genes during differentiation Homo sapiens R-HSA-5619507	1	0.210	0.264	Down
Glycosaminoglycan metabolism Homo sapiens R-HSA-1630316	2	0.211	0.264	Up
Vesicle-mediated transport Homo sapiens R-HSA-5653656	7	0.221	0.276	Up
Cell-Cell communication Homo sapiens R-HSA-1500931	2	0.223	0.278	Up
trans-Golgi Network Vesicle Budding Homo sapiens R-HSA-199992	1	0.224	0.278	Down
Clathrin derived vesicle budding Homo sapiens R-HSA-421837	1	0.224	0.278	Down
Transport of Mature Transcript to Cytoplasm Homo sapiens R-HSA-72202	1	0.233	0.287	Down
Meiosis Homo sapiens R-HSA-1500620	1	0.233	0.287	Down
Adaptive Immune System Homo sapiens R-HSA-1280218	10	0.234	0.288	Up
Signaling by NOTCH1 Homo sapiens R-HSA-1980143	1	0.236	0.290	Down
RIG-I/MDA5 mediated induction of IFN-alpha/beta pathways Homo sapiens R-HSA-168928	1	0.238	0.291	Down
Metabolism of polyamines Homo sapiens R-HSA-351202	1	0.239	0.292	Down
Signaling by the B Cell Receptor (BCR) Homo sapiens R-HSA-983705	3	0.248	0.302	Up
Glucose metabolism Homo sapiens R-HSA-70326	1	0.250	0.304	Down
Mitochondrial translation termination Homo sapiens R-HSA-5419276	1	0.253	0.306	Down
Phase II conjugation Homo sapiens R-HSA-156580	1	0.253	0.306	Down
Membrane Trafficking Homo sapiens R-HSA-199991	6	0.254	0.306	Up
Transcriptional Regulation by TP53 Homo sapiens R-HSA-3700989	5	0.257	0.309	Up
Mitochondrial translation initiation Homo sapiens R-HSA-5368286	1	0.259	0.309	Down
Mitochondrial translation elongation Homo sapiens R-HSA-5389840	1	0.259	0.309	Down
Nonsense-Mediated Decay (NMD) Homo sapiens R-HSA-927802	1	0.260	0.309	Down
Nonsense Mediated Decay (NMD) enhanced by the EJC; Homo sapiens R-HSA-975957	1	0.260	0.309	Down
PI3K Cascade Homo sapiens R-HSA-109704	1	0.261	0.310	Down
Ion channel transport Homo sapiens R-HSA-983712	3	0.264	0.312	Up
Metabolism of vitamins and cofactors Homo sapiens R-HSA-196854	2	0.264	0.312	Up
Regulation of insulin secretion Homo sapiens R-HSA-422356	1	0.267	0.315	Down
Platelet homeostasis Homo sapiens R-HSA-418346	1	0.281	0.330	Down

Class I MHC mediated antigen processing & presentation Homo sapiens R-HSA-983169	4	0.282	0.331	Up
Signaling by Rho GTPases Homo sapiens R-HSA-194315	5	0.283	0.331	Up
Fatty acid, triacylglycerol, and ketone body metabolism Homo sapiens R-HSA-535734	3	0.285	0.331	Up
PI3K/AKT Signaling in Cancer Homo sapiens R-HSA-2219528	1	0.285	0.331	Down
Mitochondrial translation Homo sapiens R-HSA-5368287	1	0.284	0.331	Down
Peptide hormone metabolism Homo sapiens R-HSA-2980736	1	0.292	0.338	Down
Opioid Signalling Homo sapiens R-HSA-111885	1	0.293	0.339	Down
Diseases of glycosylation Homo sapiens R-HSA-3781865	1	0.298	0.343	Down
RHO GTPase Effectors Homo sapiens R-HSA-195258	3	0.327	0.375	Up
Cellular response to heat stress Homo sapiens R-HSA-3371556	1	0.327	0.375	Down
Metabolism of water-soluble vitamins and cofactors Homo sapiens R-HSA-196849	1	0.328	0.375	Down
Stimuli-sensing channels Homo sapiens R-HSA-2672351	1	0.330	0.377	Down
Visual phototransduction Homo sapiens R-HSA-2187338	1	0.332	0.378	Down
MHC class II antigen presentation Homo sapiens R-HSA-2132295	1	0.339	0.385	Down
Signal Transduction Homo sapiens R-HSA-162582	30	0.344	0.390	Up
Asparagine N-linked glycosylation Homo sapiens R-HSA-446203	3	0.348	0.393	Up
Homo sapiens R-HSA-198933	1	0.363	0.409	Down
Homo sapiens R-HSA-425366	1	0.372	0.419	Down
Downstream signaling events of B Cell Receptor (BCR) Homo sapiens R-HSA-1168372	2	0.375	0.421	Down
Organelle biogenesis and maintenance Homo sapiens R-HSA-1852241	4	0.387	0.431	Up
Interferon Signaling Homo sapiens R-HSA-913531	2	0.386	0.431	Down
TCF dependent signaling in response to WNT Homo sapiens R-HSA-201681	2	0.387	0.431	Down
Signaling by NOTCH Homo sapiens R-HSA-157118	1	0.401	0.446	Down
Assembly of the primary cilium Homo sapiens R-HSA-5617833	2	0.404	0.448	Down
Integration of energy metabolism Homo sapiens R-HSA-163685	1	0.405	0.448	Down
PI-3K cascade:FGFR1 Homo sapiens R-HSA-5654689	1	0.424	0.463	Down
PI-3K cascade:FGFR3 Homo sapiens R-HSA-5654710	1	0.424	0.463	Down
PI3K events in ERBB4 signaling Homo sapiens R-HSA-1250342	1	0.424	0.463	Down
PIP3 activates AKT signaling Homo sapiens R-HSA-1257604	1	0.424	0.463	Down
PI-3K cascade:FGFR4 Homo sapiens R-HSA-5654720	1	0.424	0.463	Down
PI-3K cascade:FGFR2 Homo sapiens R-HSA-5654695	1	0.424	0.463	Down
PPARA activates gene expression Homo sapiens R-HSA-1989781	1	0.427	0.465	Down
Metabolism of lipids and lipoproteins Homo sapiens R-HSA-556833	8	0.430	0.467	Up
GAB1 signalosome Homo sapiens R-HSA-180292	1	0.433	0.469	Down
Metabolism of carbohydrates Homo sapiens R-HSA-71387	3	0.437	0.472	Up
Diseases of signal transduction Homo sapiens R-HSA-5663202	3	0.438	0.472	Up
PI3K/AKT activation Homo sapiens R-HSA-198203	1	0.438	0.472	Down
Regulation of lipid metabolism by PPARalpha; Homo sapiens R-HSA-400206	1	0.441	0.474	Down
Signaling by Wnt Homo sapiens R-HSA-195721	3	0.444	0.476	Up
ER to Golgi Anterograde Transport Homo sapiens R-HSA-199977	1	0.451	0.482	Down
Mitotic G1-G1/S phases Homo sapiens R-HSA-453279	1	0.457	0.487	Down
Role of LAT2/NTAL/LAB on calcium mobilization Homo sapiens R-HSA-2730905	1	0.456	0.487	Down
RNA Polymerase II Transcription Homo sapiens R-HSA-73857	1	0.461	0.490	Down
Generic Transcription Pathway Homo sapiens R-HSA-212436	9	0.510	0.541	Up
Metabolism of amino acids and derivatives Homo sapiens R-HSA-71291	3	0.512	0.542	Up
G alpha (i) signalling events Homo sapiens R-HSA-418594	2	0.519	0.548	Down
Major pathway of rRNA processing in the nucleolus Homo sapiens R-HSA-6791226	1	0.523	0.551	Down
Beta-catenin independent WNT signaling Homo sapiens R-HSA-3858494	1	0.527	0.554	Down
G alpha (s) signalling events Homo sapiens R-HSA-418555	1	0.546	0.572	Down
rRNA processing Homo sapiens R-HSA-72312	1	0.570	0.596	Down
Phospholipid metabolism Homo sapiens R-HSA-1483257	1	0.583	0.609	Down
Disease Homo sapiens R-HSA-1643685	7	0.611	0.635	Down
G2/M Transition Homo sapiens R-HSA-69275	1	0.610	0.635	Down
Mitotic G2-G2/M phases Homo sapiens R-HSA-453274	1	0.616	0.639	Down
Peptide ligand-binding receptors Homo sapiens R-HSA-375276	1	0.670	0.693	Down
G alpha (q) signalling events Homo sapiens R-HSA-416476	1	0.681	0.703	Down
Post-translational protein modification Homo sapiens R-HSA-597592	4	0.749	0.770	Down
Class A/1 (Rhodopsin-like receptors) Homo sapiens R-HSA-373076	2	0.748	0.770	Down

Signaling by GPCR Homo sapiens R-HSA-372790	11	0.806	0.825	Down
Metabolism of proteins Homo sapiens R-HSA-392499	9	0.806	0.825	Down
Metabolism Homo sapiens R-HSA-1430728	18	0.821	0.839	Down
SLC-mediated transmembrane transport Homo sapiens R-HSA-425407	1	0.846	0.862	Down
DNA Repair Homo sapiens R-HSA-73894	1	0.857	0.872	Down
Transmembrane transport of small molecules Homo sapiens R-HSA-382551	4	0.889	0.902	Down
Infectious disease Homo sapiens R-HSA-5663205	1	0.904	0.916	Down
GPCR ligand binding Homo sapiens R-HSA-500792	2	0.908	0.918	Down
Gene Expression Homo sapiens R-HSA-74160	13	0.912	0.920	Down
Cellular responses to stress Homo sapiens R-HSA-2262752	1	0.914	0.920	Down
Cell Cycle Homo sapiens R-HSA-1640170	2	0.959	0.963	Down
Cell Cycle, Mitotic Homo sapiens R-HSA-69278	1	0.962	0.964	Down
GPCR downstream signaling Homo sapiens R-HSA-388396	4	0.987	0.987	Down

Table C.4: Signalling pathways along with the adjusted p-values and the number of overlapping genes obtained after pathway enrichment analysis applied to the full scale analysis results using the Reactome database.

Gene Set Name	**# Genes in Gene Set (K)**	**Description**	**# Genes in Overlap (k)**	**k/K**	**p-value**	**FDR q-value**
P53_DN.V1_DN	193	Genes down-regulated in NCI-60 panel of cell lines with mutated TP53 [Gene ID=7157].	13	0.067	1.13E-17	2.14E-15
IL21_UP.V1_UP	185	Genes up-regulated in Sez-4 cells (T lymphocyte) that were first starved of IL2 [Gene ID=3558] and then stimulated with IL21 [Gene ID=59067].	6	0.032	7.52E-07	7.1E-05
E2F1_UP.V1_DN	186	Genes down-regulated in mouse fibroblasts over-expressing E2F1 [Gene ID=1869] gene.	4	0.022	0.000	0.018

Table C.5: Overlaps have been detected using gene set enrichment analysis performed using a hypergeometric distribution. The false discovery rate analog of the hypergeometric p-value is displayed after correction for multiple hypothesis testing according to Benjamini and Hochberg.

Gene Name	**Sign**	**Spearman Correlation**	**Area**	**SD**	**Mean fold change of BRAF mutation with respect to wild type**	**Protein-protein interaction network distance to BRAF**
MYO5A	+	0.955	0.531	0.261	1.358	4
UCN2	-	-0.999	0.517	0.184	1.016	NI
TSPYL5	+	0.886	0.513	0.146	0.838	NI
S100A1	+	0.812	0.488	0.189	1.263	NI
GPNMB	+	1	0.424	0.196	1.169	3
ACP5	-	-0.998	0.359	0.149	1.039	NI
FCGR2A	-	-0.588	0.341	0.158	1.250	3
GAS7	-	-0.513	0.285	0.096	1.067	NI
CITED1	0	-0.603	0.28	0.348	1.630	3
SPRY4	-	-0.611	0.274	0.127	1.228	2
HPS4	0	-0.801	0.245	0.196	1.410	NI
C3orf70	-	-0.957	0.239	0.124	1.145	NI
CD44	+	0.868	0.239	0.164	1.413	3
ACOT7	0	0.999	0.239	0.123	1.184	NI
RAP2B	0	0.927	0.236	0.179	1.254	NI
KCNJ13	0	-0.604	0.205	0.094	1.101	3
ALX1	-	-1	0.202	0.099	1.104	NI
PLAT	-	-0.405	0.201	0.121	1.312	4
RETSAT	0	0.689	0.201	0.142	1.127	NI
GSN	+	0.588	0.196	0.109	1.079	4
CDH19	0	0.943	0.185	0.102	0.933	NI
ATP1B3	-	-1	0.178	0.115	1.063	NI

BAZ1A	+	-0.29	0.173	0.105	1.109	4
FAM78B	0	-0.732	0.173	0.163	1.231	NI
SLC16A4	-	-0.298	0.166	0.117	1.234	NI
DSTYK	+	0.937	0.166	0.136	0.943	NI
ST6GALNAC2	0	-0.815	0.164	0.102	1.264	NI
MFSD12	0	-0.788	0.16	0.148	1.130	NI
MTARC1	-	-0.846	0.158	0.098	1.289	NI
GJA3	0	-0.85	0.157	0.075	1.071	NI
CYP27A1	-	-0.743	0.156	0.09	1.373	NI
ZNF292	+	-0.474	0.154	0.118	1.194	NI
CRYL1	0	0.84	0.152	0.13	1.330	NI
EGLN1	-	-0.442	0.15	0.119	1.053	3
TRPV2	0	0.769	0.147	0.118	1.074	4
MITF	+	1	0.146	0.106	0.743	2
TBC1D7	0	-0.603	0.146	0.118	1.304	3
SLC6A8	0	-0.263	0.144	0.111	0.941	NI
PTPRZ1	-	-0.808	0.139	0.138	1.074	4
PLOD3	0	0.696	0.132	0.135	1.166	4
IGSF11	0	-0.723	0.132	0.094	1.266	NI
ANKRD7	+	0.92	0.131	0.12	1.241	NI
KIAA1549L	0	-0.986	0.128	0.109	1.231	NI
LINC00518	+	0.754	0.127	0.13	1.093	NI
RTN3P1	0	0.824	0.127	0.087	1.176	NI
PRSS33	0	0.507	0.118	0.275	1.635	NI
RAB38	0	0.039	0.113	0.118	1.377	NI
TRAK2	0	-0.42	0.109	0.123	1.296	NI
KANK1	0	-0.493	0.107	0.113	1.345	3
GYPC	+	-0.3	0.105	0.092	1.072	4
HCCAT5	0	0.654	0.103	0.131	1.180	NI
TYR	-	0.467	0.1	0.098	1.110	4
BFSP1	-	-0.804	0.1	0.113	1.406	NI
TYRP1	0	0.457	0.1	0.097	1.326	3
STEAP1B	0	-0.301	0.091	0.116	1.202	NI
IGSF8	0	-0.668	0.09	0.129	1.313	5
ASB9	0	0.513	0.086	0.115	1.142	NI
SPRED1	0	-0.556	0.067	0.116	1.239	4
TBC1D16	0	0.757	0.065	0.097	1.117	NI
ITGA9	0	0.785	0.056	0.111	1.154	3
KREMEN1	0	-0.555	0.053	0.086	1.123	4
LAMA4	-	0.344	0.038	0.083	1.151	4
MLANA	0	0.534	0.037	0.097	1.147	NI
ZCCHC24	0	-0.858	0.035	0.097	1.233	NI
KLF9	0	0.932	0.011	0.074	1.064	NI

Table C.7: Gene rankings of all selected genes based on the magnitude of the genetic effect on drug response. A positive (+) sign translates to a positive effect on cells survival after drug administration, a negative (-) sign translates to a negative effect on cells survival and a mixed (0) effect translates to a varying effect on cells survival which depends on drug dosage. Spearman's correlation is calculated between drug dosage and gene estimated coefficient function values as an indicator of the magnitude change of the gene effect over the increasing dosage. Area corresponds to the area under the estimated coefficient curve and the SD corresponds to the standard deviation of the area based on bootstrapping. Mean fold change is calculated between the selected gene expression values of the cell lines carrying *BRAF* mutations with respect to wild type. Protein-protein interaction network distance is computed based on the shortest interaction path between the BRAF gene and each of the selected genes. Here, NI denotes absence of any interaction.

Pathway	Number of genes	P-value	Adjusted p-value
Senescence and Autophagy in Cancer WP615	3	0.000	0.000
Vitamin D Metabolism WP1531	1	0.000	0.000
Dopamine metabolism WP2436	1	0.000	0.000
NOTCH1 regulation of human endothelial cell calcification WP3413	1	0.000	0.000
Hippo-Merlin Signaling Dysregulation WP4541	3	0.001	0.007
RANKL/RANK (Receptor activator of NFKB (ligand)) Signaling Pathway WP2018	2	0.001	0.007
miRNA targets in ECM and membrane receptors WP2911	1	0.001	0.007
Blood Clotting Cascade WP272	1	0.001	0.007
Hereditary leiomyomatosis and renal cell carcinoma pathway WP4206	1	0.002	0.010
Globo Sphingolipid Metabolism WP1424	1	0.002	0.010
Osteoclast Signaling WP12	1	0.003	0.014
Type 2 papillary renal cell carcinoma WP4241	1	0.007	0.028
Photodynamic therapy-induced HIF-1 survival signaling WP3614	1	0.007	0.028
Spinal Cord Injury WP2431	2	0.008	0.030
Fibrin Complement Receptor 3 Signaling Pathway WP4136	1	0.011	0.030
IL-5 Signaling Pathway WP127	1	0.010	0.030
Vitamin A and Carotenoid Metabolism WP716	1	0.012	0.030
Prostaglandin Synthesis and Regulation WP98	1	0.011	0.030
Regulation of Microtubule Cytoskeleton WP2038	1	0.012	0.030
Exercise-induced Circadian Regulation WP410	1	0.011	0.030
Prader-Willi and Angelman Syndrome WP3998	1	0.011	0.030
Calcium Regulation in the Cardiac Cell WP536	2	0.014	0.030
Vitamin B12 Metabolism WP1533	1	0.014	0.030
Oxidation by Cytochrome P450 WP43	1	0.014	0.030
Hepatitis C and Hepatocellular Carcinoma WP3646	1	0.018	0.037
Endochondral Ossification WP474	1	0.019	0.038
Kit receptor signaling pathway WP304	1	0.020	0.039
Complement and Coagulation Cascades WP558	1	0.021	0.039
Folate Metabolism WP176	1	0.023	0.041
Selenium Micronutrient Network WP15	1	0.025	0.043
ncRNAs involved in Wnt signaling in hepatocellular carcinoma WP4336	1	0.027	0.045
PPAR signaling pathway WP3942	1	0.029	0.047
Focal Adhesion WP306	2	0.036	0.055
Arrhythmogenic Right Ventricular Cardiomyopathy WP2118	1	0.036	0.055
LncRNA involvement in canonical Wnt signaling and colorectal cancer WP4258	1	0.037	0.055
Pathways Regulating Hippo Signaling WP4540	1	0.039	0.056
Wnt Signaling Pathway and Pluripotency WP399	1	0.040	0.056
Neural Crest Differentiation WP2064	1	0.045	0.062
Integrin-mediated Cell Adhesion WP185	1	0.050	0.065
Wnt Signaling WP428	1	0.050	0.065
NRF2 pathway WP2884	1	0.074	0.094
Ebola Virus Pathway on Host WP4217	1	0.077	0.094
TGF-beta Signaling Pathway WP366	1	0.078	0.094
Endoderm Differentiation WP2853	1	0.080	0.095
Focal Adhesion-PI3K-Akt-mTOR-signaling pathway WP3932	2	0.086	0.099
Ectoderm Differentiation WP2858	1	0.089	0.101
PI3K-Akt Signaling Pathway WP4172	2	0.112	0.119
Regulation of Actin Cytoskeleton WP51	1	0.112	0.119
Metapathway biotransformation Phase I and II WP702	1	0.112	0.119
Circadian rhythm related genes WP3594	1	0.153	0.159
VEGFA-VEGFR2 Signaling Pathway WP3888	1	0.208	0.212
Nuclear Receptors Meta-Pathway WP2882	1	0.303	0.303

Table C.6: Signalling pathways along with the adjusted p-values and the number of overlapping genes obtained after pathway enrichment analysis applied to the resistant cell line analysis results.

Bibliography

Abbas-Aghababazadeh, F., Lu, P., and Fridley, B. L. (2019). Nonlinear mixed-effects models for modeling in vitro drug response data to determine problematic cancer cell lines. *Scientific reports*, 9(1):1–9.

Abrahamsen, P. (1997). A review of gaussian random fields and correlation functions. *Norsk Regnesentral/Norwegian Computing Center*, page 64.

Abramson, I. and Müller, H.-G. (1994). Estimating direction fields in autonomous equation models, with an application to system identification from cross-sectional data. *Biometrika*, 81(4):663–672.

Andrade, D. F. and Tavares, H. R. (2005). Item response theory for longitudinal data: Population parameter estimation. *Journal of Multivariate Analysis*, 95:1–22.

Apanasovich, T. V. and Genton, M. G. (2010). Cross-covariance functions for multivariate random fields based on latent dimensions. *Biometrika*, 97(1):15–30.

Aykul, S. and Martinez-Hackert, E. (2016). Determination of half-maximal inhibitory concentration using biosensor-based protein interaction analysis. *Analytical biochemistry*, 508:97–103.

Baltes, M. M. and Carstensen, L. L. (1996). The process of successful ageing. *Ageing & Society*, 16:397–422.

Baltes, P. B. and Baltes, M. M. (1990). *Successful Aging*. Cambridge University Press.

Banack, H. R., Kaufman, J. S., Wactawski-Wende, J., Troen, B. R., and Stovitz, S. D. (2019). Investigating and remediating selection bias in geriatrics research: the selection bias toolkit. *Journal of the American Geriatrics Society*, 67(9):1970–1976.

Barretina, J., Caponigro, G., Stransky, N., Venkatesan, K., Margolin, A. A., Kim, S., Wilson, C. J., Lehár, J., Kryukov, G. V., Sonkin, D., et al. (2012). The cancer cell line encyclopedia enables predictive modelling of anticancer drug sensitivity. *Nature*, 483(7391):603–607.

Bartholomew, D. J. (1980). Factor analysis for categorical data. *Journal of the Royal Statistical Society: Series B (Methodological)*, 42(3):293–312.

Beard, J. R., Jotheeswaran, A. T., Cesari, M., and Carvalho, I. A. D. (2019). The structure and predictive value of intrinsic capacity in a longitudinal study of ageing. *BMJ Open*, 9.

Ben-David, U., Siranosian, B., Ha, G., Tang, H., Oren, Y., Hinohara, K., Strathdee, C. A., Dempster, J., Lyons, N. J., Burns, R., Nag, A., Kugener, G., Cimini, B., Tsvetkov, P., Maruvka, Y. E., O'Rourke, R., Garrity, A., Tubelli, A. A., Bandopadhayay, P., Tsherniak, A., Vazquez, F., Wong, B., Birger, C., Ghandi, M., Thorner, A. R., Bittker, J. A., Meyerson, M., Getz, G., Beroukhim, R., and Golub, T. R. (2018). Genetic and transcriptional evolution alters cancer cell line drug response. *Nature*, 560:325–330.

Benjamini, Y. and Hochberg, Y. (1995). Controlling the false discovery rate: a practical and powerful approach to multiple testing. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 57(1):289–300.

Biemer, P. P., Groves, R. M., Lyberg, L. E., Mathiowetz, N. A., and Sudman, S. (2013). *Measurement errors in surveys*, volume 548. John Wiley & Sons.

Billingsley, P. (2008). *Probability and measure*. John Wiley & Sons.

Blackwell, M., Honaker, J., and King, G. (2017). A unified approach to measurement error and missing data: overview and applications. *Sociological Methods & Research*, 46(3):303–341.

Blattman, C., Jamison, J., Koroknay-Palicz, T., Rodrigues, K., and Sheridan, M. (2016). Measuring the measurement error: A method to qualitatively validate survey data. *Journal of Development Economics*, 120:99–112.

Boehm, J. S., Garnett, M. J., Adams, D. J., Francies, H. E., Golub, T. R., Hahn, W. C.,

Iorio, F., McFarland, J. M., Parts, L., and Vazquez, F. (2021). Cancer research needs a better map.

Bollen, K. A. and Curran, P. J. (2006). *Latent curve models: A structural equation perspective*, volume 467. John Wiley & Sons.

Bornn, L., Shaddick, G., and Zidek, J. V. (2012). Modeling nonstationary processes through dimension expansion. *Journal of the American Statistical Association*, 107(497):281–289.

Börsch-Supan, A., Brandt, M., Hunkler, C., Kneip, T., Korbmacher, J., Malter, F., Schaan, B., Stuck, S., and Zuber, S. (2013). Data resource profile: the survey of health, ageing and retirement in europe (share). *International journal of epidemiology*, 42(4):992–1001.

Boudiny, K. (2013). 'active ageing': From empty rhetoric to effective policy tool. *Ageing and Society*, 33:1077–1098.

Boudiny, K. and Mortelmans, D. (2011). A critical perspective: Towards a broader understanding of 'active ageing'. *Electronic Journal of Applied Psychology*, 7:8–14.

Bousquet, J., Kuh, D., Bewick, M., Standberg, T., Farrell, J., Pengelly, R., Joel, M.-E., Mañas, L. R., Mercier, J., Bringer, J., et al. (2015). Operational definition of active and healthy ageing (aha): A conceptual framework. *The journal of nutrition, health & aging*, 19:955–960.

Bowling, A. (1993). The concepts of successful and positive ageing. *Family Practice*, 10:449–453.

Bowling, A. (2007). Aspirations for older age in the 21st century: What is successful aging? *International Journal of Aging and Human Development*, 64:263–297.

Bowling, A. and Dieppe, P. (2005). What is successful ageing and who should define it? *Bmj*, 331:1548–1551.

Brehm, W. (1987). Competent aging with sports. *Zeitschrift fur Gerontologie*, 20(6):336–344.

Burant, C. J. (2016). Latent growth curve models: Tracking changes over time. *The International Journal of Aging and Human Development*, 82:336–350.

Byrne, B. M. and Crombie, G. (2010). Modeling and testing change: An introduction to the latent growth curve model. *Understanding Statistics*, 2:177–203.

Bülow, M. H. and Söderqvist, T. (2014). Successful ageing: A historical overview and critical analysis of a successful concept. *Journal of Aging Studies*, 31:139–149.

Caballero, F. F., Soulis, G., Engchuan, W., Sánchez-Niubó, A., Arndt, H., Ayuso-Mateos, J. L., Haro, J. M., Chatterji, S., and Panagiotakos, D. B. (2017). Advanced analytical methodologies for measuring healthy ageing and its determinants, using factor analysis and machine learning techniques: the athlos project. *Scientific Reports*, 7(1):1–13.

Cagnone, S., Moustaki, I., and Vasdekis, V. (2009). Latent variable models for multivariate longitudinal ordinal responses. *British Journal of Mathematical and Statistical Psychology*, 62:401–415.

Caprara, M., Molina, M. Á., Schettini, R., Santacreu, M., Orosa, T., Mendoza-Núñez, V. M., Rojas, M., and Fernández-Ballesteros, R. (2013). Active aging promotion: results from the vital aging program. *Current Gerontology and Geriatrics Research*, 2013.

Chang, J. C., Wooten, E. C., Tsimelzon, A., Hilsenbeck, S. G., Gutierrez, M. C., Elledge, R., Mohsin, S., Osborne, C. K., Chamness, G. C., Allred, D. C., et al. (2003). Gene expression profiling for the prediction of therapeutic response to docetaxel in patients with breast cancer. *The Lancet*, 362(9381):362–369.

Chen, E. Y., Tan, C. M., Kou, Y., Duan, Q., Wang, Z., Meirelles, G. V., Clark, N. R., and Ma'ayan, A. (2013). Enrichr: interactive and collaborative html5 gene list enrichment analysis tool. *BMC Bioinformatics*, 14(1):128.

Chen, J. and Zhang, S. (2016). Integrative analysis for identifying joint modular patterns of gene-expression and drug-response data. *Bioinformatics*, 32(11):1724–1732.

Chen, X. (2015). Relative deprivation and individual well-being: Low status and a feeling of relative deprivation are detrimental to health and happiness. *IZA world of labor: evidence-based policy making*, 2015.

Chen, Y. and Zhang, S. (2020). A latent gaussian process model for analysing intensive longitudinal data. *British Journal of Mathematical and Statistical Psychology*, 73(2):237–260.

Christensen, K., Doblhammer, G., Rau, R., and Vaupel, J. W. (2009). Ageing populations: the challenges ahead. *The Lancet*, 374:1196–1208.

Chu, W., Li, R., and Reimherr, M. (2016). Feature screening for time-varying coefficient models with ultrahigh dimensional longitudinal data. *The Annals of Applied Statistics*, 10(2):596.

Clegg, A., Bates, C., Young, J., Ryan, R., Nichols, L., Ann Teale, E., Mohammed, M. A., Parry, J., and Marshall, T. (2016). Development and validation of an electronic frailty index using routine primary care electronic health record data. *Age and Ageing*, 45(3):353–360.

Clegg, A., Young, J., Iliffe, S., Rikkert, M. O., and Rockwood, K. (2013). Frailty in elderly people. *The Lancet*, 381(9868):752–762.

Colomer, C., Margalef, P., Villanueva, A., Vert, A., Pecharroman, I., Solé, L., González-Farré, M., Alonso, J., Montagut, C., Martinez-Iniesta, M., et al. (2019). Ikkα kinase regulates the dna damage response and drives chemo-resistance in cancer. *Molecular Cell*, 75(4):669–682.

Comunanza, V., Corà, D., Orso, F., Consonni, F. M., Middonti, E., Di Nicolantonio, F., Buzdin, A., Sica, A., Medico, E., Sangiolo, D., et al. (2017). Vegf blockade enhances the antitumor effect of brafv 600e inhibition. *EMBO Molecular Medicine*, 9(2):219–237.

Cook, D., Brown, D., Alexander, R., March, R., Morgan, P., Satterthwaite, G., and Pangalos, M. N. (2014). Lessons learned from the fate of astrazeneca's drug pipeline: a five-dimensional framework. *Nature Reviews Drug Discovery*, 13(6):419–431.

Corrie, P. G. (2008). Cytotoxic chemotherapy: clinical aspects. *Medicine*, 36(1):24–28.

Cosco, T. D., Prina, A. M., Perales, J., Stephan, B. C. M., and Brayne, C. (2014). Operational definitions of successful aging: a systematic review. *International Psychogeriatrics*, 26:373.

Crawford, R. and Mei, P. (2018). An overview of the elsa'end of life'data. *The Institute for Fiscal Studies*.

Davis-Plourde, K. L., Mayeda, E. R., Lodi, S., Filshtein, T., Beiser, A., Gross, A. L., Seshadri, S., Glymour, M. M., and Tripodis, Y. (2022). Joint models for estimating determinants of cognitive decline in the presence of survival bias. *Epidemiology*, 33(3):362–371.

Dawson, M. and Müller, H.-G. (2018). Dynamic modeling of conditional quantile trajectories, with application to longitudinal snippet data. *Journal of the American Statistical Association*, 113(524):1612–1624.

De São José, J. M., Timonen, V., Amado, C. A. F., and Santos, S. P. (2017). A critique of the active ageing index. *Journal of aging studies*, 40:49–56.

Dean, C., Lin, X., Neuhaus, J., Wang, L., Wu, L., and Yi, G. (2009). Emerging issues in the analysis of longitudinal datathe. In *Banff International Research Workshop, Report by Organizers*. Citeseer.

Delle Fave, A., Bassi, M., Boccaletti, E. S., Roncaglione, C., Bernardelli, G., and Mari, D. (2018). Promoting well-being in old age: The psychological benefits of two training programs of adapted physical activity. *Frontiers in psychology*, 9:828.

Delpuech, O., Rooney, C., Mooney, L., Baker, D., Shaw, R., Dymond, M., Wang, D., Zhang, P., Cross, S., Veldman-Jones, M., Wilson, J., Davies, B. R., Dry, J. R., Kilgour, E., and Smith, P. D. (2016). Identification of pharmacodynamic transcript biomarkers in response to fgfr inhibition by azd4547. *Molecular Cancer Therapeutics*, 15(11):2802–2813.

Demakakos, P., Hacker, E., and Gjonça, E. (2006). 11. perceptions of ageing. *Retirement, health and relationships of the older population in England*, page 339.

Depp, C. A., Glatt, S. J., and Jeste, D. V. (2007). Recent advances in research on successful or healthy aging. *Current psychiatry reports*, 9:7–13.

Depp, C. A. and Jeste, D. V. (2006). Definitions and predictors of successful aging: a comprehensive review of larger quantitative studies. *The American Journal of Geriatric Psychiatry*, 14:6–20.

Diener, E., Emmons, R. A., Larsen, R. J., and Griffin, S. (1985). The satisfaction with life scale. *Journal of personality assessment*, 49(1):71–75.

Diggle, P. J., Heagerty, P., Liang, K.-Y., and Zeger, S. L. (2002). *Analysis of longitudinal data*. Oxford university press.

DiMatteo, I., Genovese, C. R., and Kass, R. E. (2001). Bayesian curve-fitting with free-knot splines. *Biometrika*, 88(4):1055–1071.

Duncan, T. E., Duncan, S. C., and Strycker, L. A. (2013). *An introduction to latent variable growth curve modeling: Concepts, issues, and application*. Routledge.

Dunteman, G. H. (1989). *Principal components analysis*. Sage.

Esses, G., Andreopoulos, E., Lin, H.-M., Arya, S., and Deiner, S. (2018). A comparison of three frailty indices in predicting morbidity and mortality after on-pump aortic valve replacement. *Anesthesia and analgesia*, 126(1):39.

Faber, M. V., Wiel, A. B.-V. D., Exel, E. V., Gussekloo, J., Lagaay, A. M., Dongen, E. V., Knook, D. L., Geest, S. V. D., and Westendorp, R. G. (2001). Successful aging in the oldest old. *Archives of Internal Medicine*, 161:2694–2700.

Falcetta, F., Lupi, M., Colombo, V., and Ubezio, P. (2013). Dynamic rendering of the heterogeneous cell response to anticancer treatments. *PLoS Computational Biology*, 9(10):e1003293.

Fan, J., Feng, Y., and Song, R. (2011). Nonparametric independence screening in sparse ultra-high-dimensional additive models. *Journal of the American Statistical Association*, 106(494):544–557.

Fan, J., Ma, Y., and Dai, W. (2014). Nonparametric independence screening in sparse ultra-high-dimensional varying coefficient models. *Journal of the American Statistical Association*, 109(507):1270–1284.

Feng, L., Nyunt, M. S. Z., Gao, Q., Feng, L., Yap, K. B., and Ng, T.-P. (2017). Cognitive frailty and adverse health outcomes: findings from the singapore longitudinal ageing studies (slas). *Journal of the American Medical Directors Association*, 18(3):252–258.

Fisher, A. L. (2005). Just what defines frailty? *Journal of the American Geriatrics Society*, 53:2229–2230.

Fitzmaurice, G., Davidian, M., Verbeke, G., and Molenberghs, G. (2008). *Longitudinal data analysis*. CRC press.

Flatt, T. (2012). A new definition of aging? *Frontiers in genetics*, 3:148.

Foster, L. and Walker, A. (2015). Active and successful aging: A european policy perspective. *Gerontologist*, 55:83–90.

Franco, O. H., Karnik, K., Osborne, G., Ordovas, J. M., Catt, M., and van der Ouderaa, F. (2009). Changing course in ageing research: the healthy ageing phenotype. *Maturitas*, 63:13–19.

Franse, C. B., van Grieken, A., Qin, L., Melis, R. J., Rietjens, J. A., and Raat, H. (2017). Socioeconomic inequalities in frailty and frailty components among community-dwelling older citizens. *PloS one*, 12(11):e0187946.

Gale, C. R., Cooper, C., and Aihie Sayer, A. (2014). Prevalence of frailty and disability: findings from the english longitudinal study of ageing. *Age and ageing*, 44(1):162–165.

Gao, X., Shahbaba, B., and Ombao, H. (2018). Modeling binary time series using gaussian processes with application to predicting sleep states. *Journal of Classification*, 35:549–579.

Garfein, A. J. and Herzog, A. R. (1995). Robust aging among the young-old, old-old, and oldest-old. *The Journals of Gerontology Series B: Psychological Sciences and Social Sciences*, 50(2):S77–S87.

Geenens, G. et al. (2011). Curse of dimensionality and related issues in nonparametric functional regression. *Statistics Surveys*, 5:30–43.

Genton, M. G. and Kleiber, W. (2015). Cross-covariance functions for multivariate geostatistics. *Statistical Science*, 30(2):147–163.

Glass, T. A. (2003). Assessing the success of successful aging. *Annals of Internal Medicine*, 139:382–383.

Gneiting, T., Kleiber, W., and Schlather, M. (2010). Matérn cross-covariance functions for multivariate random fields. *Journal of the American Statistical Association*, 105(491):1167–1177.

Gonçalves, F. B. and Gamerman, D. (2018). Exact bayesian inference in spatiotemporal cox processes driven by multivariate gaussian processes. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 80(1):157–175.

Greenacre, M. and Blasius, J. (2006). *Multiple correspondence analysis and related methods*. CRC press.

Greven, S. and Scheipl, F. (2017). A general framework for functional regression modelling. *Statistical Modelling*, 17:1–35.

Hall, P., Müller, H.-G., and Yao, F. (2008). Modelling sparse generalized longitudinal observations with latent gaussian processes. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 70:703–723.

Hall, P., Racine, J., and Li, Q. (2004). Cross-validation and the estimation of conditional probability densities. *Journal of the American Statistical Association*, 99(468):1015–1026.

Hao, Q., Dong, B., Yang, M., Dong, B., and Wei, Y. (2018). Frailty and cognitive impairment in predicting mortality among oldest-old people. *Frontiers in aging neuroscience*, 10:295.

Hastie, T. and Tibshirani, R. (1993). Varying-coefficient models. *Journal of the Royal Statistical Society: Series B (Methodological)*, 55(4):757–779.

He, B. and Luo, S. (2016). Joint modeling of multivariate longitudinal measurements and survival data with applications to parkinson's disease. *Statistical methods in medical research*, 25(4):1346–1358.

Hedeker, D. and Gibbons, R. D. (2006). *Longitudinal data analysis*. Wiley-Interscience.

Hogan, D. B. (2003). Models, definitions, and criteria of frailty. *Aging clinical and experimental research*, 15:1–29.

Hooper, D., Coughlan, J., and Mullen, M. (2008). Equation modelling: Guidelines for determining model fit. *Electronic Journal of Business Research Methods*, 6(1):53–60.

Hornby, A. S. (2005). *Oxford Advanced Learner's Dictionary of Current English*. Oxford University Press, 7 edition.

Hsieh, C.-A., von Eye, A. A., and Maier, K. S. (2010). Using a multivariate multilevel polytomous item response theory model to study parallel processes of change: The dynamic association between adolescents' social isolation and engagement with delinquent peers in the national youth survey. *Multivariate Behavioral Research*, 45(3):508–552.

Hu, L.-t. and Bentler, P. M. (1999). Cutoff criteria for fit indexes in covariance structure analysis: Conventional criteria versus new alternatives. *Structural equation modeling: a multidisciplinary journal*, 6(1):1–55.

Huang, J. Z., Wu, C. O., and Zhou, L. (2004). Polynomial spline estimation and inference for varying coefficient models with longitudinal data. *Statistica Sinica*, pages 763–788.

Hung, L. W., Kempen, G. I., and Vries, N. K. D. (2010). Cross-cultural comparison between academic and lay views of healthy ageing: A literature review. *Ageing and Society*, 30:1373–1391.

Hyde, M., Wiggins, R. D., Higgs, P., and Blane, D. B. (2003). A measure of quality of life in early old age: the theory, development and properties of a needs satisfaction model (casp-19). *Aging & mental health*, 7(3):186–194.

Hyman, D. M., Taylor, B. S., and Baselga, J. (2017). Implementing genome-driven oncology. *Cell*, 168(4):584–599.

Ibrahim, J. G., Zhu, H., and Tang, N. (2008). Model selection criteria for missing-data problems using the em algorithm. *Journal of the American Statistical Association*, 103(484):1648–1658.

Iorio, F., Knijnenburg, T. A., Vis, D. J., Bignell, G. R., Menden, M. P., Schubert, M., Aben, N., Gonçalves, E., Barthorpe, S., Lightfoot, H., et al. (2016). A landscape of pharmacogenomic interactions in cancer. *Cell*, 166(3):740–754.

Jabrayilov, R., Emons, W. H., and Sijtsma, K. (2016). Comparison of classical test theory and item response theory in individual change assessment. *Applied psychological measurement*, 40(8):559–572.

Jassal, B., Matthews, L., Viteri, G., Gong, C., Lorente, P., Fabregat, A., Sidiropoulos, K., Cook, J., Gillespie, M., Haw, R., et al. (2020). The reactome pathway knowledgebase. *Nucleic Acids Research*, 48(D1):D498–D503.

Ji, R.-R., de Silva, H., Jin, Y., Bruccoleri, R. E., Cao, J., He, A., Huang, W., Kayne, P. S., Neuhaus, I. M., Ott, K.-H., et al. (2009). Transcriptional profiling of the dose response: a more powerful approach for characterizing drug activities. *PLoS Computational Biology*, 5(9).

Johnson, F. B., Sinclair, D. A., and Guarente, L. (1999). Molecular biology of aging. *Cell*, 96:291–302.

Juster, F. T. and Suzman, R. (1995). An overview of the health and retirement study. *Journal of Human Resources*, pages S7–S56.

Kaiser, H. F. (1960). The application of electronic computers to factor analysis. *Educational and psychological measurement*, 20(1):141–151.

Karch, J. D., Brandmaier, A. M., and Voelkle, M. C. (2020). Gaussian process panel modeling—machine learning inspired analysis of longitudinal panel data. *Frontiers in psychology*, 11:351.

Kerschner, H. and Pegues, J. O. M. (1998). Productive aging: A quality of life agenda. *Journal of the American Dietetic Association*, 98:1445–1448.

Keshava, N., Toh, T. S., Yuan, H., Yang, B., Menden, M. P., and Wang, D. (2019). Defining subpopulations of differential drug response to reveal novel target populations. *NPJ Systems Biology and Applications*, 5:36.

Kline, R. B. (2015). *Principles and practice of structural equation modeling*. Guilford publications.

Kneip, A. and Liebl, D. (2020). On the optimal reconstruction of partially observed functional data. *The Annals of Statistics*, 48(3):1692–1717.

Koenker, R. (2004). Quantile regression for longitudinal data. *Journal of Multivariate Analysis*, 91(1):74–89.

Koenker, R., Ng, P., and Portnoy, S. (1994). Quantile smoothing splines. *Biometrika*, 81(4):673–680.

Koukouli, E., Wang, D., Dondelinger, F., and Park, J. (2021). A regularized functional regression model enabling transcriptome-wide dosage-dependent association study of cancer drug response. *PLoS Computational Biology*, 17(1):e1008066.

Kraus, D. (2015). Components and completion of partially observed functional data. *Journal of the Royal Statistical Society: Series B: Statistical Methodology*, pages 777–801.

Kuh, D., Karunananthan, S., Bergman, H., and Cooper, R. (2014). A life-course approach to healthy ageing: Maintaining physical capability. *Proceedings of the Nutrition Society*, 73:237–248.

Kuleshov, M. V., Jones, M. R., Rouillard, A. D., Fernandez, N. F., Duan, Q., Wang, Z., Koplev, S., Jenkins, S. L., Jagodnik, K. M., Lachmann, A., et al. (2016). Enrichr: a comprehensive gene set enrichment analysis web server 2016 update. *Nucleic Acids Research*, 44(W1):W90–W97.

Kusumastuti, S., Derks, M. G. M., Tellier, S., Nucci, E. D., Lund, R., Mortensen, E. L., and Westendorp, R. G. J. (2016). Successful ageing: A study of the literature using citation network analysis. *Maturitas*, 93:4–12.

Kutalik, Z., Beckmann, J. S., and Bergmann, S. (2008). A modular approach for integrative analysis of large-scale gene-expression and drug-response data. *Nature Biotechnology*, 26(5):531.

Lamb, J., Crawford, E. D., Peck, D., Modell, J. W., Blat, I. C., Wrobel, M. J., Lerner, J., Brunet, J.-P., Subramanian, A., Ross, K. N., et al. (2006). The connectivity map: using gene-expression signatures to connect small molecules, genes, and disease. *Science*, 313(5795):1929–1935.

Lara, J., Godfrey, A., Evans, E., Heaven, B., Brown, L. J., Barron, E., Rochester, L.,

Meyer, T. D., and Mathers, J. C. (2013a). Towards measurement of the healthy ageing phenotype in lifestyle-based intervention studies. *Maturitas*, 76(2):189–199.

Lara, J., Godfrey, A., Evans, E., Heaven, B., Brown, L. J. E., Barron, E., Rochester, L., Meyer, T. D., and Mathers, J. C. (2013b). Towards measurement of the healthy ageing phenotype in lifestyle-based intervention studies. *Maturitas*, 76:189–199.

Lee, T. K., Wickrama, K. K. A., and O'Neal, C. W. (2018). Application of latent growth curve analysis with categorical responses in social behavioral research. *Structural Equation Modeling*, 25:294–306.

Li, H., Zhang, Y., Carroll, R. J., Keadle, S. K., Sampson, J. N., and Matthews, C. E. (2017). A joint modeling and estimation method for multivariate longitudinal data with mixed types of responses to analyze physical activity data generated by accelerometers. *Statistics in Medicine*, 36:4028–4040.

Li, N., Elashoff, R. M., Li, G., and Tseng, C.-H. (2012). Joint analysis of bivariate longitudinal ordinal outcomes and competing risks survival times with nonparametric distributions for random effects. *Statistics in medicine*, 31(16):1707–1721.

Li, P. and Chen, S. (2016). A review on gaussian process latent variable models. *CAAI Transactions on Intelligence Technology*, 1:366–376.

Li, Q. and Racine, J. S. (2008). Nonparametric estimation of conditional cdf and quantile functions with mixed categorical and continuous data. *Journal of Business & Economic Statistics*, 26(4):423–434.

Li, Y., Schoufour, J., Wang, D. D., Dhana, K., Pan, A., Liu, X., Song, M., Liu, G., Shin, H. J., Sun, Q., Al-Shaar, L., Wang, M., Rimm, E. B., Hertzmark, E., Stampfer, M. J., Willett, W. C., Franco, O. H., and Hu, F. B. (2020). Healthy lifestyle and life expectancy free of cancer, cardiovascular disease, and type 2 diabetes: Prospective cohort study. *The BMJ*, 368.

Liebl, D. and Rameseder, S. (2019). Partially observed functional data: The case of systematically missing parts. *Computational Statistics & Data Analysis*, 131:104–115.

Little, R. J. (1988). A test of missing completely at random for multivariate data with missing values. *Journal of the American statistical Association*, 83(404):1198–1202.

Lupien, S. J. and Wan, N. (2004). Successful ageing: From cell to self. *Philosophical Transactions of the Royal Society B: Biological Sciences*, 359:1413–1426.

Manzano, J. L., Layos, L., Bugés, C., de los Llanos Gil, M., Vila, L., Martinez-Balibrea, E., and Martínez-Cardús, A. (2016). Resistant mechanisms to braf inhibitors in melanoma. *Annals of Translational Medicine*, 4(12).

Marioni, R. E., Davies, G., Hayward, C., Liewald, D., Kerr, S. M., Campbell, A., Luciano, M., Smith, B. H., Padmanabhan, S., Hocking, L. J., et al. (2014). Molecular genetic contributions to socioeconomic status and intelligence. *Intelligence*, 44:26–32.

Marshall, A., Nazroo, J., Tampubolon, G., and Vanhoutte, B. (2015). Cohort differences in the levels and trajectories of frailty among older people in england. *J Epidemiol Community Health*, 69(4):316–321.

Martin, F. C. and Romero Ortuño, R. (2019). Longitudinal studies of ageing: from insights to impacts: commentary to accompany themed collection on longitudinal studies. *Age and Ageing*, 48(4):481–485.

Martin, P., Kelly, N., Kahana, B., Kahana, E., Willcox, B. J., Willcox, D. C., and Poon, L. W. (2015). Defining successful aging: A tangible or elusive concept? *The Gerontologist*, 55:14–25.

Martinson, M. and Berridge, C. (2015). Successful aging and its discontents: A systematic review of the social gerontology literature. *The gerontologist*, 55(1):58–69.

Marvelde, J. M. T., Glas, C. A., Landeghem, G. V., and Damme, J. V. (2006). Application of multidimensional item response theory models to longitudinal data. *Educational and Psychological Measurement*, 66:5–34.

Masyn, K. E., Petras, H., and Liu, W. (2014). Growth curve models with categorical outcomes. *Encyclopedia of criminology and criminal justice*, pages 2013–2025.

McPhee, J. S., French, D. P., Jackson, D., Nazroo, J., Pendleton, N., and Degens, H. (2016). Physical activity in older age: perspectives for healthy ageing and frailty. *Biogerontology*, 17(3):567–580.

Mekli, K., Marshall, A., Nazroo, J., Vanhoutte, B., and Pendleton, N. (2015). Genetic variant of interleukin-18 gene is associated with the frailty index in the english longitudinal study of ageing. *Age and ageing*, 44(6):938–942.

Mendoza-Núñez, V. M., Sarmiento-Salmorán, E., Marín-Cortés, R., Martínez-Maldonado, M. D. l. L., and Ruiz-Ramos, M. (2018). Influence of the self-perception of old age on the effect of a healthy aging program. *Journal of clinical medicine*, 7(5):106.

Michel, J.-P. and Sadana, R. (2017). "healthy aging" concepts and measures. *Journal of the American Medical Directors Association*, 18:460–464.

Mindell, J., Biddulph, J. P., Hirani, V., Stamatakis, E., Craig, R., Nunn, S., and Shelton, N. (2012). Cohort profile: the health survey for england. *International journal of epidemiology*, 41(6):1585–1593.

Mishra, S. and Carleton, R. N. (2015). Subjective relative deprivation is associated with poorer physical and mental health. *Social Science & Medicine*, 147:144–149.

Mitnitski, A. B., Mogilner, A. J., and Rockwood, K. (2001). Accumulation of deficits as a proxy measure of aging. *The Scientific World Journal*, 1:323–336.

Molenberghs, G., Beunckens, C., Sotto, C., and Kenward, M. G. (2008). Every missingness not at random model has a missingness at random counterpart with equal fit. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 70(2):371–388.

Molenberghs, G., Fitzmaurice, G., Kenward, M. G., Tsiatis, A., and Verbeke, G. (2014). *Handbook of missing data methodology*. CRC Press.

Mortezaee, K., Salehi, E., Mirtavoos-mahyari, H., Motevaseli, E., Najafi, M., Farhood, B., Rosengren, R. J., and Sahebkar, A. (2019). Mechanisms of apoptosis modulation by curcumin: Implications for cancer therapy. *Journal of Cellular Physiology*, 234(8):12537–12550.

Moser, C., Spagnoli, J., and Santos-Eggimann, B. (2011). Self-perception of aging and vulnerability to adverse outcomes at the age of 65–70 years. *Journals of Gerontology Series B: Psychological Sciences and Social Sciences*, 66(6):675–680.

Muthén, B. (2002). Beyond sem: general latent variable modeling. *Behaviormetrika*, 29:81–117.

Niederstrasser, N. G., Rogers, N. T., and Bandelow, S. (2019). Determinants of frailty development and progression using a multidimensional frailty index: Evidence from the english longitudinal study of ageing. *PLoS One*, 14(10):e0223799.

Nielsen, S. F. (2000). The stochastic em algorithm: estimation and asymptotic results. *Bernoulli*, pages 457–489.

O'Donovan, K. J., Ma, K., Guo, H., Wang, C., Sun, F., Han, S. B., Kim, H., Wong, J. K., Charron, J., Zou, H., et al. (2014). B-raf kinase drives developmental axon growth and promotes axon regeneration in the injured mature cns. *Journal of Experimental Medicine*, 211(5):801–814.

Paek, I., Li, Z., and Park, H.-J. (2016). Specifying ability growth models using a multidimensional item response model for repeated measures categorical ordinal item response data. *Multivariate behavioral research*, 51(4):569–580.

Pattinson, B. L. (2018). *Measurement models for understanding the social challenges of caring for the elderly in Brazil and England*. Lancaster University (United Kingdom).

Peel, N., Bartlett, H., and McClure, R. (2004). Healthy ageing: How is it defined and measured? *Australasian Journal on Ageing*, 23.

Phelan, E. A. and Larson, E. B. (2002). "successful aging"—where next? *Journal of the American Geriatrics Society*, 50:1306–1308.

Pinkhasov, R. M., Wong, J., Kashanian, J., Lee, M., Samadi, D. B., Pinkhasov, M. M., and Shabsigh, R. (2010). Are men shortchanged on health? perspective on health care utilization and health risk behavior in men and women in the united states. *International journal of clinical practice*, 64(4):475–487.

Proust, C., Jacqmin-Gadda, H., Taylor, J. M., Ganiayre, J., and Commenges, D. (2006). A nonlinear model with latent process for cognitive evolution using multivariate longitudinal data. *Biometrics*, 62:1014–1024.

Proust-Lima, C., Amieva, H., and Jacqmin-Gadda, H. (2013). Analysis of multivariate mixed longitudinal data: A flexible latent process approach. *British Journal of Mathematical and Statistical Psychology*, 66:470–487.

Qu, A. and Li, R. (2006). Quadratic inference functions for varying-coefficient models with longitudinal data. *Biometrics*, 62(2):379–391.

Quaranta, S. and Thomas, F. (2017). Pharmacogenetics of anti-cancer drugs: State of the art and implementation–recommendations of the french national network of pharmacogenetics. *Therapies*, 72(2):205–215.

Radloff, L. S. (1977). The ces-d scale: A self-report depression scale for research in the general population. *Applied psychological measurement*, 1(3):385–401.

Rangaswami, H., Bulbule, A., and Kundu, G. C. (2006). Osteopontin: role in cell signaling and cancer progression. *Trends in Cell Biology*, 16(2):79–87.

Rasmussen, C. E. and Williams, C. K. I. (2006). *Gaussian processes for machine learning*. The MIT Press.

Raudenbush, S. W. and Chan, W.-S. (1992). Growth curve analysis in accelerated longitudinal designs. *Journal of Research in Crime and Delinquency*, 29(4):387–411.

Reckase, M. (2009). *Multidimensional Item Response Theory*. Springer.

Relling, M. V. and Dervieux, T. (2001). Pharmacogenetics and cancer therapy. *Nature Reviews Cancer*, 1(2):99.

Rizopoulos, D. (2012). *Joint models for longitudinal and time-to-event data: With applications in R*. CRC press.

Rockwood, K., Andrew, M., and Mitnitski, A. (2007). A comparison of two approaches to measuring frailty in elderly people. *The Journals of Gerontology Series A: Biological Sciences and Medical Sciences*, 62(7):738–743.

Rockwood, K. and Howlett, S. E. (2018). Fifteen years of progress in understanding frailty and health in aging.

Rockwood, K. and Mitnitski, A. (2007). Frailty in relation to the accumulation of deficits. *The Journals of Gerontology Series A: Biological Sciences and Medical Sciences*, 62:722–727.

Rogers, N. T., Marshall, A., Roberts, C. H., Demakakos, P., Steptoe, A., and Scholes, S. (2017). Physical activity and trajectories of frailty among older adults: Evidence from the english longitudinal study of ageing. *PloS one*, 12:e0170878.

Rose, M. (1991). *Evolutionary Biology of Aging*. Oxford University Press, New York.

Rose, M., Flatt, T., Graves Jr, J. L., Greer, L. F., Martinez, D. E., Matos, M., Mueller, L., Shmookler Reis, R. J., and Shahrestani, P. (2012). What is aging? *Frontiers in genetics*, 3:134.

Ross, D. T., Scherf, U., Eisen, M. B., Perou, C. M., Rees, C., Spellman, P., Iyer, V., Jeffrey, S. S., Van de Rijn, M., Waltham, M., et al. (2000). Systematic variation in gene expression patterns in human cancer cell lines. *Nature genetics*, 24(3):227–235.

Rowe, J. W. and Kahn, R. L. (1987). Human aging: Usual and successful. *Science*, 237:143–149.

Rowe, J. W. and Kahn, R. L. (1997). Successful aging. *The Gerontologist*, 37.

Rudnicka, E., Napierala, P., Podfigurna, A., Meczekalski, B., Smolarczyk, R., and Grymowicz, M. (2020). The world health organization (who) approach to healthy ageing. *Maturitas*, 139:6–11.

Rummel, R. J. (1988). *Applied factor analysis*. Northwestern University Press.

Sanchez-Niubo, A., Forero, C. G., Wu, Y.-T., Giné-Vázquez, I., Prina, M., Fuente, J. D. L., Daskalopoulou, C., Critselis, E., Torre-Luque, A. D. L., Panagiotakos, D., Arndt, H., Ayuso-Mateos, J. L., Bayes-Marin, I., Bickenbach, J., Bobak, M., Caballero, F. F., Chatterji, S., Egea-Cortés, L., García-Esquinas, E., Leonardi, M., Koskinen, S., Koupil, I., Mellor-Marsá, B., Olaya, B., Pajak, A., Prince, M., Raggi, A., Rodríguez-Artalejo, F., Sanderson, W., Scherbov, S., Tamosiunas, A., Tobias-Adamczyk, B., Tyrovolas, S., and Haro, J. M. (2020). Development of a common scale for measuring healthy ageing across the world: results from the athlos consortium. *International Journal of Epidemiology*.

Sanders, J. L., Minster, R. L., Barmada, M. M., Matteini, A. M., Boudreau, R. M., Christensen, K., Mayeux, R., Borecki, I. B., Zhang, Q., Perls, T., et al. (2014). Heritability of and mortality prediction with a longevity phenotype: the healthy aging

index. *Journals of Gerontology Series A: Biomedical Sciences and Medical Sciences*, 69(4):479–485.

Schulz, R. and Heckhausen, J. (1996). A life span model of successful aging. *American psychologist*, 51(7):702.

Searle, S. D., Mitnitski, A., Gahbauer, E. A., Gill, T. M., and Rockwood, K. (2008). A standard procedure for creating a frailty index. *BMC Geriatrics*, 8:1–10.

Sharma, N. and Jha, S. (2016). Nlr-regulated pathways in cancer: opportunities and obstacles for therapeutic interventions. *Cellular and Molecular Life Sciences*, 73(9):1741–1764.

Sharma, S. P. (2012). Ras mutations and the development of secondary tumours in patients given braf inhibitors. *The Lancet Oncology*, 13(3):e91.

Shashikumar, S. A., Huang, K., Konetzka, R. T., and Maddox, K. E. J. (2020). Claims-based frailty indices: A systematic review. *Medical care*, 58(9):815–825.

Shi, J. Q. and Choi, T. (2011). *Gaussian process regression analysis for functional data*. CRC Press.

Siegel, R. L., Miller, K. D., Fuchs, H. E., and Jemal, A. (2022). Cancer statistics, 2022. *CA: A Cancer Journal for Clinicians*, 72(1):7–33.

Silverbush, D., Grosskurth, S., Wang, D., Powell, F., Gottgens, B., Dry, J., and Fisher, J. (2017). Cell-specific computational modeling of the pim pathway in acute myeloid leukemia. *Cancer Research*, 77(4):827–838.

Skrondal, A. and Rabe-Hesketh, S. (2007). Latent variable modelling: A survey. *Scandinavian Journal of Statistics*, 34(4):712–745.

Slenter, D. N., Kutmon, M., Hanspers, K., Riutta, A., Windsor, J., Nunes, N., Mélius, J., Cirillo, E., Coort, S. L., Digles, D., Ehrhart, F., Giesbertz, P., Kalafati, M., Martens, M., Miller, R., Nishida, K., Rieswijk, L., Waagmeester, A., Eijssen, L. M. T., Evelo, C. T., Pico, A. R., and Willighagen, E. L. (2017). WikiPathways: a multifaceted pathway database bridging metabolomics to other omics research. *Nucleic Acids Research*, 46(D1):D661–D667.

Smith, H. J. and Huo, Y. J. (2014). Relative deprivation: How subjective experiences of inequality influence social behavior and health. *Policy Insights from the Behavioral and Brain Sciences*, 1(1):231–238.

Smith, W. N., Del Rossi, G., Adams, J. B., Abderlarahman, K., Asfour, S. A., Roos, B. A., and Signorile, J. F. (2010). Simple equations to predict concentric lower-body muscle power in older adults using the 30-second chair-rise test: a pilot study. *Clinical interventions in aging*, 5:173.

Solit, D. B. and Rosen, N. (2011). Resistance to braf inhibition in melanomas. *New England Journal of Medicine*, 364(8):772–774.

Somel, M., Thornton, J. M., and Dönertaş, H. M. (2020). Temporal changes in the gene expression heterogeneity during brain development and aging. *Scientific reports*, 10(1):1–15.

Song, R., Yi, F., and Zou, H. (2014). On varying-coefficient independence screening for high-dimensional varying-coefficient models. *Statistica Sinica*, 24(4):1735.

Stein, M. L. (2012). *Interpolation of spatial data: some theory for kriging*. Springer Science & Business Media.

Stephen, J., James, N., Bram, V., and Tarani, C. (2014). Aging and subjective well-being in later life. *1079-5014*.

Stephens, T. (1988). Physical activity and mental health in the united states and canada: evidence from four population surveys. *Preventive medicine*, 17(1):35–47.

Steptoe, A., Breeze, E., Banks, J., and Nazroo, J. (2013). Cohort profile: the english longitudinal study of ageing. *International Journal of pidemiology*, 42(6):1640–1648.

Steptoe, A., Deaton, A., and Stone, A. A. (2015). Psychological wellbeing, health and ageing. *Lancet*, 385(9968):640.

Strawbridge, W. J., Wallhagen, M. I., and Cohen, R. D. (2002). Successful aging and well-being. *The Gerontologist*, 42:727–733.

Strehler, B. L. B. L. (1962). *Time, Cells and Ageing*. Academic Press, New York.

Subbiah, V., Kreitman, R. J., Wainberg, Z. A., Cho, J. Y., Schellens, J. H., Soria, J. C., Wen, P. Y., Zielinski, C., Cabanillas, M. E., Urbanowitz, G., et al. (2018). Dabrafenib and trametinib treatment in patients with locally advanced or metastatic braf v600–mutant anaplastic thyroid cancer. *Journal of Clinical Oncology*, 36(1):7.

Subramanian, A., Tamayo, P., Mootha, V. K., Mukherjee, S., Ebert, B. L., Gillette, M. A., Paulovich, A., Pomeroy, S. L., Golub, T. R., Lander, E. S., et al. (2005). Gene set enrichment analysis: a knowledge-based approach for interpreting genome-wide expression profiles. *Proceedings of the National Academy of Sciences*, 102(43):15545–15550.

Szanton, S. L., Seplaki, C. L., Thorpe, R. J., Allen, J. K., and Fried, L. P. (2010). Socioeconomic status is associated with frailty: the women's health and aging studies. *Journal of Epidemiology & Community Health*, 64(01):63–67.

Tabachnick, B. G. and Fidell, L. S. (2001). Using multivariate statistics. allyn and bacon. *Needham Heights, MA*.

Tampubolon, G. (2016a). Trajectories of the healthy ageing phenotype among middle-aged and older britons, 2004–2013. *Maturitas*, 88:9–15.

Tampubolon, G. (2016b). Trajectories of the healthy ageing phenotype among middle-aged and older britons, 2004–2013. *Maturitas*, 88:9–15.

Tansey, W., Li, K., Zhang, H., Linderman, S. W., Rabadan, R., Blei, D. M., and Wiggins, C. H. (2018). Dose-response modeling in high-throughput cancer drug screenings: An end-to-end approach. *arXiv preprint arXiv:1812.05691*.

Tavassoly, I., Hu, Y., Zhao, S., Mariottini, C., Boran, A., Chen, Y., Li, L., Tolentino, R. E., Jayaraman, G., Goldfarb, J., Gallo, J., and Iyengar, R. (2019). Genomic signatures defining responsiveness to allopurinol and combination therapy for lung cancer identified by systems therapeutics analyses. *Molecular Oncology*, 13(8):1725–1743.

Thorndike, R. L. (1953). Who belongs in the family. In *Psychometrika*. Citeseer.

Tomczak, K., Czerwińska, P., and Wiznerowicz, M. (2015). The cancer genome atlas (tcga): an immeasurable source of knowledge. *Contemporary oncology*, 19(1A):A68.

Türei, D., Korcsmáros, T., and Saez-Rodriguez, J. (2016). Omnipath: guidelines and gateway for literature-curated signaling pathway resources. *Nature Methods*, 13(12):966.

United Nations (2021). *Shifting Demographics*.

Vamathevan, J., Clark, D., Czodrowski, P., Dunham, I., Ferran, E., Lee, G., Li, B., Madabhushi, A., Shah, P., Spitzer, M., et al. (2019). Applications of machine learning in drug discovery and development. *Nature reviews Drug discovery*, 18(6):463–477.

Van den Hout, A. and Muniz-Terrera, G. (2016). Joint models for discrete longitudinal outcomes in aging research. *Journal of the Royal Statistical Society: Series C (Applied Statistics)*, 65(1):167–186.

Van der Linden, W. J. (2016). *Handbook of Item Response Theory, Volume One, Models*. CRC Press, first edition. edition.

van der Linden, W. J. and Hambleton, R. K. (2013). *Handbook of modern item response theory*. Springer Science & Business Media.

van Oest, R. (2021). Unconstrained cholesky-based parametrization of correlation matrices. *Communications in Statistics-Simulation and Computation*, 50(11):3607–3613.

Vandenberg-Rodes, A. and Shahbaba, B. (2015). Dependent mat\'ern processes for multivariate time series. *arXiv preprint arXiv:1502.03466*.

Verbeke, G., Fieuws, S., Molenberghs, G., and Davidian, M. (2014). The analysis of multivariate longitudinal data: A review. *Statistical Methods in Medical Research*, 23:42–49.

Vittinghoff, E., Malani, H. M., and Jewell, N. P. (1994). Estimating patterns of cd4 lymphocyte decline using data from a prevalent cohort of hiv infected individuals. *Statistics in Medicine*, 13(11):1101–1118.

Walston, J. D. and Bandeen-Roche, K. (2015). Frailty: a tale of two concepts. *BMC Medicine*, 13(1):1–3.

Wang, B. and Shi, J. Q. (2014). Generalized gaussian process regression model for non-gaussian functional data. *Journal of the American Statistical Association*, 109(507):1123–1133.

Wang, C. and Nydick, S. W. (2020). On longitudinal item response theory models: A didactic. *Journal of Educational and Behavioral Statistics*, 45:339–368.

Wang, D., Hensman, J., Kukaite, G., Toh, T. S., Dry, J. R., Saez-Rodriguez, J., Garnett, M., Menden, M. P., Dondelinger, F., et al. (2020). A statistical framework for assessing pharmacological response and biomarkers with confidence. *BioRxiv*.

Whyte, J., Bergin, O., Bianchi, A., McNally, S., and Martin, F. (2009). Key signalling nodes in mammary gland development and cancer. mitogen-activated protein kinase signalling in experimental models of breast cancer progression and in mammary gland development. *Breast Cancer Research*, 11(5):209.

Wilke, C. O. (2012). Bringing molecules back into molecular evolution. *PLoS Computational Biology*, 8(6):e1002572.

Willcox, D. C., Willcox, B. J., Sokolovsky, J., and Sakihara, S. (2007). The cultural context of "successful aging" among older women weavers in a northern okinawan village: The role of productive activity. *Journal of cross-cultural gerontology*, 22(2):137–165.

World Health Organization (2002). *Active Ageing: A Policy Framework*.

World Health Organization (2017). *Global strategy and action plan on ageing and health*.

Wu, C. O. and Chiang, C.-T. (2000). Kernel smoothing on varying coefficient models with longitudinal dependent variable. *Statistica Sinica*, pages 433–456.

Wu, C. O., Chiang, C.-T., and Hoover, D. R. (1998). Asymptotic confidence regions for kernel smoothing of a varying-coefficient model with longitudinal data. *Journal of the American Statistical Association*, 93(444):1388–1402.

Xue, L. and Qu, A. (2012). Variable selection in high-dimensional varying-coefficient models with global optimality. *Journal of Machine Learning Research*, 13(Jun):1973–1998.

Xue, L., Qu, A., and Zhou, J. (2010). Consistent model selection for marginal generalized additive model for correlated data. *Journal of the American Statistical Association*, 105(492):1518–1530.

Yang, W., Soares, J., Greninger, P., Edelman, E. J., Lightfoot, H., Forbes, S., Bindal, N., Beare, D., Smith, J. A., Thompson, I. R., et al. (2012). Genomics of drug sensitivity in cancer (gdsc): a resource for therapeutic biomarker discovery in cancer cells. *Nucleic Acids Research*, 41(D1):D955–D961.

Yang, X., Hu, F., Liu, J. A., Yu, S., Cheung, M. P. L., Liu, X., Ng, I. O.-L., Guan, X.-Y., Wong, K. K., Sharma, R., et al. (2020). Nuclear dlc1 exerts oncogenic function through association with foxk1 for cooperative activation of mmp9 expression in melanoma. *Oncogene*, 39(20):4061–4076.

Yao, F., Müller, H.-G., and Wang, J.-L. (2005). Functional data analysis for sparse longitudinal data. *Journal of the American statistical association*, 100(470):577–590.

You, L. and Qiu, P. (2021). Joint modeling of multivariate nonparametric longitudinal data and survival data: A local smoothing approach. *Statistics in medicine*, 40(29):6689–6706.

Zaidi, A., Gasior, K., Zolyomi, E., Schmidt, A., Rodrigues, R., and Marin, B. (2017). Measuring active and healthy ageing in europe. *Journal of European Social Policy*, 27:138–157.

Zhang, J., Baran, J., Cros, A., Guberman, J. M., Haider, S., Hsu, J., Liang, Y., Rivkin, E., Wang, J., and Whitty, B. (2011). International cancer genome consortium data portal—a one-stop shop for cancer genomics data. *Database*, 2011.

Zhang, S., Liu, C.-C., Li, W., Shen, H., Laird, P. W., and Zhou, X. J. (2012). Discovery of multi-dimensional modules by integrative analysis of cancer genomic data. *Nucleic acids research*, 40(19):9379–9391.

Zhu, C., Byrd, R., Lu, P., and Nocedal, J. (1995). A limited memory algorithm for bound constrained optimisation. *SIAM J. Sci. Stat. Comp*, 16:1190–1208.